Modern education and development
ISSN: 3060-4567

CHATGPT VA GPT TEXNOLOGIYASI: TRANSFORMER ARXITEKTURASI, O‘QITISH BOSQICHLARI VA ILMIY-NAZARIY TAHLIL

ТЕХНОЛОГИЯ CHATGPT И GPT: АРХИТЕКТУРА TRANSFORMER, ЭТАПЫ ОБУЧЕНИЯ И НАУЧНО-ТЕОРЕТИЧЕСКИЙ АНАЛИЗ

CHATGPT AND GPT TECHNOLOGY: TRANSFORMER ARCHITECTURE, TRAINING STAGES, AND A SCIENTIFIC-THEORETICAL ANALYSIS

Navoiy davlat universiteti Aniq fanlar fakulteti
talabasi ***Ruxshona Ziyodullayeva***
Navoiy davlat universiteti Aniq fanlar fakulteti
talabasi ***Toshmurodova Gulsevar***

Annotatsiya. *Ushbu maqolada ChatGPT va GPT oilasining shakllanishi, transformer arxitekturasi, autoregressiv o‘qitish prinsipi, instruktsiyaga moslashtirish va RLHF bosqichlari ilmiy manbalar asosida tahlil qilinadi. Tadqiqotning maqsadi generativ pretrained transformer modellarining ishlash mexanizmini matematik va konseptual nuqtai nazardan yoritishdan iborat. Maqolada o‘z-o‘ziga e’tibor mexanizmi, keyingi token ehtimolini hisoblash, kross-entropiya yo‘qotish funksiyasi hamda inson fikri orqali moslashtirish modeli formulalar bilan izohlanadi. Natijada ChatGPT samaradorligi faqat model hajmiga emas, balki transformer arxitekturasi, pre-training sifati, fine-tuning, xavfsizlik va alignment mexanizmlarining uyg‘unligiga bog‘liq ekani asoslab beriladi.*

Abstract. *This article analyzes the evolution of ChatGPT and the GPT family, the transformer architecture, autoregressive training, instruction tuning, and RLHF based on primary scientific sources. The objective is to explain the working mechanism of generative pretrained transformer models from mathematical and*

conceptual perspectives. The paper describes self-attention, next-token probability estimation, cross-entropy loss, and human-feedback alignment using formulas and diagrams. The analysis shows that ChatGPT performance depends not only on model scale, but also on the interaction between transformer design, pre-training quality, fine-tuning, safety, and alignment mechanisms.

Аннотация. *В статье на основе первичных научных источников анализируются эволюция ChatGPT и семейства GPT, архитектура transformer, авторегрессионное обучение, настройка по инструкциям и этап RLHF. Цель работы — раскрыть механизм функционирования generative pretrained transformer моделей с математической и концептуальной точек зрения. В статье формулами и схемами объясняются self-attention, вычисление вероятности следующего токена, функция потерь кросс-энтропии и согласование модели с человеческой обратной связью. Показано, что эффективность ChatGPT определяется не только масштабом модели, но и согласованностью архитектуры, предобучения, fine-tuning, безопасности и alignment-механизмов.*

Kalit so‘zlar. *ChatGPT, GPT, transformer, self-attention, RLHF, pre-training, fine-tuning, token, generativ model, sun’iy intellekt.*

Keywords. *ChatGPT, GPT, transformer, self-attention, RLHF, pre-training, fine-tuning, token, generative model, artificial intelligence.*

Ключевые слова. *ChatGPT, GPT, transformer, self-attention, RLHF, pre-training, fine-tuning, токен, генеративная модель, искусственный интеллект.*

KIRISH

ChatGPT — bu dialog formatida ishlashga moslashtirilgan generativ pretrained transformer oilasiga mansub yirik til modeli bo‘lib, uning paydo bo‘lishi transformer arxitekturasi, masshtabli pre-training va keyinchalik instruction tuning hamda RLHF kabi alignment usullarining bosqichma-bosqich rivojlanishi bilan bog‘liq [1], [5], [6]. 2017-yilda transformer modeli taklif qilingach, ketma-ketliklarni qayta ishlashda rekurrent tarmoqlarga tayanmasdan e’tibor mexanizmi asosida uzoq

bog‘lanishlarni samarali o‘rganish imkoniyati yaratildi [1]. Shu ilmiy poydevor ustida GPT-1 generativ pre-training g‘oyasini, GPT-2 esa keng ko‘lamli nol-shot va ko‘p vazifali xatti-harakatni, GPT-3 esa few-shot o‘qitish salohiyatini namoyish etdi [2]–[4].

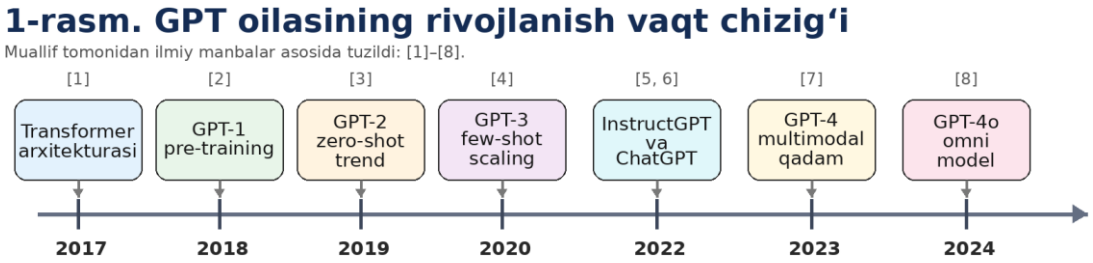
ChatGPTning alohida ilmiy va amaliy ahamiyati shundaki, u oddiy til modelidan foydalanuvchining ko‘rsatmalariga mos, dialogga tayyor tizimga aylantirildi. OpenAI ta’rifiga ko‘ra, ChatGPT InstructGPT bilan yaqin qarindosh model bo‘lib, dialog formati sababli savollarga javob berish, oldingi xatolarini tan olish, noto‘g‘ri taxminlarni rad etish va zararli so‘rovlarni cheklash imkoniyatiga ega [5], [6]. Mazkur maqolaning maqsadi GPT oilasining ilmiy asoslarini, ChatGPTning ishlash jarayonini hamda uning matematik modelini ilmiy uslubda, rasmlar va formulalar bilan izohlashdan iborat.

GPT OILASINING SHAKLLANISHI VA EVOLYUTSIYASI

GPT modeli nomidagi “Generative Pretrained Transformer” iborasi uchta tayanch g‘oyani bildiradi: model generativ xarakterga ega, ya’ni matnni tokenma-token yaratadi; u katta korpusda oldindan o‘qitiladi; va uning asosida transformer arxitekturasi yotadi [1], [2]. GPT-1 generativ pre-training yondashuvini sistemalashtirib, katta belgilanmagan korpusda til modeli sifatida o‘qitilgan parametrlarning keyinchalik aniq vazifaga fine-tuning qilinishi katta foyda berishini ko‘rsatdi [2]. GPT-2 esa masshtab oshirilganda til modelining yangi xususiyatlari paydo bo‘lishini, jumladan nol-shot va ko‘p vazifali xatti-harakat shakllanishini namoyon etdi [3].

GPT-3 modeli miqyosning ahamiyatini yanada ravshan ko‘rsatdi: ulkan parametrlar soni va kengroq ma’lumot asosida model ko‘plab vazifalarda an’anaviy fine-tuningsiz, ya’ni faqat promptdagi namunalarga tayangan holda natija bera boshladi [4]. Shundan keyin tadqiqotlar model faqat “kuchli” bo‘lishi emas, balki “foydali, xavfsiz va ko‘rsatmaga mos” bo‘lishi kerakligini ko‘rsatdi. InstructGPT ishida inson namunalari va preferensiya baholari asosida til modelini moslashtirish yondashuvi taklif etildi [5]. ChatGPT aynan shu liniyaning amaliy ko‘rinishi bo‘lib,

dialog interfeysida instruction-following sifatlarini kuchaytiradi [6]. GPT-4 va GPT-4o esa multimodal rivojlanish bosqichini boshlab, matndan tashqari tasvir va audio bilan ishlashga yo‘l ochdi [7], [8].



1-rasm. GPT oilasining rivojlanish vaqt chizig‘i

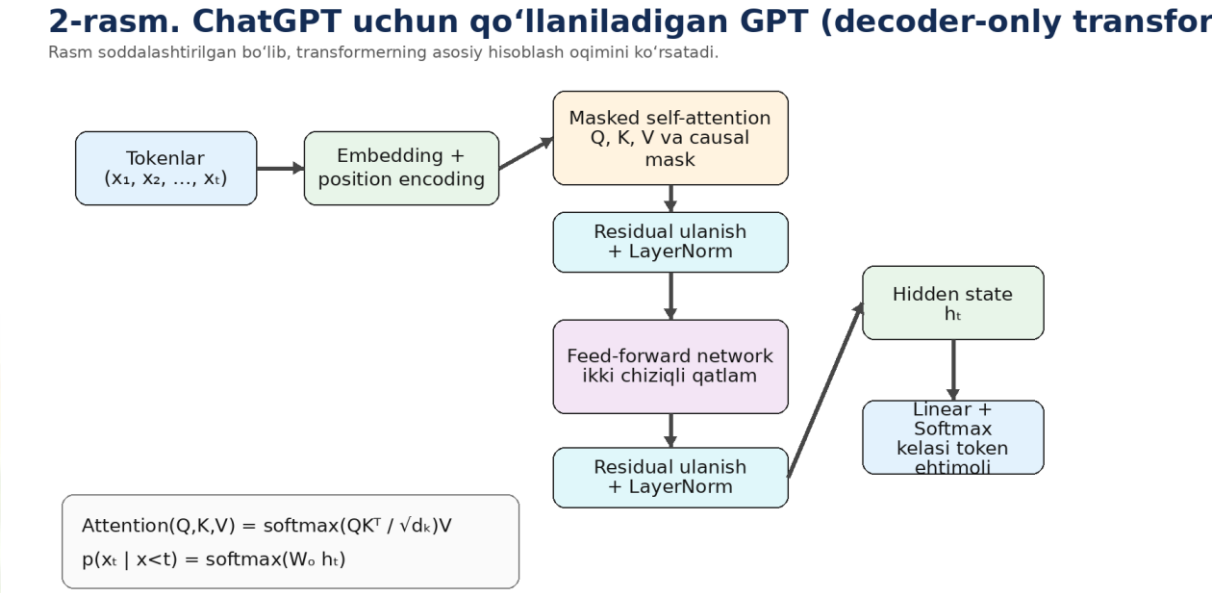
1-rasmda ko‘rinib turganidek, GPT oilasining har bir bosqichi alohida ilmiy muammoni hal qildi: transformer ketma-ketlikni parallel qayta ishlashni, GPT-1 pre-trainingni, GPT-2 noldan vazifani o‘rganish tendensiyasini, GPT-3 esa masshtabning few-shot salohiyatga ta’sirini ko‘rsatdi. InstructGPT va ChatGPT esa modelni inson intizomiga yaqinlashtirish, xavfsizlik va foydalanuvchi niyatiga moslikni oshirishga xizmat qildi [5], [6].

CHATGPTNING ISHLASH PRINSIPI: TRANSFORMER ARXITEKTURASI

GPT oilasi decoder-only transformer arxitekturasiga tayangan. Bu arxitekturada kiruvchi tokenlar embedding qatlamida vektor ko‘rinishga o‘tkaziladi va pozitsion axborot bilan boyitiladi. So‘ngra har bir blokda causal (ya’ni faqat oldingi tokenlarga qarovchi) masked self-attention hamda feed-forward tarmoq qo‘llanadi [1], [9]. Decoder-only tuzilma kelasi tokenni bashoratlash vazifasi uchun juda qulay bo‘lib, matnni chapdan o‘ngga izchil generatsiya qilishga imkon beradi.

Self-attention mexanizmi har bir tokenning boshqa tokenlar bilan bog‘lanishini o‘rganadi. Matematik jihatdan bu jarayon so‘rovlar Q, kalitlar K va qiymatlar V matritsalari yordamida ifodalanadi. E’tibor og‘irliklari tokenlar o‘rtasidagi semantik va sintaktik aloqalarni aniqlaydi, shuning uchun model uzoq

masofadagi bog‘liqliklarni ham samarali ushlaydi [1]. Keyingi bosqichda residual ulanishlar, LayerNorm va feed-forward qatlamlar hisoblashni barqarorlashtiradi hamda nolinearlik kiritadi [9], [10].



2-rasm. GPT modelining soddalashtirilgan decoder-only transformer blok sxemasi

2-rasm GPT blokining asosiy oqimini ko‘rsatadi: tokenlar embedding ko‘rinishga o‘tkaziladi, masked self-attention orqali kontekst olinadi, undan keyin feed-forward tarmoq hamda residual-layer normalization amallari bajariladi. Yakunda yashirin holat h_t vektori linear qatlam va softmax orqali ehtimollik taqsimotiga aylantiriladi. ChatGPT javob berayotganda aynan shu mexanizm takroran ishlaydi: har gal model o‘zining keyingi token ehtimolini hisoblab, eng ma’qul tokenni tanlaydi yoki namunaviy tanlash mexanizmini qo‘llaydi [1], [4], [7].

MATEMATIK MODEL VA O‘QITISH BOSQICHLARI

GPT tipidagi til modeli uchun asosiy vazifa — berilgan oldingi tokenlar ketma-ketligidan keyingi token ehtimolini hisoblash. Agar x = (x₁, x₂, …, x_T) matn ketma-ketligi bo‘lsa, model quyidagi taqsimotni o‘rganadi: pθ(x) = Π(t=1…T) pθ(x_t | x_<t). Bu yerda θ — model parametrlari, x_<t — t-pog‘onagacha bo‘lgan kontekst. Trening jarayonida modelning maqsadi to‘g‘ri tokenlarning log-ehtimolini

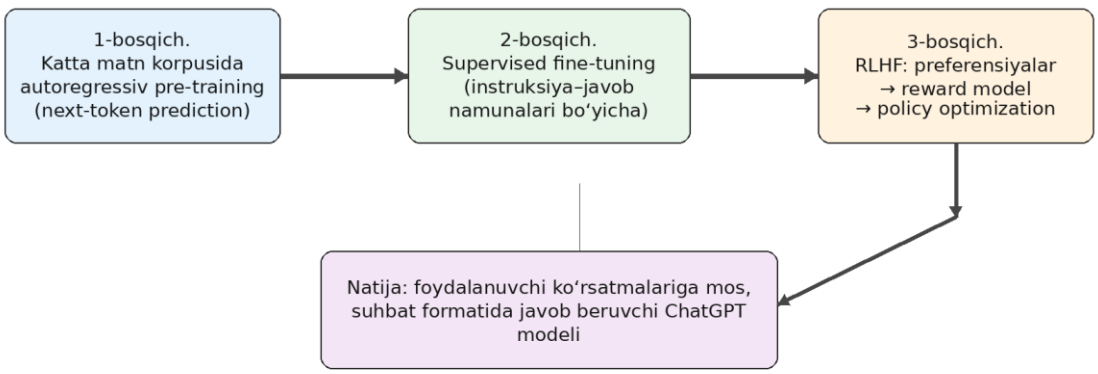
maksimal qilish, ya’ni kross-entropiya yo‘qotish funksiyasini minimallashtirishdan iborat [2], [4], [9].

Transformerning yuragi hisoblangan self-attention quyidagi formula bilan yoziladi: Attention(Q, K, V) = softmax(QK^T / √d_k) V. Ushbu ifodada QK^T tokenlar orasidagi moslikni, √d_k esa miqyoslashni, softmax esa e’tibor og‘irliklari normallashuvini ifodalaydi [1]. Til modelining odatiy yo‘qotish funksiyasi $L(\theta) = -\Sigma(t=1\ldots T) \log p\theta(x_t \mid x_{<t})$ ko‘rinishida bo‘ladi. Katta korpusda shu maqsad bilan pre-training o‘tkazilgach, model umumiy til statistikasi va dunyo haqidagi ko‘plab qonuniyatlarni o‘rganadi [4].

ChatGPTga olib keluvchi keyingi muhim bosqich — instruction tuning va RLHF. Avval modelga odamlar yozgan “savol–javob” juftliklari ko‘rsatiladi va supervised fine-tuning amalga oshiriladi [5]. Shundan so‘ng bir nechta javob variantlari bo‘yicha inson preferensiyalari yig‘ilib, reward model o‘qitiladi. Oxirgi bosqichda esa siyosat (policy) modeli reward’ni oshirish bilan birga asosiy modeldan haddan tashqari chetlashmaslikka harakat qiladi. Bu g‘oya soddalashtirilgan holda $L_{RLHF}(\varphi) = E[r(x, y)] - \beta\, KL(\pi\varphi \,||\, \pi ref)$ tarzida yozilishi mumkin; bu yerda KL had modelning mos yozuvlarni saqlab qolishini nazorat qiladi [5], [6].

3-rasm. ChatGPT modeli tayyorlanishining uch bosqichli sxemasi

Mazkur sxema InstructGPT va ChatGPT haqidagi manbalardagi RLHF pipeline’ni umumlashtiradi [5], [6].



3-rasm. ChatGPTni tayyorlashning umumlashtirilgan uch bosqichli RLHF pipelinei

3-rasmda ko‘rsatilgan pipeline ChatGPTning amaliy ustunligini tushuntiradi. Oddiy yirik til modeli ko‘p bilimga ega bo‘lishi mumkin, biroq bu bilimning foydalanuvchi niyatiga mos ifodalanishi uchun alignment kerak. InstructGPT tadqiqoti shuni ko‘rsatadiki, inson ko‘rsatmalariga moslashtirilgan kichikroq model ba’zan ancha katta, lekin alignment qilinmagan modeldan afzal bo‘lishi mumkin [5]. Demak, ChatGPTning sifati faqat parametrlar soni bilan emas, balki trening jarayoni dizayni bilan ham belgilanadi.

TAHLIL VA ILMIY-AMALIY AHAMIYAT

GPT oilasining ilmiy ahamiyati uchta asosiy natijada ko‘rinadi. Birinchidan, transformer arxitekturasi tabiiy tilni qayta ishlashda universal platformaga aylandi [1]. Ikkinchidan, katta miqyosli pre-training umumiy til kompetensiyasini shakllantirib, zero-shot va few-shot xatti-harakatni paydo qildi [3], [4]. Uchinchidan, instruction tuning va RLHF modelni foydalanuvchi uchun foydali, nisbatan xavfsiz va boshqariladigan tizimga aylantirishning amaliy yo‘lini ko‘rsatdi [5], [6].

ChatGPTning amaliy ahamiyati ta’lim, dasturlash, ma’lumotlarni tahlil qilish, hujjat bilan ishlash, tarjima, konsultatsiya va ilmiy yozuv kabi ko‘plab sohalarda namoyon bo‘lmoqda. Biroq ilmiy nuqtai nazardan uning cheklovlari ham mavjud: model ba’zan ishonchli ko‘rinadigan, lekin xato yoki uydirma javoblar berishi, trening ma’lumotlaridagi og‘ishlarni takrorlashi, kontekst uzunligi va xavfsizlik chegaralariga bog‘liq bo‘lishi mumkin [6]–[8]. Shuning uchun GPT modellarini qo‘llashda verifikatsiya, tashqi manbalar bilan tekshirish va muammo kontekstiga mos boshqaruv muhimdir.

1-jadval. GPT oilasi bosqichlarining ilmiy qiyosi

Model	Yil	Asosiy g‘oya	Arxitektura / trening	Ilmiy ahamiyati
Transformer	2017	Self-attention asosida ketma-	Encoder–decoder transformer	GPT oilasi uchun nazariy poydevor [1]

Model	Yil	Asosiy g‘oya	Arxitektura / trening	Ilmiy ahamiyati
		ketlikni parallel qayta ishlash		
GPT-1	2018	Generativ pre-training va keyingi fine-tuning	Decoder-only LM	Umumiy pre-training paradigmasini ko‘rsatdi [2]
GPT-2	2019	Zero-shot va ko‘p vazifali xatti-harakat	Katta hajmli decoder-only LM	Masshtab yangi xususiyatlar berishini ko‘rsatdi [3]
GPT-3	2020	Few-shot learning va promptga sezgirlik	175B parametrli autoregressiv LM	Katta til modellarining imkoniyatini ochdi [4]
InstructGPT / ChatGPT	2022	Instruction tuning va RLHF	SFT + reward model + policy optimization	Foydalilik va alignmentni kuchaytirdi [5], [6]
GPT-4 / GPT-4o	2023–2024	Multimodal rivojlanish va xavfsizlik tizimlari	Text + image, keyin omni yo‘nalish	Yangi multimodal bosqichni boshladi [7], [8]

1-jadvaldan ko‘rinadiki, GPT oilasining rivojlanishi faqat parametrlar ko‘payishi emas; har bir bosqichda yangi ilmiy g‘oya paydo bo‘lgan. Transformer masalani parallel e’tibor mexanizmi bilan yechgan bo‘lsa, GPT-1 pre-trainingni umumlashtirdi, GPT-2 va GPT-3 masshtabning ta’sirini ko‘rsatdi, InstructGPT va

ChatGPT esa modelni inson maqsadlariga yaqinlashtirish muammosini hal qildi [1]–[6].

XULOSA

Tahlil natijalariga ko‘ra, ChatGPT — bu alohida bir “sehrli” model emas, balki transformer arxitekturasi, keng ko‘lamli generativ pre-training, supervised instruction tuning va RLHF kabi bir necha ilmiy yondashuvlarning integratsiyasi mahsulidir. Uning matematik asosi self-attention, ehtimollik taqsimoti va kross-entropiya yo‘qotishiga tayanadi; amaliy samaradorligi esa alignment, xavfsizlik va foydalanuvchi niyatiga moslik bilan kuchayadi [1], [5]–[8].

Kelajak tadqiqotlar uchun ikki yo‘nalish alohida ahamiyatga ega: birinchisi — generativ modellarni yanada ishonchli, tushuntiriladigan va tekshiriladigan qilish; ikkinchisi — multimodal GPT tizimlarini ta’lim, ilm-fan va sanoatdagi maxsus vazifalarga chuqurroq moslashtirish. Shu ma’noda ChatGPT va GPT oilasining o‘rganilishi nafaqat sun’iy intellekt amaliyoti, balki zamonaviy matematik modellashtirish va hisoblash nazariyasi uchun ham muhimdir.

FOYDALANILGAN ADABIYOTLAR

1. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017.

2. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving Language Understanding by Generative Pre-Training. OpenAI, 2018.

3. Radford A., Wu J., Child R. et al. Language Models are Unsupervised Multitask Learners. OpenAI, 2019.

4. Brown T. B., Mann B., Ryder N. et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020.

5. Ouyang L., Wu J., Jiang X. et al. Training Language Models to Follow Instructions with Human Feedback // Advances in Neural Information Processing Systems. 2022.

6. OpenAI. Introducing ChatGPT. OpenAI, 2022.

7. OpenAI. GPT-4 Technical Report. OpenAI, 2023.

8. OpenAI. GPT-4o System Card. OpenAI, 2024.

9. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.

10. LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. 2015. Vol. 521. P. 436–444.