202401 Co z tym AI?

Praktyczny słownik na początek

model uczenia maszynowego - algorytm wytrenowany na zbiorze danych

Generative AI - sztuczna inteligencja zdolna do generowania tekstu, obrazów lub innych mediów, wykorzystując modele generatywne. Modele Generative AI uczą się wzorców i struktury swoich danych wejściowych, a następnie generują nowe dane o podobnych cechach.

GPT (Generative Pretrained Transformer) - typ dużego modelu językowego (LLM) i wybitnego frameworka dla generatywnej sztucznej inteligencji. Są to sztuczne sieci neuronowe, które są używane w zadaniach przetwarzania języka naturalnego.

GPT-3 - przełomowy duży model językowy opracowany przez OpenAI.

transformer - architektura głębokiego uczenia oparta na mechanizmie wielogłowicowej uwagi. Jest znana z tego, że nie zawiera żadnych jednostek rekurencyjnych, a więc wymaga mniej czasu na szkolenie niż poprzednie architektury rekurencyjne, takie jak LSTM.

temperatura - parametr sterujący losowością generowanych przewidywań. Jeśli temperatura jest niska, prawdopodobieństwo wyboru innych tokenów niż ten z najwyższym logarytmicznym prawdopodobieństwem będzie małe, a model prawdopodobnie wygeneruje najbardziej poprawny tekst, ale raczej nudny, z małą różnorodnością. Z drugiej strony, wyższa wartość temperatury, na przykład 1.0, prowadzi do większej losowości i kreatywności. W skrócie, temperatura determinuje, jak zachłanny jest generatywny model1.

token - część tekstu, najczęściej słowa. Ilość tokenów jest używana do miary "wysiłku" (jak i kosztu), który LLM ma włożyć w odpowiedź na nasze pytanie. Dla modelu GPT3.5 w języku angielskim, jeden token to ~0.75 słowa, w polskim jedynie ~0.5 słowa. Taniej zatem jest generować tekst po angielsku.

halucynacja - w dziedzinie sztucznej inteligencji, halucynacja lub sztuczna halucynacja (również nazywana konfabulacją lub urojeniem) to odpowiedź generowana przez AI, która zawiera fałszywe lub mylące informacje przedstawiane jako fakt.

prompt - inaczej tekst wejściowy, czyli nasze polecenie.

prompt engineering - optymalizacja tekstu wejściowego, poleceń, której celem jest otrzymanie na wyjściu dokładnie tego czego oczekujemy. Wymaga umiejętności i rozumienia w jaki sposób dany model LLM procesuje tekst

Mid Journey - jeden z pierwszych i najpopularniejszy model do generowania obrazów. Dostępny na Discord. Płatny i dostępny jedynie jako usługa w abonamencie.

Stable Diffusion - otwarty model do generowania obrazów. Można go pobrać na dysk i używać lokalnie wykorzystując moc obliczeniową karty graficznej komputera lub nawet smartfona.

LlaMa 2 - otwarty model LLM opublikowany przez Meta (spółkę matkę Facebooka) i dostępny do pobrania. Porównywalny z komercyjnie dostępnymi rozwiązaniami. Szeroko stosowany do "fine-tuningu" przez biznesy.

Retrieval Augmentation Generation - metoda wzbogacania modelu LLM o nasz własny kontekst, najczęściej jest to prywatne repozytorium wiedzy, wewnętrznych dokumentów, pracowników w firmie. W ten sposób LLM będzie mógł odpowiadać na pytania, również na podstawie tych informacji

Agent - komponent wykonawczy w architekturze LLMów np. LangChain. Podstawową ideą agentów jest wykorzystanie modelu języka do wyboru sekwencji działań do podjęcia. W łańcuchach LangChaina sekwencja działań jest zakodowana. W agentach model języka jest używany jako silnik rozumowania do określania, które działania podjąć i w jakiej kolejności, np. jeśli jest do działanie matematyczne może skierować je do WolframAlpha.

Linki

https://voicebot.ai/large-language-models-history-timeline/ - historia LLMów
https://platform.openai.com/playground?mode=chat&model=gpt-3.5-turbo - możliwość przetestowania GPT-3.5
http://platform.openai.com/playground?mode=chat&model=gpt-4 - możliwość przetestowania GPT-4
https://www.typingmind.com/ - płatny interfejs do modeli OpenAI oraz Anthropic z bogatą biblioteką promptów
https://chatpad.ai - wygodny interfejs do OpenAI (wymaga klucza do API)
https://www.aiprm.com/ - dodatek do przeglądarki do szybszego używania modeli GPT od OpenAI
https://learnprompting.org/ - strona poświecona nauce promptowania
https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/advanced-prompt-engineering - kurs promptowania od Microsoftu
https://gpus.llm-utils.org/llama-2-prompt-template/ - szablon prompta do użycia z otwartym modelem Llama-2 (techniczne)
https://llm-utils.org/ - skarbnica wiedzy o LLMach, różne notatki (techniczne)
https://getimg.ai/guides - poradniki od polskiego startupu. Mają zastosowanie do Stable Diffusion i podobnych modeli
https://gpus.llm-utils.org/stable-diffusion-guides/ - poradniki Stable Diffusion
https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard - ranking otwartych modeli LLM

Znalazłeś błąd lub czegoś zabrakło? Wypełnij https://tally.so/r/wzYqbM lub skontaktuj się z mną: pawel@bogumilo.co