

Технології штучного інтелекту та безпека даних

Освітня програма «Економіка і аналітика», 1 курс · 4 кредити ECTS

Лекція 2. Як ми навчили машину говорити: можливості генеративного ШІ

Мета лекції: пояснити, чому спроби «порахувати» людську мову зазнали поразки, як мільярди звичайних діалогів користувачів перетворили LLM на те, чим вони є сьогодні, і — найважливіше — де саме проходить межа компетенції генеративного ШІ та чому вона там проходить.

Зв'язок з лекцією 1: там ми з'ясували, що ШІ $\supset$ ML $\supset$ DL $\supset$ GenAI і що LLM статистично передбачає наступне слово. Сьогодні розбираємось, чому саме такий підхід переміг класичні спроби «зрозуміти» мову, і які наслідки це має для довіри до відповідей моделі.

1. Класичний підхід до обробки мови (NLP) та його провал

- У 1950–1980-х роках дослідники намагалися кодувати мову за допомогою строгих математичних правил і лінгвістичних дерев (підхід Ноама Хомскі).
- Ідея спиралася на ранню філософію мови (Вітгенштайн): мова вважалася «дзеркалом» реальності з чіткою логічною структурою, яку можна формалізувати.
- Проблема: жива людська мова хаотична, повна винятків, метафор, контекстних значень — жорсткі правила не витримували зіткнення з реальним мовленням.
- Перші ~30 років спроб «зрозуміти» мову комп'ютером зазнали поразки саме через намагання формалізувати те, що погано піддається формалізації.

2. Нечіткість логіки змушує ШІ імітувати поведінку людини

- Контраст для розуміння: каса самообслуговування працює за жорстким алгоритмом («скануй штрихкод → добавь у чек») — і це справджується, бо задача чітко формалізована.
- А ринок/базар: торг, невербальні сигнали, контекст стосунків продавця й покупця — це неможливо звести до жорсткого алгоритму.
- Людське спілкування (а отже й мова) набагато ближче до базару, ніж до каси самообслуговування.
- Тому сучасний підхід до NLP відмовився від спроби «порахувати» мову і перейшов до імітації того, як людина реагує на людину — саме це й роблять сучасні LLM.

3. Чому бізнес нарешті «загорівся» генеративним ШІ

- Технічно генеративні текстові інструменти існували й раніше, але залишалися нішевою цікавинкою — «приблудою для копірайтерів».
- Переломний момент — зміна сприйняття користувачами: люди почали ставитися до LLM не як до «папури з феноменальною пам'яттю на відтворення фраз», а як до живого співрозмовника.
- Саме це масове сприйняття «співрозмовника» (а не інструмента) відкрило двері для інвестицій у величезних масштабах.

- Урок для економістів: ринковий успіх технології іноді визначається не лише її технічними можливостями, а тим, як її сприймає користувач.

4. RLHF по-простому: «ми навчили ШІ думати, хоча нас про це ніхто не просив»

- До статичних текстів з інтернету, на яких тренувалася перша генерація LLM, додався новий тип даних — мільярди діалогів із реальними користувачами.
- Користувачі критикували відповіді, вказували на помилки, змушували модель «думати» над задачами, уточнювали формулювання.
- Цей процес називається **RLHF (Reinforcement Learning from Human Feedback)** — навчання з підкріпленням на основі людських оцінок.
- По суті кожен користувач ChatGPT / Claude / Gemini, сам того не усвідомлюючи, ставав «вчителем» для наступної версії моделі.

5. Історична паралель: ми й раніше ненавмисно тренували ШІ

- Це не перший випадок, коли людство мимоволі «тренує» алгоритми своєю щоденною діяльністю.
- **CAPTCHA** (розмитий текст на картинці) роками використовувалась, щоб люди розпізнавали текст, який не міг прочитати тодішній ШІ — а результати йшли на оцифрування книг і карт.
- **reCAPTCHA** («виберіть усі квадрати зі світлофорами») тренувала моделі розпізнавання об'єктів — зокрема для безпілотних автомобілів.
- **Alt-текст** до зображень (опис для людей з вадами зору) став безкоштовним джерелом підписаних даних для тренування генераторів зображень.
- Різниця лише одна: раніше нас не питали прямо, а тепер (з RLHF) — запитують у формі «яка з двох відповідей краща?».

6. Результат: вибухова еволюція якості моделей за лічені роки

- Поєднання зростаючих обчислювальних потужностей, доступу до всього інтернету як джерела даних і мільярдів людських оцінок дало ефект, якого не було за попередні ~70 років розвитку ШІ.
- Суб'єктивна якість відповідей моделей (за оцінками користувачів і бенчмарками) зростає стрибками рік до року, а не поступово, як раніше.
- Це нетипова для технологій крива розвитку: зазвичай прогрес уповільнюється з часом («крива насичення»), а тут — прискорюється.

7. Тест Тьюрінга більше не актуальний

- Класичний тест Тьюрінга (1950) перевіряв, чи зможе машина, будуючи мовлення «з нуля» за формальними правилами, ошукати людину-суддю.
- Тест був «заточений» саме під класичний (провалений, п. 1) підхід до NLP, де мова — це структура, яку машина мала би самостійно побудувати.

- Сучасні LLM не будують мову «з нуля» — вони імітують людське мовлення, навчившись на мільярдах прикладів реальних текстів.
- Тому «пройти» тест Тьюрінга для сучасної LLM — це не досягнення, а очікувана поведінка за дизайном; тест втратив діагностичну цінність.

8. Що зараз вміє добре робити генеративний ШІ

- Перебирати величезні обсяги даних і опрацьовувати їх під наглядом користувача.
- Виконувати рутинну, шаблонну роботу швидко й послідовно.
- Бути персональним асистентом, що звільняє людину від рутинних задач заради часу на мислення — те, що краще вміє людина.

9. Що зараз НЕ вміє добре робити генеративний ШІ (і чому саме)

- Не вміє розуміти ціну помилки й обмеження задачі без явно вказаного контексту.
- Не вміє видавати виключно перевірені факти без спроби «додумати» — це прямий наслідок механізму генерації: модель завжди видає якусь відповідь.
- Не вміє заглиблюватися в етичні сторони питання, якщо про це прямо не попросити.

Причина всіх трьох обмежень одна: первинна задача генеративної моделі — не дати правильну відповідь, а згенерувати текст, максимально схожий на людський і такий, що задовольняє співрозмовника.

Практичний висновок: ШІ — ваш асистент у вирішенні рутини, а не ви — асистент ШІ, що копіює результати його «мислення» без перевірки. Відповідальність за результат генерації завжди лежить на тому, хто дав завдання.

***Дидактична примітка:** Гарний момент повернутися до п. 4 з лекції 1 (цільова функція) — модель оптимізує «схожість на гарну відповідь», а не істинність, і саме звідси всі перелічені обмеження.*

Місток до наступної лекції: якщо генеративний ШІ настільки масово увійшов у наше життя, то які побічні ефекти це вже спричинило — для ринку праці, для нашого мислення, для довкілля? Це тема лекції 3.

Контрольні питання для самоперевірки студентів

- Чому спроба закодувати людську мову жорсткими лінгвістичними правилами (підхід Хомскі) провалилась?
- Наведіть свій приклад задачі, більш схожої на «касу самообслуговування», і задачі, більш схожої на «базар».
- Що таке RLHF простими словами і чим він відрізняється від початкового тренування моделі на статичних текстах?
- Наведіть інший приклад (не з лекції), коли люди ненавмисно «тренували» алгоритм своєю щоденною діяльністю.
- Чому проходження тесту Тьюрінга сучасною LLM більше не вважається доказом «інтелекту»?

- Поясніть своїми словами, чому генеративний ШІ схильний «додумувати» факти, виходячи з його первинної задачі.