ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 2 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 2 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 78-91

KATTA TIL MODELLARI

Botir Elov[1]
Zilola Xusainova[2]
Sarvinoz Qosimova[3]

Annotatsiya. Bugungi kunda katta til modellari (Large Language Models, LLM) bir qancha sohalarning rivojiga o‘z hissasini qo‘shmoqda. Katta til modellari aniqroq va ilg‘or NLP tizimlarini yaratishga yordam beradi. Katta til modellari – bu katta hajmdagi matnli ma’lumotlarga o‘qitilgan *sun’iy intellekt (artificial intelligence, AI)* dasturiy ta’minoti bo‘lib, til ilovalarini tushunish va tahlil qilish uchun chuqur o‘rganish – *mashinali o‘qitish (machine learning, ML)* algoritmlari deb ataladigan ilg‘or dasturlash texnologiyasidan foydalanadi. Ushbu til modellari turli NLP vazifalarini bajarish uchun qo‘llanilishi mumkin, masalan, *matnni tarjima qilish, kontent yaratish, hissiyotlarni tahlil qilish* va boshqalar. Ushbu maqolada LLMlarni ishlab chiqish bosqichlari va NLP ilovalariga qo‘llash usullari keltiriladi.

Kalit so‘zlar: *Katta til modellari, Large Language Models, LLM, n-gram til modeli, unigram, bigram, modelni baholash, mashinali o‘qitish.*

Kirish

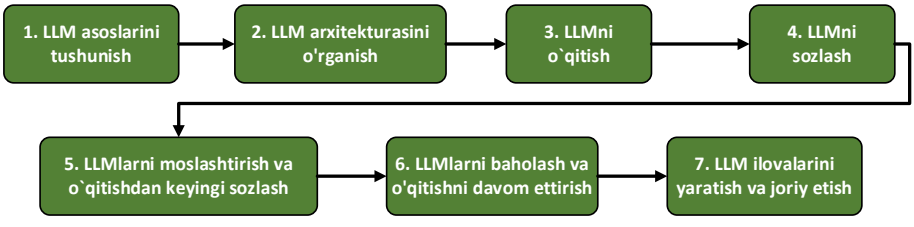
Katta til modellari – bu katta hajmdagi matnli ma’lumotlarga o‘rgatilgan sun’iy intellekt (artificial intelligence, AI) dasturiy ta’minoti. Katta til modellari katta hajmdagi til ma’lumotlarini, jumladan, til ilovalarini tushunish va tahlil qilish uchun chuqur

[1]*Elov Botir Boltayevich – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.*
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2]*Xusainova Zilola Yuldashevna – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti o‘qituvchisi.*
E-pochta: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515
[3]*Qosimova Sarvinoz Furqat qizi – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxasisligi II kurs magistranti.*
E-pochta: sarvinozq54@gmail.com

o‘rganish – mashinali o‘qitish (machine learning, ML) algoritmlari deb ataladigan ilg‘or dasturlash texnologiyasidan foydalanadi.

LLM vositasida xuddi odam qanday yozishi yoki gapirishi kabi (og‘zaki va yozma) nutq yaratishga yordam beradi. Shu jumladan, mashina tilini tushunish, inson va mashina tomonidan yaratilgan til o‘rtasidagi farqni aniqlashni murakkablashtiradi.

Katta til modellari asosida LLM ilovalarini ishlab chiqish uchun quyidagi 7 ta qadamni bajarish lozim [Huang va boshq., 2022; Naveed va boshq., 2023].



1-rasm. LLM ilovalarini ishlab chiqish qadamlari

Bugungi kunda GPT-4 [Achiam va boshq., 2023], Llama [Inan va boshq., 2023], Falcon [Penedo va boshq., 2023] va boshqa ko‘plab katta tilli modellarni chat interfeysi yoki API orqali ishlatish mumkin. Ushbu LLMlar arxitekturasi, ularning ishlash tamoyillari va ulardan qaysi maqsadlarda foydalanish mumkinligi haqida ma’lumotlar keltiriladi.

Katta til modellari yoki LLMlar *matn ma’lumotlarining katta korpusida o‘qitiladigan chuqur o‘rganish modellarining kichik to‘plamidir*. Ular katta hajmga va o‘n milliardlab parametrlarga ega bo‘lib, *tabiiy til vazifalarini* hal qilishda keng miqyosida qo‘llaniladi.

LLMlar izchil, kontekstga mos keladigan va grammatik jihatdan aniq matnni tushunish va yaratish qobiliyatiga ega bo‘lib, ularning ommabopligi sabablari quyidagilardan iborat:

– Turli xildagi til vazifalarini samarali hal qilishi;

– AI asosida tabiiy tilni tushunish va yaratish uchun oldindan tayyorlangan LLMlarning mavjudligi.

LLMlarni quydagi NLP vazifalarini hal qilishda qo‘llash tavsiya qilinadi:

– ***Tabiiy tilni tushunish:*** *LLMlardan his-tuyg‘ularni tahlil qilish, NER obyektni tanib olish va savol-javob tizimlarini ishlab chiqish kabi NLP vazifalarda juda yaxshi natijalarni taqdim qiladi.*

– ***Matn yaratish:*** *Ular chatbotlar va boshqa kontent yaratish NLP vazifalarni hal qilish uchun insonga o‘xshash matn yaratishi mumkin.*

– ***Mashina tarjimasi:*** *LLMlar mashina tarjimasi sifatini*

sezilarli darajada yaxshilagan.

– ***Kontentni umumlashtirish:*** *LLMlar katta hajmli hujjatlarning qisqacha xulosalarini shakllantirishi mumkin. Bunga YouTube video transkriptlarini umumlashtirishni misol sifatida keltirish mumkin.*

Unigram tili modeli

Tabiiy tilni qayta ishlashda n-gramm n ta so‘zdan iborat ketma-ketlikdir [Vaswani va boshq., 2017]. Masalan, “NLP” – unigram (n = 1), “kompyuter lingvistikasi” – bigram (n = 2), “sun’iy intellekt metodlari” – trigram (n = 3).

Modelni o‘qitish

Til modeli gapdagi so‘zning ehtimolligini baholaydi, odatda undan oldin kelgan so‘zlarga asoslanadi. Masalan, **“Men yangi kitobni”** so‘zlar birikmasi uchun bizning maqsadimiz shu jumladagi oldingi so‘zlar asosida gapdagi har bir so‘zning ehtimolini baholashdir:

Berilgan gap:	**"Men yangi kitobni o‘qidim </s>"**
Bashorat qilish:	P("men")
	P("yangi" | "men")
	P("kitobni" | "men yangi")
	P("o‘qidim") | "men yangi kitobni")
	P("</s>") | "men yangi kitobni o‘qidim")

Odatda, til modelida barcha so‘zlar kichik harflar bilan yoziladi va tinish belgilari e’tiborga olinmaydi. **</s>** belgisi gapning oxirini bildiradi.

Unigram tili modeli quyidagi taxminlarga asoslanadi:

1. Har bir so‘zning ehtimoli o‘zidan oldingi so‘zlarga bog‘liq emas.

2. Har bir so‘z ehtimoli mashg‘ulot matnidagi barcha so‘zlar orasidagi uchrash qiymatiga bog‘liq. Ya’ni, modelni o‘rgatish o‘quv matnidagi barcha unigramlar uchun ushbu qiymatlarni hisoblashdan iborat.

$$P_{o'quv}(\text{"o‘qidim"} \mid \text{"men yangi kitobni"}) = P_{o'quv}(\text{"o‘qidim"}) = \frac{n_{o'quv}(\text{"o‘qidim"})}{N_{o'quv}}$$

o'quv matnidagi so'zlarning umumiy soni

1-rasm. O‘quv matnidan “o‘qidim” unigramining taxminiy ehtimoli

Modelni baholash

Barcha unigram ehtimolliklari aniqlanganidan so‘ng, matndagi har bir gapning ehtimolini hisoblash uchun ushbu ehtimollik qiymatlaridan foydalanamiz: *har bir gapning ehtimoli undagi so‘zlar ehtimolliklarining yig‘indisidan iborat.*

$$P_{umumiy}(\text{"Men yangi kitobni o‘qidim </s>"}) =$$

$$P_{o'quv}(\text{"men"})P_{o'quv}(\text{"yangi"}|\text{"men"})P_{o'quv}(\text{"kitobni"}|\text{"men yangi"})P_{o'quv}(\text{"o‘qidim"}|\text{"men yangi kitobni"})$$

$$P_{o'quv}(\text{"[END]"}|\text{"men yangi kitobni o‘qidim"}) =$$

$$P_{o'quv}(\text{"men"})P_{o'quv}(\text{"yangi"})P_{o'quv}(\text{"kitobni"})P_{o'quv}(\text{"o‘qidim"})P_{o'quv}(\text{"</s>"})$$

Yuqorida keltirilgan qoidani *uzb_text1* yoki *uzb_text2* kabi matnlarning umumiy ehtimolini hisoblash mumkin. Matndagi har bir gapning boshqa gaplardan mustaqil, degan sodda taxmin asosida, bu ehtimolni matndagi gaplar ehtimollarining yig‘indisi sifatida aniqlanadi.

Berilgan matn:	**"Men yangi kitobni o‘qidim. U juda qiziqarli ekan"**
$P_{baholash}(T) =$	$= P_{umumiy}(\text{"men yangi kitobni o‘qidim}$ $P_{umumiy}(\text{"u juda qiziqarli ekan </s>"})$
$P_{baholash}(\text{"men yangi kitobni o‘qidim}$ $=$	$P_{o'quv}(\text{"men"})P_{o'quv}(\text{"yangi"})P_{o'quv}(\text{"ki}$ $P_{o'quv}(\text{"o‘qidim"})P_{o'quv}(\text{"</s>"})$
$P_{baholash}(\text{"u juda qiziqarli ekan </s>"})$ $=$	$P_{o'quv}(\text{"u"})P_{o'quv}(\text{"juda"})P_{o'quv}(\text{"qiziqa}$ $P_{o'quv}(\text{"ekan"})P_{o'quv}(\text{"</s>"})$
$\rightarrow P_{baholash}(T) =$	$P_{o'quv}(\text{"men"})P_{o'quv}(\text{"yangi"})P_{o'quv}(\text{"ki}$ $P_{o'quv}(\text{"o‘qidim"})P_{o'quv}(\text{"</s>"})$ $P_{o'quv}(\text{"u"})P_{o'quv}(\text{"juda"})P_{o'quv}(\text{"qiziqa}$ $P_{o'quv}(\text{"ekan"})P_{o'quv}(\text{"</s>"})$

Tabiiy til qoliplarini n-gram metodi vositasida aniqlash

N-gram metodi – matn ma’lumotlaridagi qolip va munosabatlarni aniqlash uchun tabiiy tilni qayta ishlash (Natural Language Processing, NLP)da qo‘llaniladigan matnni tahlil qilish usuli [Elov va boshq., 2024; Elov, 2022]. Ushbu metod matnni n-gram deb ataladigan kichikroq birliklarga bo‘lish va matn ma’lumotlari haqida tushunchaga ega bo‘lish uchun ushbu n-grammlarning chastotasi va ularning korpusda tarqalishini tahlil qilishni o‘z ichiga oladi. N-grammlar so‘zlar, belgilar yoki boshqa har qanday mazmunli matn birliklaridan iborat bo‘lishi mumkin. Til korpusidagi N-grammlar

tahlili muhim ahamiyatga ega bo‘lib, u matn ma’lumotlarini tahlil qilish va ma’lumotlar ichidagi qolip va munosabatlarni aniqlashning sodda, ammo samarali usulini taklif qiladi. N-gramm metodi tilni modellashtirish, matnni tasniflash va his-tuyg‘ularni tahlil qilish kabi turli xil NLP ilovalarini ishlab chiqish uchun foydali bo‘lishi mumkin. N-gram tahlili tilni modellashtirishda matn ma’lumotlaridagi qolip va munosabatlarni aniqlash hamda tabiiy tilni qayta ishlash vazifalari uchun bashoratli modellarni yaratish uchun qo‘llaniladi. Shungdek, N-gramm tahlili matnni tasniflashda matnning asosiy xususiyatlarini aniqlash va matnni oldindan belgilangan toifalarga ajratish uchun ishlatiladi.

LLMlar haqida boshlang‘ich bilimlarga ega bo‘lganidan so‘ng, ushbu LLMlarni asoslaydigan transformer modeli arxitekturasini [Vaswani, 2017] ko‘rib chiqamiz. 2017- yilda Ashish Vaswani tomonidan yozilgan “Attention Is All You Need” maqolasida transformer usuli arxitekturasi tabiiy tilni qayta ishlashda muhim qadam hisoblandi [Vaswani, 2017]:

Asosiy xususiyatlari Self-attention layers, multi-head attention, feed-forward neural networks, encoder-decoder architecture.

Foydalanish holatlari: Transformerlar BERT va GPT kabi LLMlar uchun asosdir.

Transformer arxitekturasi kodlovchi-dekoder arxitekturasidan foydalangan holda, quyidagi muhim xususiyatlarga ega:

Arxitektura	Asosiy xususiyatlar	Mashhur LLMlar	Foydalanish holatlari
Faqat enkoder	Ikki tomonlama kontekstni qayta ishlaydi; Tabiiy tilni tushunish uchun javob beradi;	BERT Shuningdek, BERT arxitekturasi asosidagi RoBERTa, XLNet	• Matn tasnifi • Savol-javob tizimlari
Faqat dekoder	Bir tomonlama til modeli; Avtoregressiv generatsiya;	GPT PaLM	• Matn yaratish • Text completion
Enkoder dekoder	Matnni qayta ishlashga qaratilgan har qanday vazifa	T5 BART	• Umumlashtirish • Tarjima • Savol-javob tizimlari • Hujjatlarni tasniflash

LLM ma'lumotlarini o'qitish

Katta til modellari va transformer arxitekturasi haqida tasavvur hosil bo'lganidan so'ng, LLM ma'lumotlarini o'qitish qadamiga o'tish mumkin [Ali va boshq., 2023]. LLM ma'lumotlarini o'qitish qadami LLMlarning asosini tashkil qiladi, ularni matn ma'lumotlarining katta korpusi asosida tabiiy tilning aspektlari va o'ziga xos jihatlarini tushunishga imkon beradi. Ushbu qadamda quyidagi tamoyillar ko'rib chiqladi:

– **Ma'lumotlarini o'qitish maqsadlari:** Til qoliplari, grammatika va kontekstni o'rganish uchun LLMlarni katta hajmli matn korpusi asosida shakllantirish lozim.

– **Ma'lumotlarini o'qitish uchun matnli korpuslar:** LLMlar katta va xilma-xil matn korpuslari, jumladan, *veb-maqolalar, kitoblar* va boshqa manbalar bo'yicha o'qitiladi. Bular milliardlab, trillionlab matnli tokenlarga ega yirik ma'lumotlar to'plamidir. Ingliz tilidagi ochiq ma'lumotlar to'plamiga C4, BookCorpus, Pile, OpenWebText va boshqalar kiradi. O'zbek tili korpusi bo'yicha bir qator limiy tadqiqotlar amalga oshirilgan bo'lib, jumladan muallif rahbarligida https://uzschoolcorpara.uz/ va https://uznatcorpara.uz/ kabi 2024-yil sentabr holatida 50 millionga yaqin matnlardan iborat o'zbek tili korpuslari ishlab chiqilgan [Elov, 2023].

– **O'qitish jarayoni:** o'qitish jarayonining texnik jihatlarini, jumladan *optimallashtirish algoritmlarini* va *o'qitish davrlari (epochs)* ni tushunib olish zarur. Ma'lumotlardagi shovqinlarni bartaraf qilish kabi boshlang'ich qadamlarni amalga oshirish lozim.

Til korpuslari asosida ma'lumotlar o'qitilganidan so'ng, LLMlarni sozlash qadamini amalga oshirish lozim.

LLM ilovalari

Katta til modellari bilan turli dasturlash texnologiyalari integratsiyasi orqali yangi imkoniyatlarga ega keng doiradagi NLP ilovalari ishlab chiqildi. Quyida bugungi kunda LLMdan keng foydanaladigan NLP ilovalari keltirilgan.

Chatbotlar va virtual yordamchilar

Eng mashhur LLM ilovalaridan biri bu **chatbotlar** va **virtual yordamchilarni** ishlab chiqishdir [Bouchiha va boshq., 2024]. Ushbu modellar odamlarning qanday savol berishini tushunishi va odam aytganiga juda o'xshash javoblarni berishi mumkin. Bunday NLP ilovalari tashkilotlarga o'z mijozlariga inson aralashuvisiz 24/7 xizmat ko'rsatish tizimi orqali mijozlarga xizmat ko'rsatishni yaxshilash imkonini berdi. LLM ilovalari tabiiy tilning nuanslarini tushunishda ish faoliyatini yaxshilash uchun katta hajmdagi

matn ma'lumotlari bo'yicha o'qitilishi mumkin. Ular, shuningdek, so'rovlarga moslashtirilgan javoblarni taqdim etish uchun *moliya, sog'liqni saqlash* va *ta'lim* kabi ma'lum sohalarga moslashish uchun sozlanishi mumkin.

Shuningdek, chatbotlar va virtual yordamchilarni tizim foydalanuvchilari va tizim o'rtasida uzluksiz va tabiiy o'zaro aloqani ta'minlash uchun nutqni aniqlash va tushunish kabi boshqa AI ilovalari bilan integratsiya qilinishi mumkin. Biroq chatbotlar va virtual yordamchilarda LLMdan foydalanishda ham cheklovlar mavjud. Misol uchun, ular har doim ham foydalanuvchi so'rovining kontekstini tushuna olmaydi va ahamiyatsiz/mos bo'lmagan javoblarni taqdim etishi mumkin. Bundan tashqari, ushbu modellarda inson hissiy intellektning yo'qligi foydalanuvchi bilan kamroq *empatik aloqaga* olib kelishi mumkin. Shu sababli, katta til modellaridan foydalanish va mijozlarga xizmat ko'rsatishda inson aralashuvi o'rtasida muvozanatni saqlash juda muhimdir.

Odatda *chatbotlar va virtual yordamchilar* ko'p so'raladigan savollarga javob berish, mahsulot tavsiyalarini berish va hatto tibbiy tashxis qo'yish uchun ishlatilishi mumkin. Bu esa biznes uchun vaqt va resurslarni tejash, shuningdek, mijozlarga xizmat ko'rsatish sifatini yaxshilash imkonini beradi.

Ko'p afzalliklariga qaramay, chatbotlar va virtual yordamchilar ham ba'zi cheklovlarga ega. Asosiy qiyinchiliklardan biri ularning murakkab so'rovlarni tushunish va javob berish qobiliyatidir. Ular oddiy savollarga javob berish va asosiy ma'lumotlarni taqdim etishda samarali bo'lsa-da, ular murakkabroq so'rovlarga javob berishlari ancha murakkab qo'shimcha amallarni talab qiladi.

Kontent yaratish

Yangi kontent yaratish NLP ilovalari rivojida LLM sezilarli ta'sir ko'rsatgan. Bunda **Generativ AI** ilovalari yordamida yozma yoki multimedia kontentini yaratish jarayonini nazarda tutadi.

LLM ilovasi *maqolalar, blog postlari, ijtimoiy media sarlavhalari, mahsulot tavsiflari* va boshqa shu turdagi kontentlarni yaratishi mumkin. Katta hajmdagi matn ma'lumotlarini o'qitish orqali ushbu modellar turli *janrlarning uslubi, ohangi* va *tuzilishini* qamrab olishi va izchil va kontekstga mos keladigan kontentni yaratishi mumkin.

Katta til modellari, shuningdek, muayyan parametrlar yoki takliflar asosida kontent yaratish orqali kontentni avtomatlashtirishda yordam berishi mumkin. Masalan, elektron tijorat veb-sayti ushbu modellardan *mahsulotning xususiyatlari, afzalliklari* va *texnik*

xususiyatlari haqida ma'lumot berish orqali **mahsulot tavsiflarini avtomatik tarzda yaratish** uchun foydalanishi mumkin.

Shuningdek, katta til modeli kontentni *demografiya, afzalliklar* va *ko'rish tarixi* kabi omillarni hisobga olgan holda turli auditoriyalar uchun moslashtirilgan tarzda yaratishi mumkin. Bu esa tashkilotlarga yanada dolzarb va jozibador kontentni taqdim etishga, foydalanuvchi tajribasini oshirishga va mijozlar ehtiyojini qondirish darajasini oshirishga yordam beradi.

Shuni ta'kidlash kerakki, katta til modellari tomonidan yaratilgan kontent lingvist yoki mutaxassislar tomonidan diqqat bilan ko'rib chiqilishi va tahrir qilinishi kerak. Ushbu modellar matn yaratishda muayyan yutuqlarga ega bo'lsa-da, ular qo'shimcha kontekst yoki ijodkorlikni talab qiladigan *xatolar* yoki *noaniqliklarga* ega kontentni yaratishi mumkin.

Inson ishtiroki yakuniy kontentning sifat standartlariga javob berishini va brendning ko'rsatmalariga mos kelishini ta'minlaydi. Biroq, avtomatlashtirilgan kontent yaratish va inson ijodi o'rtasida muvozanatni saqlash muhimdir. Katta til modellari samaradorlikni oshirishni taklif qilsa-da, kontentning haqiqiyligi va o'ziga xosligini saqlab qolish juda muhimdir.

Mashina tarjimasi

Bugungi kunda katta til modellari mashina tarjimasi sohasida ajoyib yutuqlarga erishdi va til to'siqlarini yo'q qilish hamda global muloqotni osonlashtirish jarayonida inqilob qildi. Ushbu modellar tarjima jarayonida inqilob qilib, turli tillarda yanada aniqroq va tabiiy tarjimalarni taklif qildi. LLM katta hajmdagi ko'p tilli ma'lumotlar bo'yicha o'qitilishi natijasida turli *tillarning murakkabligini tushunish* va *til tuzilmalari, idiomatik iboralar* va *madaniy nuanslarini qamrab olish* imkonini beradi. Ushbu keng qamrovli o'qitish LMM modellariga kontekstga mosroq va lingvistik jihatdan aniqroq bo'lgan yuqori sifatli tarjimalarni yaratishga imkon beradi.

Biroq, LLM ilovalari tilni tarjima qilishda hali ham ba'zi cheklovlarga ega bo'lishi mumkin. Ba'zan ular ma'lum sohalarda qo'llaniladigan so'zlarni o'z ichiga olmaydi yoki kam uchraydigan yoki juda o'ziga xos tillarda qiyinchiliklarga duch kelishi mumkin. Shuningdek, hozirda ushbu LMM modellar samaradorligi ancha yaxshilangan bo'lsa-da, tarjimada odamlar bilan bir xil aniqlik darajasiga erishish hali ham qiyin, ayniqsa *madaniy tafsilotlar* va *kontekst* e'tiborga olinishi kerak bo'lgan muhim omillar mavjud.

His-tuyg'ularni tahlil qilish va matnni tasniflash

His-tuyg'ularni tahlil qilish va matnni tasniflash LLMning

asosiy ilovalari bo‘lib, ular bizga matn ma’lumotlarida ifodalangan *fikrlar, his-tuyg‘ular* va *niyatlarni tushunish* va *tahlil qilishda* yordam beradi. Ushbu ilovalar *iste’molchilarning fikr-mulohazalarini tahlil qilish, ijtimoiy media monitoringi, bozor tadqiqotlari* va *kontentni moderatsiya qilish* kabi sohalarda keng qo‘llaniladi.

His-tuyg‘ularni tahlil qilish matn fragmentidagi *hissiyot* yoki *hissiy ohangni,* uning **ijobiy, salbiy** yoki **neytral** ekanligini aniqlashni o‘z ichiga oladi. Katta til modellari katta hajmdagi teglangan ma’lumotlar asosida o‘qitilganligi sababli, bu ularga matnda ifodalangan *his-tuyg‘ularni aniq aniqlash* va *tasniflash* imkonini beradi. Bunday ilovalar tashkilotlarga *mijozlar fikri, brendni tahlil qilish* va *jamoatchilik kayfiyati* haqida qimmatli tushunchalarga ega bo‘lish imkonini beradi. His-tuyg‘ularni tahlil qilish va matnlarni tasniflash uchun katta til modellaridan foydalanishning afzalliklaridan biri ularning matnning kontekstual ma’nosini qamrab olish qobiliyatidir.

Katta til modeli hissiyotlarni tahlil qilish va matn tasniflash vazifalarida bir nechta tillarni ham boshqarishi mumkin. Ular ko‘p tilli ma’lumotlar to‘plami bo‘yicha o‘qitilishi mumkin, bu esa turli tillardagi hissiyotlarni tahlil qilish va tasniflash imkonini beradi. Bunday ilovalar ayniqsa turli mintaqalar uchun ijtimoiy media monitoringini amalga oshirishda foydalidir.

Maxsus LLMlarni ishlab chiqish

Marketing tadqiqotlarida his-tuyg‘ularni tahlil qilish mijozlar nimani yoqtirishini va yoqtirmasligini o‘lchaydigan foydali vositadir. Shuningdek, u yangi zamonaviy tendensiyalarni aniqlashga va marketing kompaniyalari faoliyati yaxshi ketayotganligini tekshirishga yordam beradi.

Matnni tasniflash – bu katta hajmdagi matnli ma’lumotlarini tartibga solish va tahlil qilishda yordam beradigan mashinali o‘qitishning yana bir foydali usuli. Matnni tasniflash tadqiqotchilarga hujjatlar, tadqiqot hujjatlari yoki onlayn forumlar kabi muhim ma’lumotlar va shablonlarni topish imkonini beradi.

Katta til modellari his-tuyg‘ularni tahlil qilish va matn tasnifini sezilarli darajada rivojlantirgan bo‘lsa-da, hali ham hal qilinishi kerak bo‘lgan qiyinchiliklar mavjud. LMMlar ma’lum sohalardagi matn bilan ishlashda, his-tuyg‘ularni aniqlashda yoki madaniy kontekstni tushunishda qiyinchiliklarga duch kelishi mumkin. Hozirda amalga oshirilayotgan ilmiy tadqiqotlar ushbu muammolarni hal qilish va hissiyotlarni tahlil qilish hamda matnlarni tasniflash vazifalarida katta til modellarining ishlashi va umumlashtirilishini yaxshilashga qaratilgan.

Savol-javob tizimlari

Savol-javob tizimlari foydalanuvchilarga tabiiy tilda savollar berish orqali aniq ma'lumotlarni olish imkonini beruvchi katta til modellarining muhim NLP ilovasi hisoblanadi. Ushbu tizimlar keng ko'lamli so'rovlarga to'g'ri va mos javoblar berish uchun katta til modelining tilni tushunishning samarali imkoniyatlaridan foydalanadi. Savol-javob tizimlari kiritilgan *savolni tahlil qilish*, *uning maqsadini tushunish* va *keng bilimlar bazasidan tegishli ma'lumotlarni olish* orqali ishlaydi.

Katta til modellari turli xil ma'lumotlar to'plamlarida, shu jumladan *Internet, kitoblar* va *boshqa manbalardan olingan matnlar* bo'yicha o'qitiladi, bu ularga siafatli javoblarni yaratish uchun katta hajmdagi ma'lumotlardan foydalanish imkonini beradi. LLM vositasida shakllantirilgan savol-javob tizimlarining asosiy afzalliklaridan biri ularning murakkab savollarni tushunish va mazmunli javob berish qobiliyatidir. Ushbu modellar tilning nuanslarini tushunishi, kontekstga bog'liq so'rovlarni boshqarishi va savolda taqdim etilgan ma'lumotlarni hisobga olgan holda javoblarni yaratishi mumkin.

Savol-javob tizimlari *qisqacha faktik javoblar*, *qisqacha javoblar* yoki hatto *batafsil tushuntirishlar berish* uchun mo'ljallangan bo'lishi mumkin. Ular umumiy bilim, tarix, sport va boshqalarni o'z ichiga olgan keng doiradagi mavzularni qamrab olishi mumkin. Foydalanuvchilar suhbat tarzida savollar berishlari mumkin, bu tizim bilan o'zaro aloqani yanada intuitiv va foydalanuvchilar uchun qulay qiladi.

Bugungi kunda savol-javob tizimlari turli sohalarda ko'plab ilovalarga ega. Ta'limda ushbu tizimlar talabalarga topshiriqlari uchun tegishli ma'lumotlarni topishda yoki tushunchalarni aniqlashtirishda yordam berishi mumkin. Ular savollarga tezkor javoblar berib, talabalar va o'qituvchilar uchun vaqt va kuchni tejaydi. Biroq, katta til modeliga asoslangan savol-javob tizimlari ajoyib imkoniyatlarni taqdim etsa-da, ular hali ham turli qiyinchiliklarga duch kelishmoqda. *Noaniq so'rovlar, to'liq bo'lmagan ma'lumotlar* va *muayyan sohaga xos* yoki *texnik savollarni hal qilish* zarurati doimiy izlanish va takomillashtirishni talab qiladi.

Qidiruv tizimlarini takomillashtirish

Qidiruv tizimi mexanizmlari foydalanuvchilarga Internetda tegishli ma'lumotlarni topishga yordam berishda hal qiluvchi rol o'ynaydi. Katta til modellari vositasida qidiruv tizimlarini yaxshilash va *qidiruv natijalarining aniqligi, dolzarbligi* va *foydalanish qulayligini*

oshirish uchun yangi imkoniyatlar taqdim etiladi.

Qidiruv tizimlarini takomillashtirishda katta til modellarining asosiy afzalliklaridan biri ularning tabiiy til so‘rovlarini qayta ishlash qobiliyatidir. Ko‘pincha foydalanuvchilar qidiruv so‘rovlarini savollar yoki to‘liq gaplar shaklida amalga oshirib, mazmunli javoblarni kutadilar.

Katta til modellari ushbu so‘rovlarni qayta ishlashlari va foydalanuvchining maqsadiga to‘g‘ridan-to‘g‘ri javob beradigan tegishli javoblar yoki ma’lumotlarni taqdim etishi mumkin. Katta til modellari qidiruv tizimlarini yaxshilashi mumkin bo‘lgan yana bir soha semantik qidiruvdir. LMMlar *so‘zlar, iboralar* va *tushunchalar* o‘rtasidagi munosabatlarni tushunib, ma’lumotni aniqroq olish imkonini beradi. Shuningdek, katta til modellari personallashtirilgan qidiruvda yordam berishi mumkin. Foydalanuvchining qidiruv tarixini va kontekstli ma’lumotlarni tahlil qilish orqali ushbu modellar qidiruv natijalarini individual qiziqish va imtiyozlarga moslashtirishi mumkin.

Ushbu yondashuv foydalanuvchilarga o‘ziga xos ehtiyojlariga eng mos keladigan kontentini topishga yordam beradi. Katta til modellari qidiruv tizimlarida sezilarli yaxshilanishlarni taklif qilsa-da, hal qilinmagan muammolar hali ham mavjud. Maxfiylik muammolari va qidiruv natijalarining dolzarbligi va xilma-xilligini muvozanatlash uchun algoritmlarni sozlash zarurati e’tibor talab qiladigan sohalardir.

Xulosa

Bugungi kunda LLMlar turli xil NLP ilovalarda imkoniyatlarni taqdim etdi. Bu modellar raqamli dunyomiz bilan o‘zaro munosabatlarimizni inqilob qilish potensialiga ega bo‘lib, biznes va foydalanuvchilarga yanada samarali va moslashtirilgan xizmatlarni taqdim etadi. Shuningdek, ushbu modellardagi turli cheklovlar va axloqiy jihatlarni ham yodda tutish kerak. Katta til modelidan foydalanish axloqiy va mas’uliyatli foydalanishni ta’minlash uchun kuzatilishi va tartibga solinishi kerak. LMMlar noto‘g‘ri va noaniqliklarga yo‘l qo‘ymaslik uchun turli xildagi ma’lumotlarga o‘qitilgan bo‘lishi kerak. Umuman olganda, katta til modellardan hozirda keng miqyosida foydalanilmoqda va biz kelgusi yillarda yanada ilg‘or hamda murakkab modellarni ko‘rishimiz va ulardan kundalik faoliyatimizda foydalanishimiz mumkin.

Foydalanilgan adabiyotlar

Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F. L., McGrew, B. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*., 2023.

Ali M., Fromm M., Thellmann K., Rutmann R., Lübbering M., Leveling J., Flores-Herr N. Tokenizer Choice For LLM Training: Negligible or Crucial?. *arXiv preprint arXiv:2310.08754*., 2023.

Bouchiha M. A., Telnoff Q., Bakkali S., Champagnat R., Rabah M., Coustaty M., Ghamri-Doudane Y. LLMChain: Blockchain-based Reputation System for Sharing and Evaluating Large Language Models. *arXiv preprint arXiv:2404.13236*., 2024.

Elov B. N-gramm til modellari vositasida o‘zbek tilida matn generatsiya qilish. C*omputer linguistics: problems, solutions, prospects*, *1*(1), 2022.

Elov B. O‘zbek-ingliz tillari parallel korpusiga qo‘yiladigan lingvistik va ekstralingvistik talablar. *Computer linguistics: problems, solutions, prospects*, *1*(1), 2023.

Elov B., Alayev R., Abdullayev A., Aloyev N. Yuqori n-gram modellarini o‘zbek tili matnlariga qo‘llash. *digital transformation and artificial intelligence*, *2*(5), 2024. 152-162.

Huang J., Gu S. S., Hou L., Wu Y., Wang X., Yu H., Han J. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*., 2022.

Inan H., Upasani K., Chi J., Rungta R., Iyer K., Mao Y., Khabsa M. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*., 2023.

Naveed H., Khan A. U., Qiu S., Saqib M., Anwar S., Usman M., Mian A. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*., 2023.

Penedo G., Malartic Q., Hesslow D., Cojocaru R., Cappelli A., Alobeidli H., Launay J. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*., 2023.

Vaswani A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Polosukhin I. Attention Is All You Need.(Nips), 2017. *arXiv preprint arXiv:1706.03762*, *10*, S0140525X16001837., 2017.

LARGE LANGUAGE MODELS

Botir Elov[1],
Zilola Xusainova[2],
Sarvinoz Kasimova[3]

Abstract. Today, large language models (Large Language Models, LLM) contribute to the development of several fields. Large language models help create more accurate and advanced NLP systems. Large language models are artificial intelligence (AI) software trained in large volumes of text data, and language applications use advanced programming technology called deep learning - machine learning (ML) algorithms for understanding and analyzing. These language models can be used to perform various NLP tasks, such as text translation, content creation, emotion analysis, etc. This article presents the stages of LLM development and methods for applying NLP to applications.

Key words: *Large language models, Large Language Models, LLM, n-gram language model, unigram, bigram, model evaluation, machine learning.*

References

Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F. L., McGrew, B. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*., 2023.

Ali M., Fromm M., Thellmann K., Rutmann R., Lübbering M., Leveling J., Flores-Herr N. Tokenizer Choice For LLM Training: Negligible or Crucial?. *arXiv preprint arXiv:2310.08754*., 2023.

Bouchiha M. A., Telnoff Q., Bakkali S., Champagnat R., Rabah M., Coustaty M., Ghamri-Doudane Y. LLMChain: Blockchain-based Reputation System for Sharing and Evaluating Large

[1]*Elov Botir Boltayevich* – doctor of philosophy in technical sciences, docent. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2]*Xusainova Zilola Yuldashevna* – doctor of philosophy in philology, senior teacher. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: *xusainovazilola@navoiy-uni.uz*
ORCID: 0000-0003-4357-7515
[3]*Qosimova Sarvinoz Furqat qizi* – Master of degree. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: *sarvinozq54@gmail.com*

Language Models. *arXiv preprint arXiv:2404.13236*., 2024.

Elov B. N-gramm til modellari vositasida o‘zbek tilida matn generatsiya qilish. C*omputer linguistics: problems, solutions, prospects*, *1*(1), 2022.

Elov B. O‘zbek-ingliz tillari parallel korpusiga qo‘yiladigan lingvistik va ekstralingvistik talablar. *Computer linguistics: problems, solutions, prospects*, *1*(1), 2023.

Elov B., Alayev R., Abdullayev A., Aloyev N. Yuqori n-gram modellarini o‘zbek tili matnlariga qo‘llash. *digital transformation and artificial intelligence*, *2*(5), 2024. 152-162.

Huang J., Gu S. S., Hou L., Wu Y., Wang X., Yu H., Han J. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*., 2022.

Inan H., Upasani K., Chi J., Rungta R., Iyer K., Mao Y., Khabsa M. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*., 2023.

Naveed H., Khan A. U., Qiu S., Saqib M., Anwar S., Usman M., Mian A. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*., 2023.

Penedo G., Malartic Q., Hesslow D., Cojocaru R., Cappelli A., Alobeidli H., Launay J. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*., 2023.

Vaswani A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Polosukhin I. Attention Is All You Need.(Nips), 2017. *arXiv preprint arXiv:1706.03762*, *10*, S0140525X16001837., 2017.

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga **.**.****-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.