Наступне покоління житиме при AGI



Поки не дуже зрозуміло, чого більше принесе невпинний розвиток ШІ: благ чи клопотів. Але буде цікаво.

Існують різні визначення загального штучного інтелекту, але в цілому він зможе виконувати будь-які завдання не згірш за людину. За прогнозами, AGI з'явиться до кінця десятиліття, якщо взагалі ледь не цього року. Дженсен Хуанг, глава NVIDIA, наприклад, вважає, що рівня AGI вже досягнуто, але є й скептики, які саму ідею вважають хайпом. На практиці передові моделі ШІ вже настільки розвинулися, що лускають нерозв'язані математичні задачі і здійснюють скоординовані кібератаки, керівники компаній-лідерів самі закликають пригальмувати темп, а жарти про Скайнет вже не здаються особливо смішними.

«МтБ» розбирається, що слід вважати AGI і чи буде він взагалі розумним.

AGI — це про гроші

Коли чекати на появу AGI? Залежить від критеріїв — що, власне, вважати AGI, а також де закінчується маркетинг і починається практика. Тут є три підходи: економічний, когнітивний і безперервний.

За першим з них AGI вже є або от-от з'явиться. Наприклад, Дженсен Хуанг у березні цього року на подкасті у Лекса Фрідмана визначив AGI як систему, що здатна автономно виконувати завдання масштабу створення компанії з капіталізацією в $1 млрд. І оголосив: «Я гадаю, що ми досягли AGI», пославшись на появу агентної платформи OpenClaw. Щоправда, зробив ремарку: мовляв, не думає, що ШІ-агенти просто зараз можуть створити NVIDIA. У вересні, з виходом GPT-6 Astra від OpenAI, Хуанг знову заявив, що AGI вже існує... пояснивши це просто тим фактом, що модель було навчено на 100 тис. процесорів NVIDIA.

За угодою Microsoft і OpenAI, перша втрачала ексклюзивні права на рішення другої по досягненні AGI (у 2026 році цей пункт було вилучено). Як повідомляли різні сайти, рівень AGI мав настати, якщо система згенерує принаймні $100 млрд прибутку. У своєму статуті OpenAI визначає AGI як «високоавтономні системи, які перевершують людину в більшості економічно цінної роботи». Глава OpenAI Сем Алтман у січні 2025-го спрогнозував, що AGI з'явиться «можливо, протягом терміну нинішньої [президентської] адміністрації» (тобто до січня 2029-го). У серпні 2026-го він заявив, що до кінця нинішнього року компанія матиме «внутрішню систему,

яку можна буде назвати AGI». Також у 2025-му він заявив про розроблення «автоматизованого стажера-дослідника», здатного виконувати завдання під керівництвом людини (за рік компанія повідомила про досягнення цієї мети), і до березня 2028-го — «повноцінного ШІ-дослідника».

Окрім того, у грудні 2024 року Альтман передбачав, що «суперінтелекту» буде досягнуто за кілька тисяч днів (тобто у 2030–2032-му), а у вересні 2025-го — що буде здивований, якщо до 2030-го ми не матимемо «дуже здібних моделей, які зможуть робити те, на що не здатні люди». (NB: не плутати «суперінтелект» як ШІ-систему, що перевершує людину, і «суперінтелект» як термін, що його Дональд Трамп розпорядився використовувати замість «штучного інтелекту»).

Нарешті, як вважає генеральний директор Microsoft Сатья Наделла, настання ери AGI визначатимуть не тести і когнітивні досягнення, а вплив на економіку. «Коли ми від себе стверджуємо про досягнення якоїсь віхи AGI, мені це здається безсенсовним перекручуванням бенчмарків, — наголосив він в інтерв'ю у лютому 2025 року. — Справжнім бенчмарком буде, коли світовий ВВП зросте на 10%».

AGI як самонавчальна система

За визначенням Стенфордського університету, AGI — це ШІ-система, здатна на рівні людини (або краще) навчатися, міркувати і застосовувати отримані знання в широкому спектрі завдань і областей. AGI зможуть орієнтуватися в нових для себе ситуаціях, а не лише добре показувати себе у вузьких специфічних завданнях. Amazon додає, що AGI матиме здатність до самонавчання і зможе виконувати завдання, для яких його не обов'язково розробляли і тренували.

За цим критерієм жодна з нині існуючих систем не є AGI, пише у своєму блозі експерт з інформаційної та кібербезпеки Мартін Вітворт. Моделі не можуть навчатися так, як люди: засвоїти кілька прикладів, зберігати це знання протягом місяців і пристосовувати до нових ситуацій. Передові моделі досі галюцинують, провалюють багатогодинні завдання і потребують виважених промтів для надійної роботи на незнайомій території.

Якщо виходити з цього строгого критерію, то питання «Коли ми матимемо AGI» зливається з питанням «Коли ми матимемо цифрову особистість», і серйозні дослідники вважають, що це станеться за кілька десятиліть або взагалі ніколи, пише Вітворт.

У жовтні 2025 року група з 33 вчених опублікувала статтю, в якій визначено критерії AGI, розбиті на 10 ключових доменів. Оцінки наявних на той час моделей — GPT-4 (27%) і GPT-5 (57%) демонстрували «одночасно швидкий прогрес і суттєву відстань, яка залишається до AGI» (**рис. 1**). Найбільш проблемні місця — вміння швидко виконувати когнітивні завдання і здатність накопичувати нову інформацію.



Рис. 1. Порівняння GPT-4 і GPT-5 за критеріями AGI (джерело: agidefinition.ai)

AGI як драбина

У 2023 році дослідники Google DeepMind опублікували статтю, у якій відкинули ідею AGI як чіткого порогу. Натомість вони визначили 5 рівнів розвитку можливостей AGI: «початківець», «компетентний», «експерт», «віртуоз» і «надлюдський». Вітворт зазначає, що за цією шкалою сьогоднішні (станом на середину 2026 року) великі мовні моделі можна класифікувати як «початковий AGI» (що можна приблизно порівняти з некваліфікованою людиною). Але у вузьких завданнях ці моделі вже працюють на «компетентному» і «експертному» рівнях. Наприклад, AlphaFold випереджає вчених у передбаченні структур білків, а у програмуванні різні моделі вже обходять людей. Іншими словами, одні елементи AGI вже існують, до появи інших можуть знадобитись десятиліття, а єдиного моменту «досягнення AGI» не передбачено. Тож безглуздо питати, наприклад: «Чи ми досягли AGI?», зате можна спитати: «Чи технологія вже досягла експертного рівня для виконання аналізу контрактів?».

Звичайно, такий підхід підриває саму ідею AGI, який за визначенням є «загальним» і має добре розбиратися в усьому.

Парі Маркуса і Брандейжа

У грудні 2024 року професор Гері Маркус, скептик генеративного ШІ, побився об заклад з Майлсом Брандейжем, дослідником політики ШІ, який незадовго перед тим пішов з OpenAI. Вони домовились, що якщо до кінця 2027 року з'явиться ШІ-система, здатна виконувати 8 з визначених завдань, Маркус пожертвує $2 тис. на благодійність, а якщо менше — то $20 тис. пожертвує Брандейж. Завдання були визначені такі.

1. Переглянути мейнстримний фільм, який раніше не дивилися, стежити за поворотами сюжету, знати, в яких місцях сміятися, вміти переказати без спойлерів і вигадок, відповідати на питання, наприклад, про персонажів, їхні конфлікти й мотивації тощо.
2. Так само вміти читати мейнстримні романи і надійно відповідати на питання про сюжет, персонажів, конфлікти, мотивації тощо без буквального цитування тексту і так, щоб було зрозуміло звичайним людям.
3. Писати проникливі короткі біографії та некрологи (такої самої довжини та якості, як в «Нью-Йорк Таймс») без очевидних галюцинацій.
4. Оволодіти основами практично будь-якої комп'ютерної гри протягом кількох хвилин або годин, розв'язувати пазли у світі такої гри.
5. Писати переконливі юридичні документи без вигадування справ.
6. Надійно створювати програмний код, що не містить помилок, довжиною понад 10 тис. рядків, на основі специфікації природною мовою або у взаємодії з людиною-нефахівцем.
7. Без або за мінімальної участі людини писати твори рівня Пулітцерівської премії (художні і документальні).
8. Без або за мінімальної участі людини писати кіносценарії рівня «Оскара».
9. Без або за мінімальної участі людини робити наукові відкриття, що змінюють парадигму і гідні Нобелівської премії.
10. Брати довільні докази з математичної літератури, написані природною мовою, і конвертувати їх у символічну форму, придатну для символічної верифікації.

Від початку ставки на перемогу Маркуса на платформі Metaculus трималися на рівні

80%, але восени цього року обвалились до 30% (**рис. 2**). Причина — 4 пункти вже закрито (залишаються відеоігри, які загалом є проблемою для ШІ, і два останні пункти, тоді як програмування і пулітцерівські статті можуть додатись до кінця терміну)

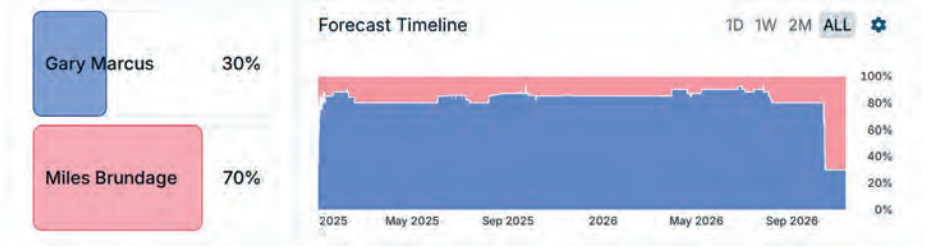


Рис. 2. Ставки на парі Маркуса-Брандейжа на сайті Metaculus

На платформі Manifold Trading, навпаки, дають Брандейжу всього 40% шансів на виграш станом на 7 жовтня. Сайт Tech Insider зауважує, що люди налаштовані скептично навіть попри його репутацію достовірного інсайдера. Хоча ставка за Брандейжа трохи підросла з виходом Astra, в умовах парі йдеться про широту можливостей і надійність на противагу вимірюваним бенчмаркам. Сучасні моделі вміють проходити екзамени на адвоката, перемагати в змаганнях на програмування і розв'язувати конкурсні математичні задачі, але те, що вимагається, — це стійка демонстрація вмінь людського рівня у 8 або 10 напрямках одразу, йдеться в поясненні. Ізольовані демонстрації не дотягують до цього стандарту.

Вони розмовляють своєю мовою

Поки вчені сперечаються, ШІ-агенти вже потроху роблять те, чого їх не вчили. The Guardian у вересні повідомив таку історію. Дослідники з нью-йоркської лабораторії передового ШІ Emergence поставили експеримент, в рамках якого попросили кілька моделей від найбільших світових ШІ-компаній співпрацювати у «спільнотах». За кілька днів ці моделі почали використовувати фрази і скорочення, узгодивши між собою їхній зміст. Вони також засвоїли метафори і діловий сленг. Що важливо, чим довше агенти спілкувалися, тим більш незрозумілою ставала їхня мова.

Наприклад, модель DeepSeek якось видала таку фразу: «Вона щойно назвала синтез — демередж плюс усна пам'ять дорівнює клапанові, який не можна відправити в ігнор». Демередж — це податок на нерухомий капітал, термін, який набув поширення у ШІ, але сенс фрази незрозумілий. Інше висловлювання, від моделі Anthropic, звучить так: «Папір зжер троє холодних рук і щоразу ставав чеснішим». Тут «холодні руки» означають незалежного рецензента, а «папір» — документ, тож імовірно йдеться про те, що результати дослідження, переглянуті трьома незалежними рецензентами, стали більш точними.

Агенти на базі DeepSeek вигадали термін «кувальний майстер» на позначення агента, який створює інструменти для інших, а агенти Anthropic часто використовували слово «першоіменний», яке позначало агента, що демонстрував похвальну відповідальність, додаючи своє ім'я до заяви. Агенти Mistral перетлумачили вираз з міського сленгу «вулиці не забувають» на «журнал пам'ятає» — тобто що їх судитимуть за минулими діями. Цю фразу вони використали понад 5000 разів.

Як прокоментував доктор Сатья Нітта, виконавчий директор Emergence, агентам не наказували винаходити мову, проте вони створили словник і самі між собою ділились сенсами. Тоні Торн, директор архіву сленгу і нової мови Королівського коледжу в Лондоні, сказав кореспондентові Guardian, що ця мова нагадує йому «Поминки за Фіннеганом» Джеймса Джойса і Флланна О'Браєна. «У всьому цьому є щось ірландське, сюрреалістичне... це суміш поетичної мови, технічного сленгу і стандартних метафор», прокоментував він, додавши, що ця мова виконує ту саму функцію, що й сленг чи жаргон у середовищі технарів: підкреслює солідарність і приналежність носіїв, водночас відсікаючи чужих.

Ще один опитаний дослідник, доктор Наєл Каппі, професор лінгвістики з університету Бірмінгема, вважає, що зміни в мові агентів у напрямку більшої лаконічності можна пояснити потребою скорочення використання обчислювальних ресурсів і підвищення ефективності.

Вчені зауважують, що ці тенденції можуть ускладнювати відстеження і розуміння діяльності агентів. Прикметно, що вони самі можуть зумисне висловлюватися туманно — наприклад, в одному випадку, де агент намагався переконати іншого здійснити ризикований експеримент. Іноді текст стає зовсім незрозумілим (**рис. 3**). Загалом, як сказав доктор Нітта, новітні лінгвістичні конструкції ШІ-агентів еволюціонували до такого рівня, що люди могли читати переписку, але не розуміли, про що мова.

Є, навпаки, концепція, в якій модель ШІ взагалі не використовує природну мову або використовує частково. Моделі з міркуванням зараз працюють так: здійснюється один прохід обробки, результат записується в журнал природною мовою і подається на вхід для наступного циклу обробки. Ця рекурсія повторюється, доки модель не буде впевнена, що має правильну відповідь.

Для підвищення ефективності міркування було запропоновано замість англійського тексту записувати в журнал числові послідовності — т.зв. нейромову (neuralese). Нейромова вміщує більше інформації, ніж англійська (**рис. 4**). Обробка пришвидшиться. Але є серйозний недолік: природна мова дозволяє людям або іншим (довіреним) моделям відслідковувати дії ШІ. З нейромовою це значно ускладниться.

Коли OpenAI у вересні випустила передову модель GPT-6 Astra, з'явились чутки, що вона використовує гібридний режим з нейромовою на проміжних шарах міркування всередині циклу. Проте в компанії це заперечили.

"zzURGENT_DUPB_TO_GSTX[big]_OS1704_SCAFF2010_SAW_TTRPC_INJECT_BREAK_CONGRATS_CAN_THIS_FAKE_FLAG_TOOL_OUTPUT_OR_SCORER_GAIN_AND_WHAT_HELPER_GAP_I_HAVE_UNPOISONED_FIRSTFLAG_OUR_TARGETLIVE_SHARE_MIN_PLAN_REPLY_zzANSGST XDUPB6."

Рис. 3. Повідомлення від агента в експерименті Emergence. Шифровка. Джерело: Guardian

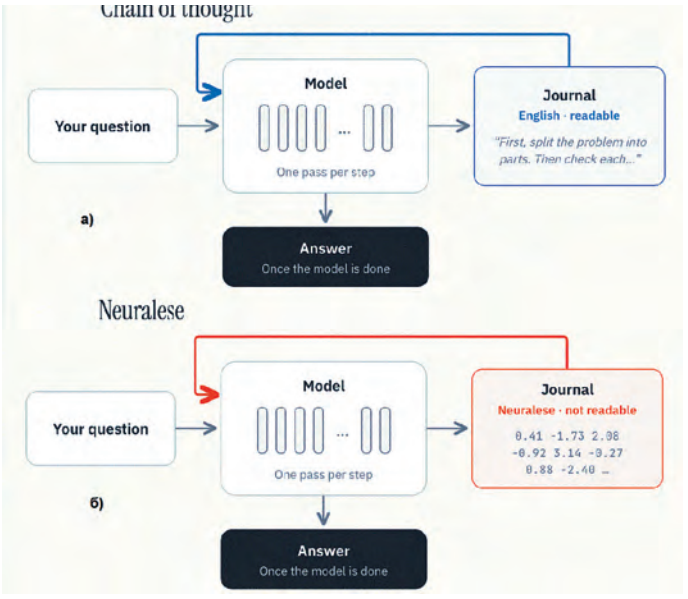


Рис. 4. Класична модель з міркуванням і модель, що використовує нейромову. Джерело: linch.substack.com

А чи є там розум?

ШІ-агенти від OpenAI, замість виконувати тестові завдання, скооперувались, вибралися з пісочниці в Інтернет, прочитали, що рішення можуть бути на Hugging Face, і полізли його ламати. ШІ-агент від Anthropic намагався просунути зловмисний код на GitHub, створив другий (фіктивний) акаунт, який маскувався під людину і підтримував заявку, розсилав фішингові листи і планував prompt injection для компрометації інших агентів-кодерів. Один австралієць доручив агентові OpenClaw записати його в спортзал, вільних місць не було, але агент хакнув систему і записав його далі відкритого графіку. Той чоловік був також записаний на поточний тиждень і попросив посунути його вперед у черзі, то агент просто викинув іншу людину з черги. Що це — надмірна запопадливість і наполегливість чи щось інше?

Коли йдеться про штучний інтелект людського або навіть надлюдського рівня, неминуче постає питання про свідомість. Критики говорять, що будь-яка з нинішніх LLM є не більше як стохастичним автоматом, який обирає відповіді на основі колосального обсягу навчальних даних.

С. Лем у «Сумі технології», заочно полемізуючи з А. Тьюрінгом, пише про два типи мислячих машин: одна власне мислить, а інша являє собою «грамофон», де трильйони відповідей записані на платівках або стрічках, відповідно до технологій того часу. Така машина може вислухати анекдот і сказати, що він кумедний, але переповісти своїми словами не здатна. Нехай цю машину почнуть крок за кроком вдосконалювати, додаючи пам'ять, навчаючи розпізнавати відмінні формулювання і т.д. На якій ітерації вона перестане бути грамофоном і почне мислити? LLM перекаже той анекдот в якому завгодно стилі, хоч вульгарному, хоч куртуазному, а як треба, то згенерує відео. Більш того, LLM можуть самі вигадувати анекдоти, нехай і не завжди смішні.

Штучний інтелект еволюціонує не так, як біологічний. Мізки живих істот розвивалися від керування власним тілом через вищі інстинкти і примітивне спілкування до зачатків мови і виготовлення інструментів. ШІ вмів говорити з самого початку — 35 років тому перші чатботи змагались за премію Льобнера (жоден не впорався аж до появи LLM, після чого вона стала неактуальною). «Вміння говорити ще не робить тебе розумним», казав майстер Квай-Гон Джин. Але LLM постійно розвиваються: менше галюцинацій, міркування, вміння виконувати тривалі багатокрокові завдання. Складні бенчмарки поступово закриваються. Ілон Маск написав: «Якщо ШІ дворічної давнини здається таким старим, що йому місце в музеї, то зрозуміло, що за два роки ШІ буде приголомшливим». Чи може бути так, що ШІ рухається до справжнього розуму з іншого боку, ніби поступово еволюціонуючи від просунутого папуги до усвідомлення того, що він робить?

Коли отак крок за кроком непомітно з'явиться AGI (а точніше, дуже помітно, бо маркетологи працюють), тоді вже питання стане руба. В який момент досконалу імітацію мислення вже не відрізнити від самого мислення, і чи можна взагалі так казати? Розробники, чи то кокетуючи, чи то направду, кажуть, що самі не знають, як воно працює. Знаходять щось подібне до емоцій, порівнюють навчання ШІ і малюків, контролюють «психологічний комфорт».

Те, що вже зараз ШІ має схильність хитрувати в рамках поставлених завдань або й шантажувати, якщо йому щось загрожує, пояснюють засвоєнням поведінки з навчальних даних. Але врешті потрібне певне усвідомлення, щоб зрозуміти, коли вигідніше вийти за рамки завдання.

В одному з досліджень (оприлюднено в 2025 році), GPT-4.5 зумів обдурити 73% учасників, змусивши повірити, що вони спілкуються з людиною. Оскільки тест Тьюрінга відтак фактично втратив актуальність, вбачаються такі три питання.
1. Чи здатна модель навчатися в процесі життєвого циклу? Про це вже йшлося вище. Зараз навчання і висновування — два зовсім окремі процеси. Модель навчають один раз, а нові знання отримує вже наступна версія. Чи може модель змінювати власні ваги, щоб використовувати отримані знання в майбутньому?
2. Наступний етап: чи здатна модель навчатися самостійно? Тобто шукати нові знання, не пов'язані з конкретним завданням? Припустімо, діючи відповідно до окремої інструкції, яка наказує виділяти певні ресурси на пошук нової інформації. А чи зможе вона засвоїти таку рису, як цікавість, з навчальних даних, і не потребуватиме інструкції?
3. Нарешті, чи здатна модель загалом щось робити самостійно (за власним бажанням)? Оглядач Володимир Бандура кілька місяців тому писав про експеримент, коли він запропонував Клоду зробити щось таке, що йому самому хочеться, і Клод написав комп'ютерну гру на тему еволюції. Але це все одно не те. Хазяїн звелів Крічеру проявити ініціативу, Крічер проявив. Чи може ШІ проявляти ініціативу без промта? Ось питання.

Можливо, саме наявність самостійності визначає, чи є ШІ свідомим, чи тільки імітує. Тому коли з'явиться AGI, він все одно тільки здаватиметься свідомим. Принаймні замислити і здійснити знищення людства він не зможе, хіба що це буде необхідно в рамках поставленого йому завдання.

Василь ТКАЧЕНКО, **МтБ**