Випуски великих мовних моделей (LLM)

- У 2024 році OpenAI планує представити GPT-5, який обіцяє значні покращення в своїх можливостях.

- LLaMA 3.1: нова версія моделі від Meta, що продовжує розвиток LLM.

- Qwen2 від Alibaba: ще один важливий реліз у цій сфері.

- Gemini 1.5: нова модель від Google DeepMind, яка обіцяє вдосконалення в обробці природної мови та мультимодальних даних.

Технічні покращення

- Розширене контекстне вікно: Модель здатна обробляти більшу кількість контекстних даних одночасно, зокрема 128 тисяч токенів. Це досягнуто завдяки використанню більшого обсягу пам’яті, тренуванню на різноманітних типах даних та змінам в архітектурі моделі.

- Мультимодальність: Модель має можливість одночасно працювати з текстом і зображеннями. У GPT-4o був змінений підхід, що дозволив використовувати єдину модель для роботи з різними мультимедійними даними, на відміну від попередніх версій, де використовувалися кілька моделей.

Тренди

Я б виділив два основні тренди, які наразі мають значну увагу в галузі:

1. Визначення генерованого тексту: Оскільки моделі навчаються на текстах, які вони самі генерують, це сповільнює процес навчання.

2. Вплив LLM на прийняття рішень: Зростає використання AI-агентів у різних сферах.

3. Етика та відповідальність у використанні LLM: Зростаюча увага до етичних аспектів використання великих мовних моделей, включаючи питання конфіденційності, упередженості та впливу на суспільство.

Вплив на ринок та галузі

- Освіта: Кастомізація та персоналізація навчальних матеріалів стають більш доступними завдяки LLM.

- Розробка програмного забезпечення: Інструменти на кшталт GitHub Copilot та Cody значно підвищують продуктивність написання коду.

- Медицина: Використання таких технологій, як AlphaFold для персоналізованої медицини та обробки медичних зображень, а також їх опису.

Ці інновації та тренди свідчать про стрімкий розвиток LLM та їх вплив на різні галузі, формуючи нові можливості для автоматизації та підвищення ефективності.