Generative AI: Large Language Models

Як працюють великі мовні моделі *

Serhii Zabolotnii
zabolotniua@gmail.com

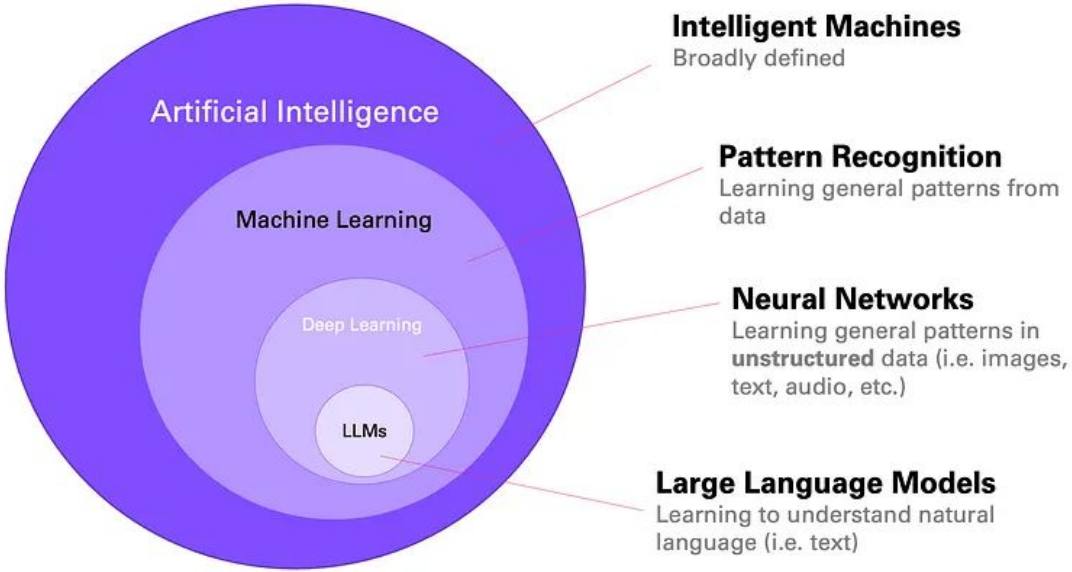
** засновано на матеріалах статті "How Large Language Models Work: From zero to ChatGPT" by Andreas Stöffelbauer (Medium)*

Вступ

1. **Важливість розуміння LLMs:**
   - 1.1. Штучний інтелект та його роль у сучасному світі.
   - 1.2. LLMs як ключова складова AI-революції.
2. **Еволюція мовної взаємодії з машинами:**
   - 2.1. Від простих команд до складної бесіди.
   - 2.2. LLMs перетворюють спілкування з машинами на природний процес.
3. **Практичне застосування LLMs:**
   - 3.1. Вплив на освіту, бізнес, розваги та інші сфери.
   - 3.2. Розширення можливостей професійної та особистої діяльності.
4. **Мета доповіді:**
   - 4.1. Глибоке розуміння механізмів роботи LLMs.
   - 4.2. Огляд тренування та використання LLMs у різних контекстах.

Ієрархія Штучного Інтелекту

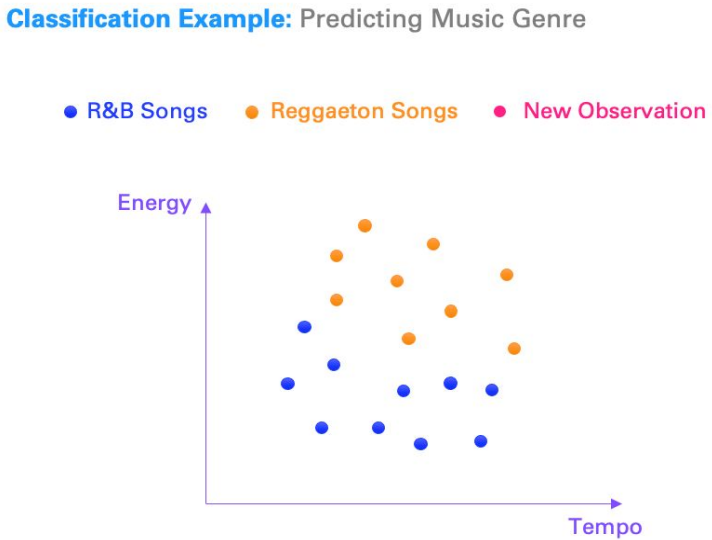
- Штучний інтелект (AI) — це дуже широкий термін, але загалом він стосується інтелектуальних машин.
- Машинне навчання (ML) — це підгалузь штучного інтелекту, яка спеціально фокусується на розпізнаванні шаблонів у даних. Як ви можете собі уявити, як тільки ви розпізнаєте шаблон, ви можете застосувати його до нових спостережень.
- Глибинне навчання – це сфера в ML, яка зосереджена на неструктурованих даних, які включають текст і зображення. Він покладається на штучні нейронні мережі, метод, який (незначно) натхненний людським мозком.
- Великі мовні моделі (LLM) стосуються конкретно тексту, і це буде в центрі уваги цієї доповіді.



Основи машинного навчання

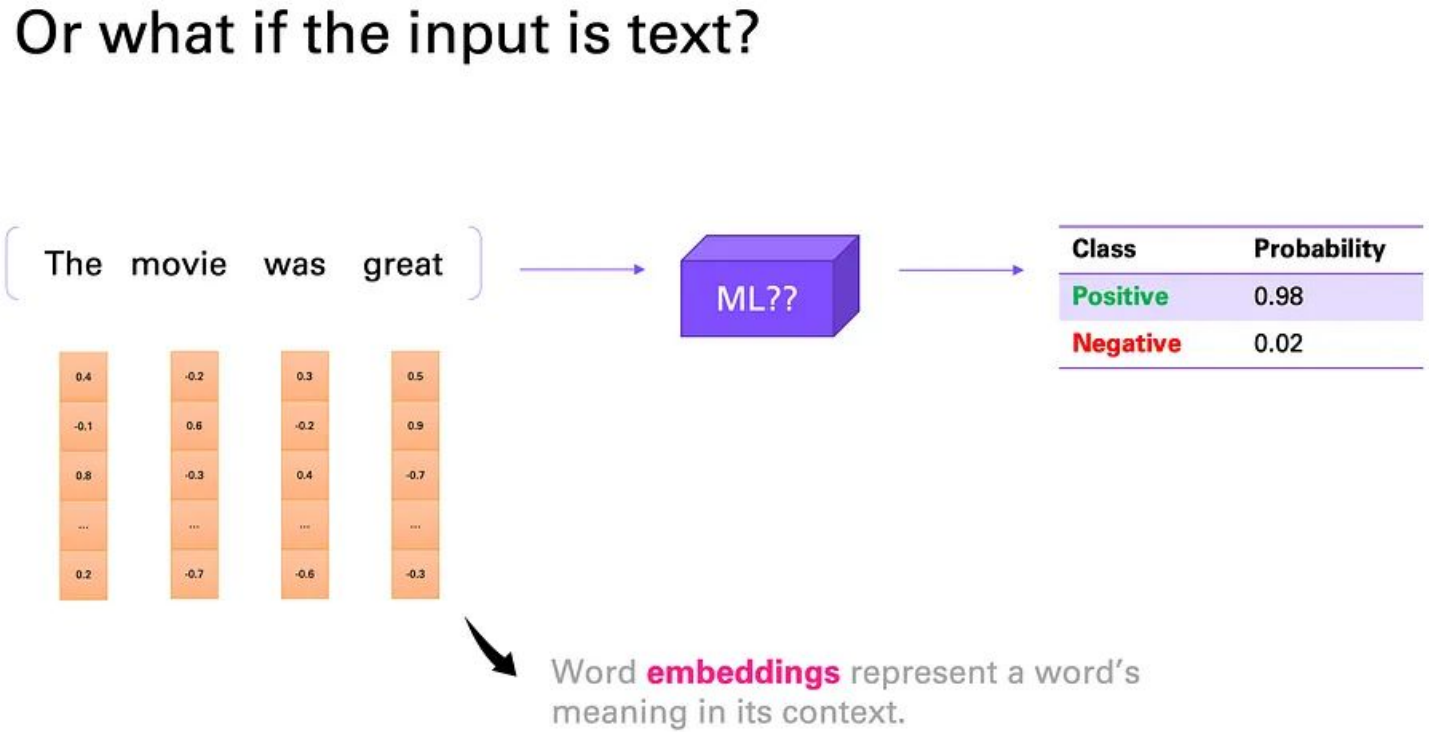
Steps

How it looks in practice



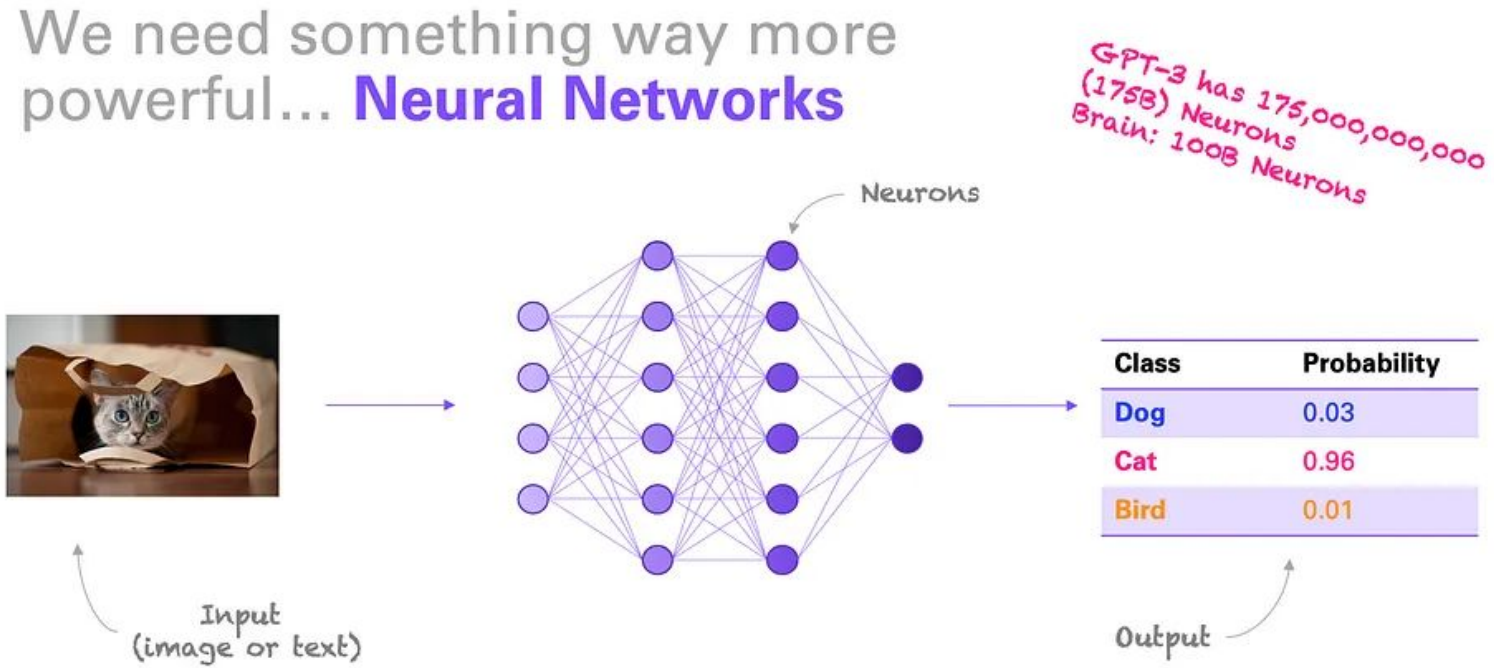
Приклад Машинного навчання: Задача визначення приналежності (класифікації) пісні до одного із двох музичних жанрів (reggaeton чи R&B) основі аналізу Темпу і Енергійності аудіозапису

Основа обробки мови (NLP): векторне представлення



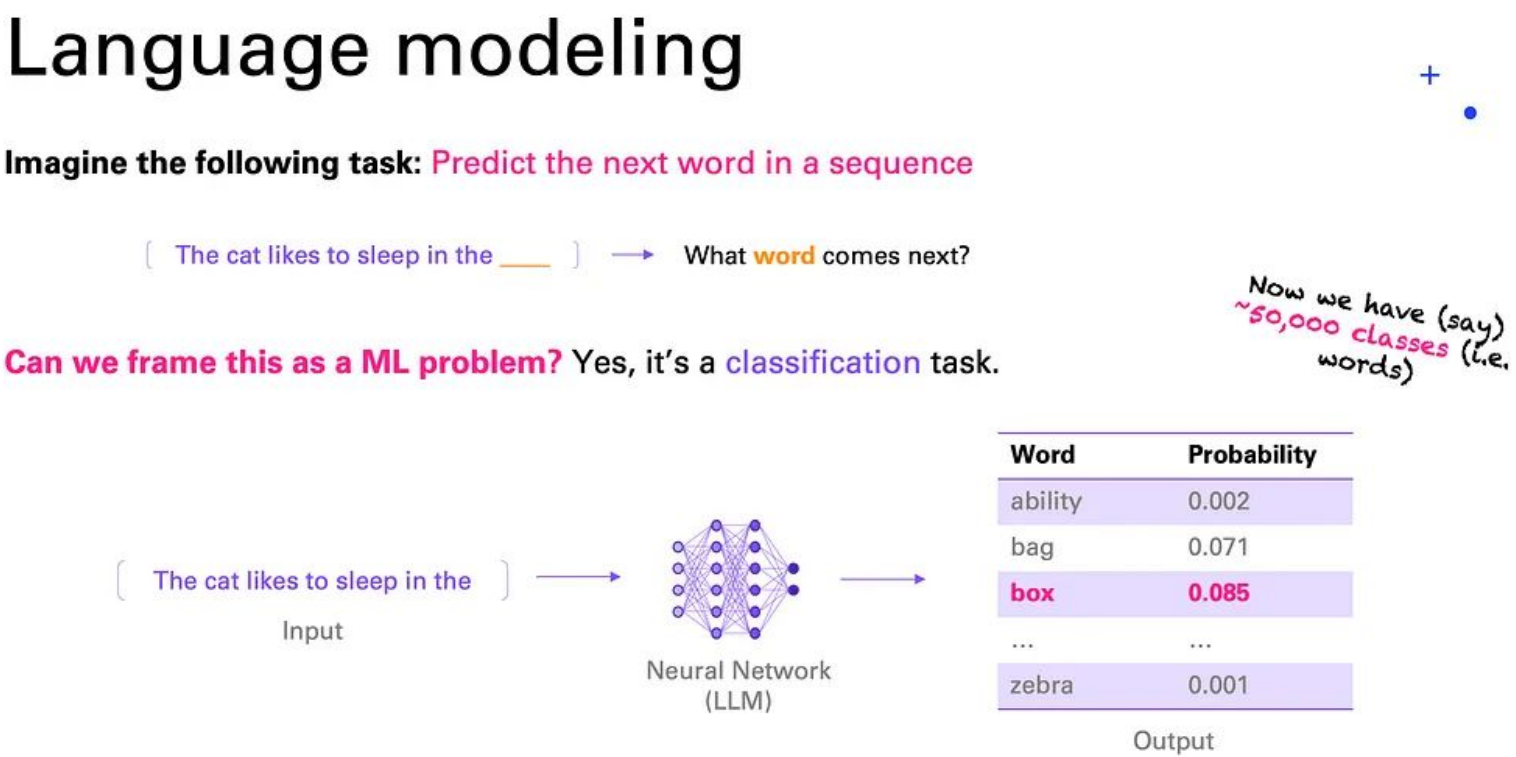
Ключова технологія NLP - Word Embedding (векторне представлення слів/речень/документів)

Нейронні мережі і глибоке навчання (Deep Learning)



Нейронні мережі — це потужні моделі машинного навчання, які дозволяють моделювати довільно складні зв'язки. Вони є двигуном, який дозволяє вивчати такі складні відносини у великому масштабі.

Мовне моделювання (Великі Мовні Моделі - LLM)



*"**Велика Мовна Модель**" (LLM, від англ. Large Language Model) - це складний (більше 1 млрд. зв'язків) тип нейронних мережу. Її особливість - у здатності глибоко розуміти мову, що дозволяє ефективно генерувати текст та відповідати на запити.*

Навчання мовних моделей

Massive training data

We can create **vast amounts of sequences** for training a language model

We do the same with much **longer sequences**. For example:

A language model is a probability distribution over sequences of words. [...] Given any sequence of words, the model predicts the next ...

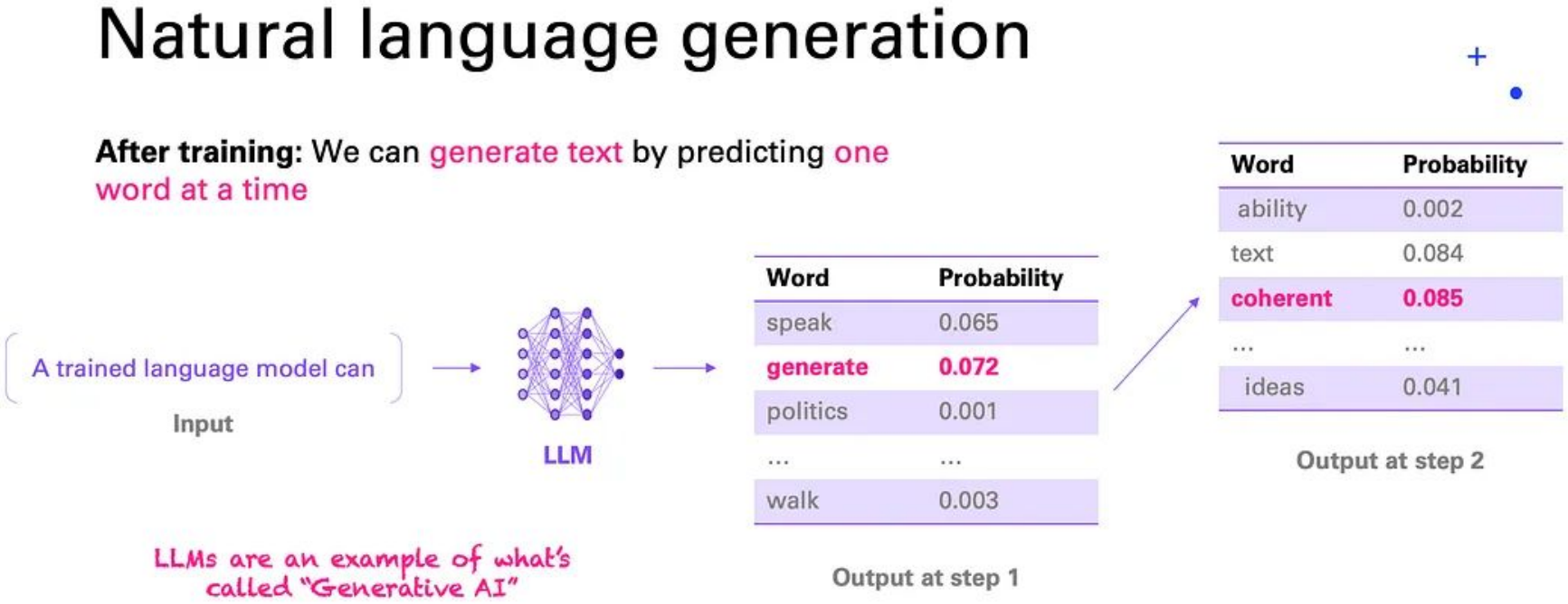
Or also with **code**:

```
def square(number):
    """Calculates the square of a number."""
    return number ** 2
```

And as a result - the model becomes incredibly good at predicting the next word in any sequence.

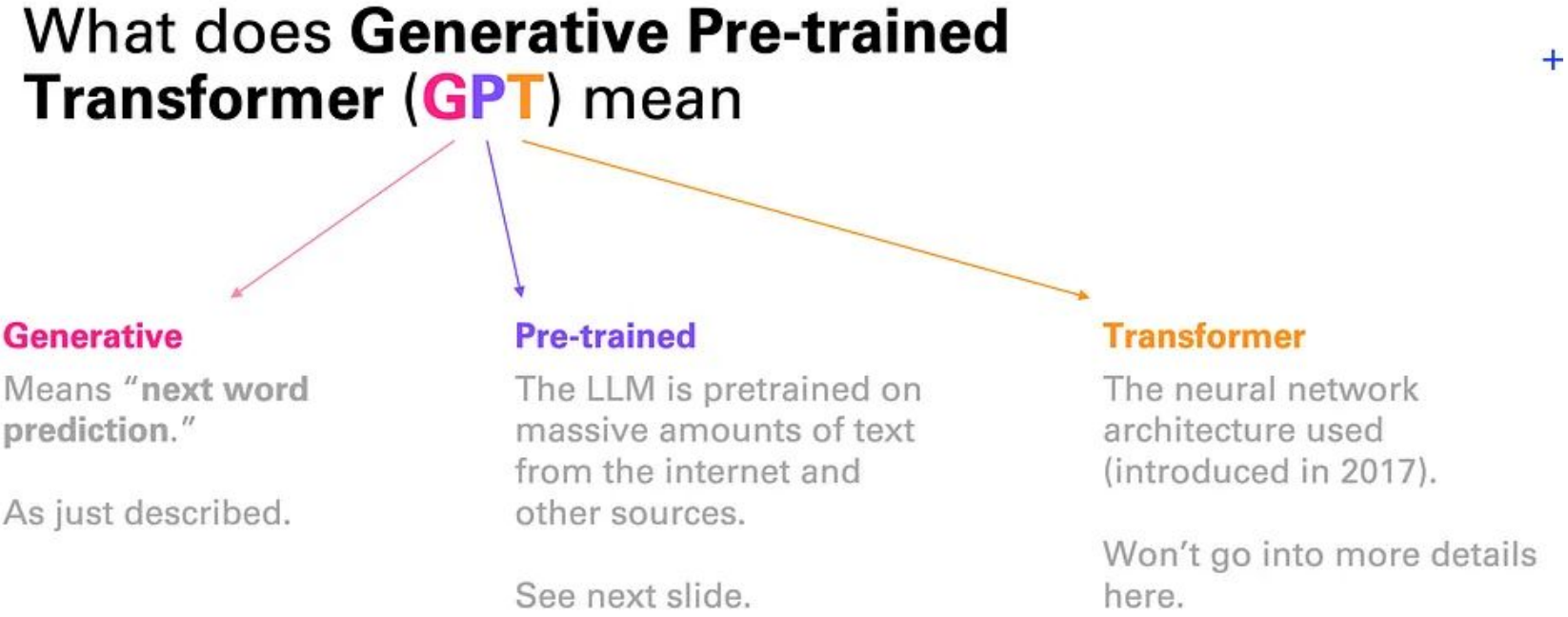
*"Основне завдання - навчити нейронну мережу (LLM) **передбачати наступне слово**. Для цього використовується велика кількість текстових даних з Інтернету, книг, досліджень тощо. Методика самоконтрольованого навчання дозволяє створювати набори даних без позначення, використовуючи наступне слово як мітку. Робимо це для різноманітних послідовностей, коротких і довгих. Навчання LLM передбачати наступне слово в будь-якій мові і контексті, чи то твіт, вірш чи код, покращує його здатність робити відповідні вибори слова, хоч і не завжди ідеально."*

Генеративні Мовні Моделі (Generative AI)



*Завдяки навчанню LLM передбачати слова, ми можемо генерувати тексти, передаючи послідовності назад у модель для передбачення наступних слів. Таким чином, LLM є прикладом **Generative AI**. Важливо, що LLM не обмежується передбаченням найбільш ймовірного слова, але може вибирати з кількох варіантів, абезпечуючи креативність у відповідях. Така стратегія використовується в ChatGPT, де відповіді не завжди однакові. ChatGPT не називається ChatLLM, оскільки мовне моделювання - лише частина його функціоналу, а GPT відображає його здатність до генерування тексту.*

GPT - Генеративний попередньо навчений трансформер.



'G' *у GPT означає 'generative', що вказує на здатність моделі генерувати мову.* ***'P'*** *означає 'попереднє навчання', що відображає етаповий процес навчання моделі.* ***'T'*** *означає 'трансформер', тип архітектури нейронної мережі, що зосереджує увагу на найважливіших частинах вхідної послідовності, схоже на людське сприйняття.*

Фази навчання LLM

Phases of training LLMs (GPT-3 & 4)

1. Pretraining

Massive amounts of data from the internet + books + etc.

Question: What is the problem with that?

Answer: We get a model that can babble on about anything, but it's probably not **aligned** with what we want it to do.

2. Instruction Fine-tuning

Teaching the model to respond to instructions.

Model learns to respond to instructions.

→ Helps **alignment**

"Alignment" is a hugely important research topic

3. Reinforcement Learning from Human Feedback

Similar purpose to instruction tuning.

Helps produce output that is closer to what humans want or like.

(1) Попередня підготовка,
(2) Точне налаштування інструкцій,
(3) Підкріплення за допомогою зворотного зв'язку людини (RLHF).

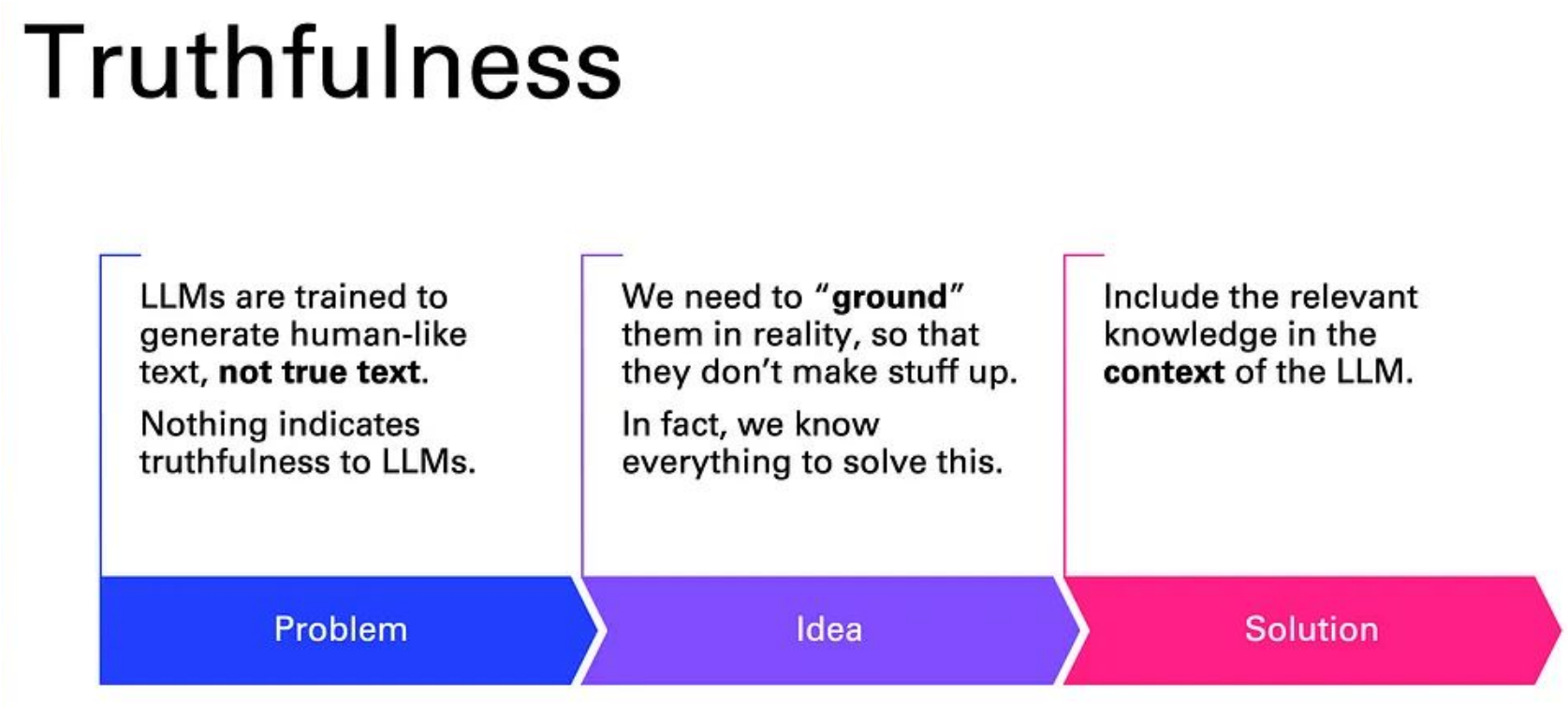
Приклади для перевірки нашого розуміння LLM

Three examples to test our understanding

Ability	Explanation
Why can an LLM perform **Text Summarization**?	Ability probably learned during **pre-training**
Why can an LLM perform **Question Answering**?	**Knowledge** acquired in pre-training, responds nicely due to fine-tuning
Why does a LLM sometimes answer wrong or even make stuff up?	Let's discuss this next...

Чому LLM може виконувати резюмування довшого фрагмента тексту?
Чому LLM може відповісти на загальновідомі запитання ?
Чому LLM іноді відповідає неправильно і навіть вигадує щось?

Правдивість VS “Галюцинації”



LLM страждають від галюцинацій, але їх можна пом’якшити, надавши додатковий контекст (додаткові дані).

Провідні провайдери та моделі генеративного ШІ

Провайдер	Провідна LLM	Контекстне вікно (токени)	Макс. вихідних токенів	Режим "роздумів"	Відкриті ваги	Інші важливі характеристики
OpenAI	o1, o3 GPT-4o	128 000	4 096	Так	Ні	Мультимодальність (текст, аудіо, зображення, відео)
Google	Gemini 2.5 Pro, Gemini 2.0 Flash	1 000 000 - 2 000 000	8 192 - 65 536	Так	Ні	Мультимодальність, інтеграція з Google Workspace, розширене міркування
Anthropic	Claude 3.7 Sonnet, Claude 3.5 Sonnet	200 000	4 096 - 128 000	Так	Ні	Акцент на безпеці та пояснюваності, розширений режим мислення
Meta	Llama 4 Scout, Llama 4 Maverick,	128 000 - 10 000 000	4 096 - 8 192	Ні	Так	Мультимодальність, велике контекстне вікно (Scout), висока продуктивність (Maverick)
Mistral AI	Mistral Large 2, Mistral Small 3.1, Codestral	131 000 - 256 000	8 192 - 131 000	Так (в деяких моделях)	Так (в деяких моделях)	Сильні можливості міркування, багатомовність, оптимізація для кодування (Codestral)

Чати vs API

Провайдер	Режим чат-боту	API (для розробників)
OpenAI	https://chatgpt.com/	https://platform.openai.com/
Google	https://gemini.google.com/	https://aistudio.google.com/
Anthropic	https://claude.ai/	https://console.anthropic.com/
Mistral AI	https://chat.mistral.ai/	https://console.mistral.ai/
DeepSeek	https://chat.deepseek.com/	https://platform.deepseek.com/
Qwen	https://chat.qwen.ai/	https://bailian.console.alibabacloud.com/

Рейтинги LLM (https://lmarena.ai/)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1437	+8/-6	7431	Google	Proprietary
2	2	ChatGPT-4o-latest (2025-03-26)	1406	+7/-8	6612	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1402	+5/-5	13919	xAI	Proprietary
2	2	GPT-4.5-Preview	1397	+5/-6	13443	OpenAI	Proprietary
5	8	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+5/-4	25266	Google	Proprietary
5	4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-5	20136	Google	Proprietary
5	4	DeepSeek-V3-0324	1370	+7/-7	4721	DeepSeek	MIT
7	5	DeepSeek-R1	1359	+5/-5	15098	DeepSeek	MIT
8	13	Gemini-2.0-Flash-001	1354	+4/-4	21065	Google	Proprietary
8	4	o1-2024-12-17	1350	+4/-5	27831	OpenAI	Proprietary
10	13	Gemma-3-27B-it	1342	+7/-6	9147	Google	Gemma
11	13	Qwen2.5-Max	1340	+4/-4	19995	Alibaba	Proprietary
11	10	o1-preview	1335	+5/-4	33175	OpenAI	Proprietary
14	13	o3-mini-high	1325	+6/-4	16889	OpenAI	Proprietary
14	15	DeepSeek-V3	1318	+4/-4	22843	DeepSeek	DeepSeek
14	20	QwQ-32B	1315	+6/-8	6729	Alibaba	Apache 2.0

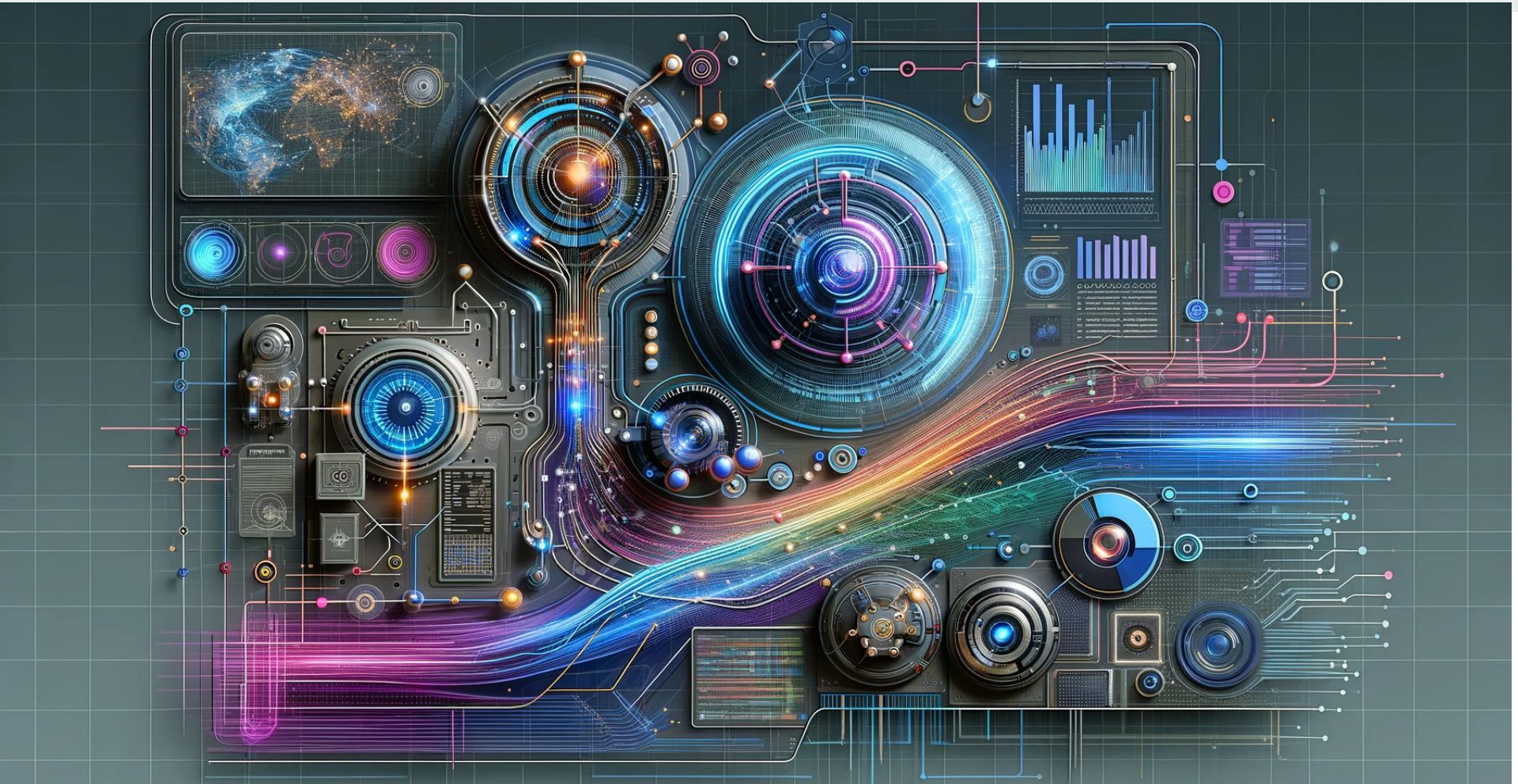
Хмарні платформи

Google Cloud Platform (GCP), Microsoft Azure та Amazon Web Services (AWS)

Які завдання можуть виконувати інструменти ШІ?

Інструменти на основі ШІ

Магія Штучного Інтелекту



Thank you