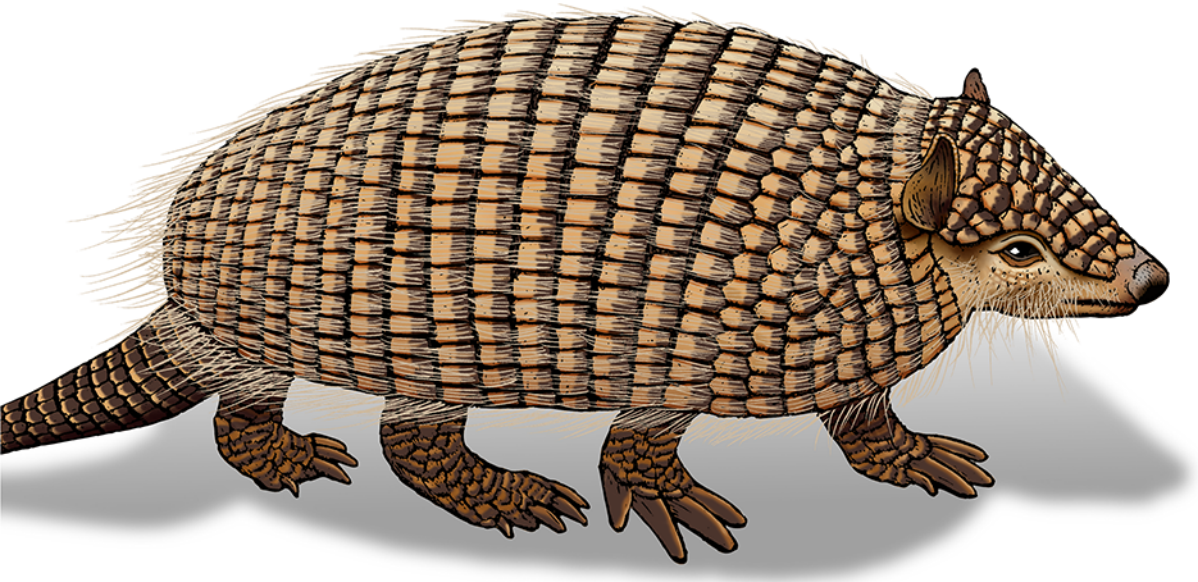
O'REILLY®

Helion

Skuteczna inżynieria promptów

Przyszłościowe rozwiązania dla rzetelnych wyników generatywnej AI



James Phoenix
Mike Taylor

Tytuł oryginału: Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs

Tłumaczenie: Krzysztof Rychlicki-Kicior

ISBN: 978-83-289-1904-4

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/skuinz
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

Spis treści

ROZDZIAŁ 2.

Wprowadzenie do dużych modeli językowych do generowania tekstu

W świecie sztucznej inteligencji niezwykle ważnym zagadnieniem w ostatnim czasie jest ewolucja dużych modeli językowych. W przeciwieństwie do mniej elastycznych poprzedników, duże modele językowe są w stanie radzić sobie ze znacznie większymi zbiorami danych i się z nich uczyć, co prowadzi do wykształcenia się zdolności do generowania tekstu niezwykle zbliżonego poziomem skomplikowania do możliwości człowieka. Modele radzą sobie z różnymi zadaniami, od pisania treści, przez automatyczne tworzenie oprogramowania, aż po tworzenie interaktywnych czatbotów odpowiadających na bieżąco użytkownikom.

Czym są modele do generowania tekstu?

Modele do generowania tekstu korzystają z zaawansowanych algorytmów w celu znalezienia znaczenia tekstu i generowania odpowiedzi często nieodróżnialnej od pracy człowieka. Jeśli miałeś doświadczenia z usługą ChatGPT (*https://chat.openai.com*) lub doświadczyłeś jej zdolności do generowania spójnych i przydatnych odpowiedzi, to poznałeś moc LLM w praktyce.

Zarówno w przetwarzaniu języka naturalnego, jak i w dużych modelach językowych podstawową jednostką językową jest token. **Tokeny** (*https://oreil.ly/3fOsM*) mogą reprezentować zdania, słowa, a nawet fragmenty słów, czyli zbiory znaków. Dobrą metodą wyrażania rozmiaru danych tekstowych jest określenie liczby tokenów danego tekstu. W języku angielskim 100 tokenów odpowiada mniej więcej 75 słowom. Warto zapamiętać tę proporcję, ponieważ każdy duży model językowy ma swoje ograniczenia wyrażone w tokenach — na podstawie tego przelicznika jesteśmy w stanie oszacować liczbę słów, którą możemy zmieścić w pojedynczym kontekście.

Tokenizacja to proces zamiany tekstu na tokeny — podstawowy etap przygotowania danych pod kątem zadań przetwarzania języka naturalnego (NLP). Tokenizację można wykonać na kilka sposób, np. z wykorzystaniem metody *Byte-Pair Encoding* (kodowanie parami bajtów, BPE — *https://oreil.ly/iSOp7*), *WordPiece* lub *SentencePiece*. Każda z metod ma swoje unikatowe zalety i jest dostosowana do konkretnych przypadków użycia. BPE używa się głównie ze względu na wydajność obsługi w przypadku rozbudowanego słownika przy jednoczesnym zachowaniu rozsądnej liczby tokenów.

Metoda BPE rozpoczyna swoje działanie od potraktowania tekstu jako ciągu pojedynczych znaków. Z czasem najczęściej występujące ciągi znaków są grupowane w pojedyncze jednostki, zwane tokenami. Aby lepiej zrozumieć ten mechanizm, rozważmy słowo *jabłko*. Na początku BPE może traktować to słowo jako sześć niezależnych znaków — *j*, *a*, *b*, *ł*, *k*, *o*. Po analizie tekstów algorytm może zauważyć, że litera *b* często pojawia się po *a*, *a* zaś po *j*, z kolei *b* występuje przed *ł*, a *ł* przed *k*, co sprawia, że ciąg *jabłk* może zostać potraktowany jako pojedynczy token.

Takie podejście może pomóc dużym modelom językowym w generowaniu słów lub fraz, nawet jeśli nie występowały one często w danych treningowych — dzięki temu modele są w stanie dostosowywać się do różnych sytuacji i są bardziej uniwersalne.

Zrozumienie zasad działania dużych modeli językowych wymaga przyswojenia podstaw rządzących nimi reguł matematycznych. Choć obliczenia potrafią być skomplikowane, możemy uprościć kluczowe elementy, aby przedstawić intuicyjne zasady działania tych modeli. Zwłaszcza w kontekście aplikacji dla biznesu dokładność i stabilność dużych modeli językowych jest kluczowa.

Istotnym czynnikiem w osiągnięciu tej stabilności są fazy wstępnego uczenia (ang. *pretraining*) i dostrajania (and. *fine-tuning*) w tworzeniu dużych modeli językowych. Na początku modele są uczone na ogromnych zbiorach danych w fazie wstępnego uczenia, co umożliwia im zrozumienie języka powszechnego, używanego w typowych, ogólnych sytuacjach. Z kolei w następującej potem fazie dostrajania modele dostosowuje się do konkretnych zadań, co pozwala zapewnić dokładne i wiarygodne wyniki dla aplikacji specjalistycznych.

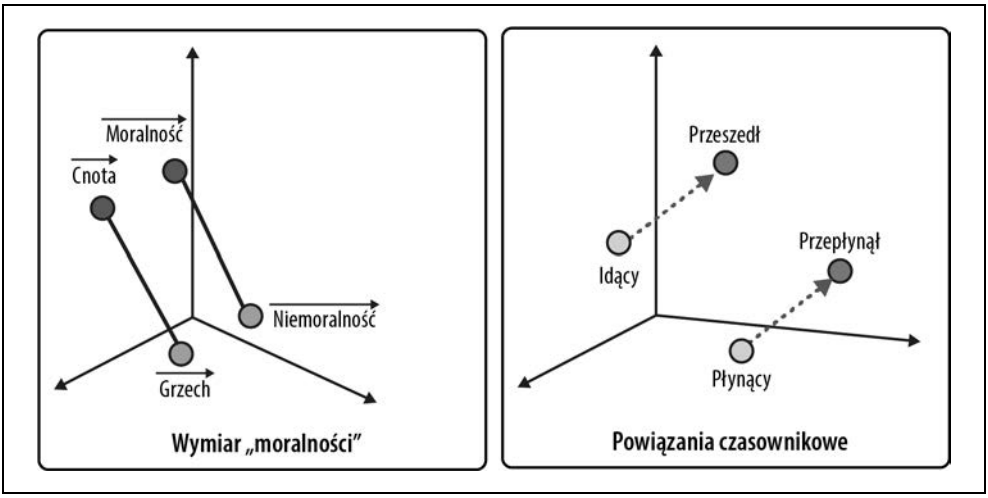
Reprezentacje wektorowe: język w liczbach

W świecie naturalnego przetwarzania języka słowa nie są jedynie zlepkami liter. Słowa mogą być poddane tokenizacji, a następnie przetwarzane w postaci liczbowej — wektorowej. Wektory to wielowymiarowe tablice liczb, które potrafią uchwycić związki semantyczne i składniowe:

$$w \rightarrow \mathbf{v} = [v_1, v_2, \ldots, v_n]$$

Tworzenie wektorów słów, zwanych także **osadzeniami słów** (ang. *word embeddings*), polega na wykrywaniu zawiłych wzorców obecnych w języku. Podczas intensywnej fazy uczenia modele mają wykrywać te wzorce i się ich uczyć, by tym samym zapewnić, że słowa o podobnych znaczeniach będą znajdować się w przestrzeni wielowymiarowej blisko siebie (rysunek 2.1).

Piękno tego podejścia polega na zdolności do uchwycenia subtelnych związków pomiędzy słowami i obliczenia odległości między nimi. Jeśli przeanalizujemy osadzenia słów, to zauważymy bez problemu, że słowa o podobnych znaczeniach, takie jak *cnota* i *moralność* lub *idący* i *przeszedł*, są zlokalizowane blisko siebie. Bliskość w przestrzeni osadzeń staje się niezwykle istotna w rozmaitych zadaniach z zakresu NLP, zapewniając modelom zrozumienie kontekstu, semantyki i skomplikowanej sieci powiązań, które tworzą język.



Rysunek 2.1. Semantyczna bliskość wektorów słów wewnątrz przestrzeni osadzeń słów

Architektura transformerów: orkiestracja związków kontekstowych

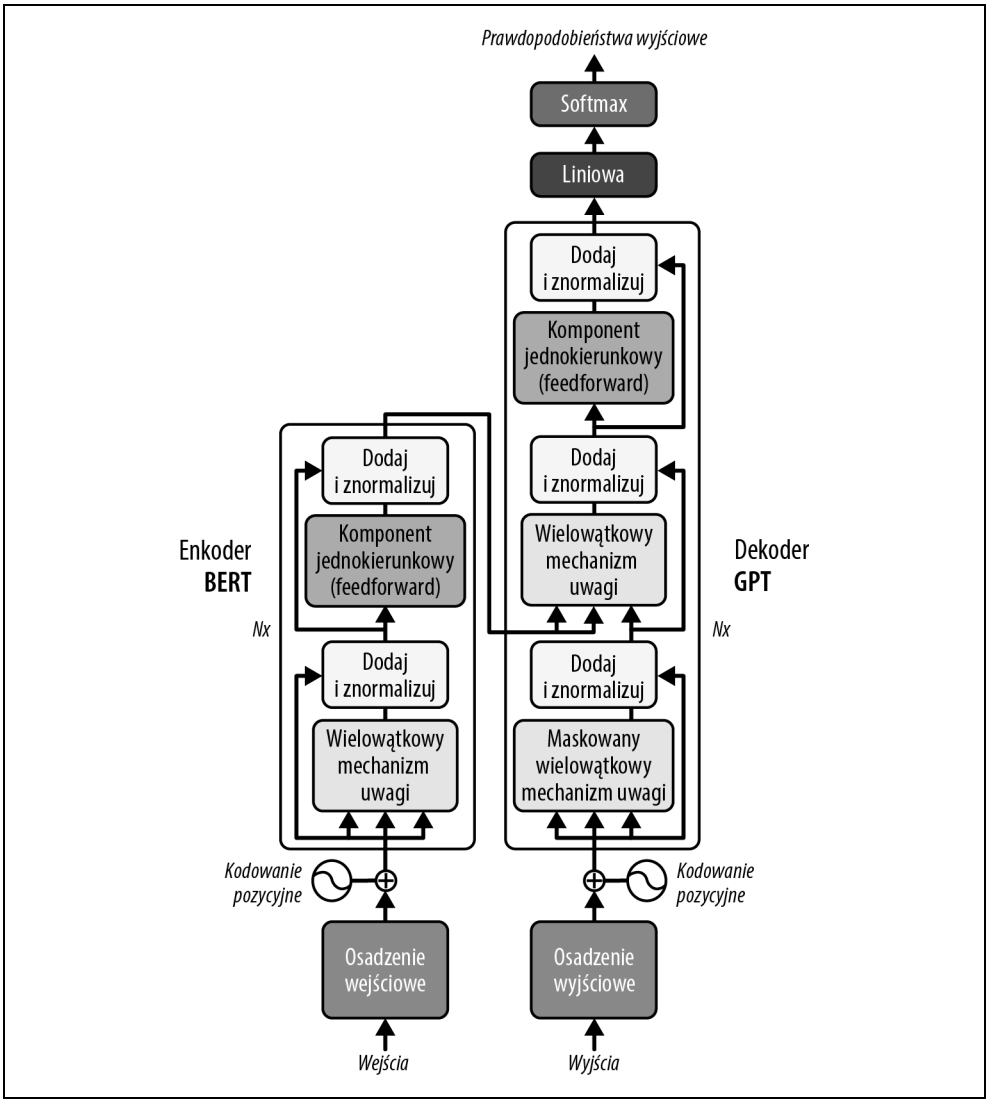
Zanim bardziej zagłębimy się w szczegóły działania architektury transformerów, musimy zrozumieć podstawy. Jeżeli mamy dowolne zdanie, np. *kot usiadł na podłodze*, to każde słowo w zdaniu zostanie przetworzone na reprezentację wektorową. Słowo *kot* zostaje przekonwertowane na serię liczbę, podobnie jak *usiadł*, *na* i *podłodze*.

Jak przekonasz się w dalszej części rozdziału, architektura transformerów przyjmuje te wektory słów i znajduje związki występujące pomiędzy nimi — zarówno pod kątem struktury (składni), jak i znaczenia (semantyki). Istnieje wiele rodzajów transformerów — rysunek 2.2 przedstawia architektury BERT i GPT. Co więcej, transformer nie rozpatruje słów niezależnie — w naszym przykładzie wie, że słowo *kot* jest w tym zdaniu w pewien sposób powiązane z *usiadł* i *podłodze*.

Transformer w trakcie przetwarzania wektorów korzysta z matematycznych operacji do zrozumienia związków między słowami i w ten sposób tworzy nowe wektory o bogatej informacji kontekstowej:

$$\mathbf{v}_i' = \text{Transformer}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$$

Jedną z ważniejszych możliwości transformerów jest zdolność do zrozumienia subtelnych znaczeń słów, zależnych od kontekstu. Mechanizm samouwagi (*https://oreil.ly/xuovP*) pozwala każdemu słowu zapoznać się z pozostałymi słowami w pobliżu, aby lepiej zrozumieć kontekst. Można powiedzieć, że każde słowo głosuje na znaczenie innych słów, aby ustalić swoje znaczenie. Analizując całe zdanie, transformery są w stanie dokładnie określić rolę i znaczenie każdego słowa, dzięki czemu *ich zdolności interpretacyjne są bardziej dopasowane do kontekstu*.



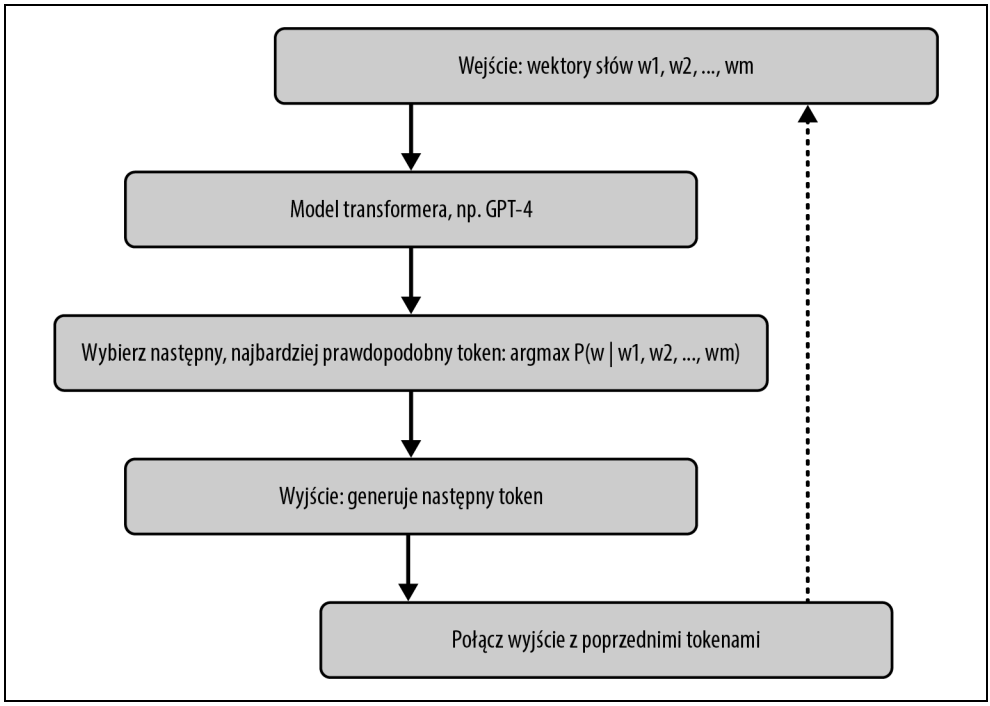
Rysunek 2.2. BERT stosuje enkoder do danych wejściowych, a GPT — dekoder do danych wyjściowych

Probabilistyczne generowanie tekstu: mechanizm podejmowania decyzji

Gdy transformer rozumie kontekst danego tekstu, może przejść do generowania nowego, korzystając z pojęcia prawdopodobieństwa (ang. *likelihood/probability*). W sensie matematycznym: model oblicza prawdopodobieństwo wystąpienia kolejnego słowa, biorąc pod uwagę bieżącą sekwencję słów, i wybiera to najbardziej prawdopodobne:

$$w_{\text{nast}} = \text{argmax}\ P(w \mid w_1, w_2, \ldots, w_m)$$

Powtarzając ten proces (rysunek 2.3), model jest w stanie wygenerować spójny i kontekstowo istotny tekst.



Rysunek 2.3. Ogólna metoda generowania tekstu przez modele transformerów, takich jak GPT-4

Mechanizmy stojące za sukcesem dużych modeli językowych pochodzą z matematyki wektorów, transformacji liniowej i modeli probabilistycznych. Pod względem implementacji operacje te są niezwykle intensywne obliczeniowo, niemniej warto pamiętać, że główne pojęcia związane z funkcjonowaniem transformerów opierają się na wspomnianych matematycznych zasadach, co z pewnością ułatwi zrozumienie technicznych aspektów działania transformerów przez interesariuszy, np. z biznesu.

Krótko o historii: wzrost znaczenia architektur transformerów

Modele językowe takie jak ChatGPT (GPT — *Generative Pretrained Transformer*, **wstępnie przeszkolony transformer generatywny**) nie pojawiły się znikąd. Ich obecność to efekt wielu lat postępu w dziedzinie przetwarzania języka naturalnego (NLP), zwłaszcza w drugiej dekadzie tego stulecia. Kluczowym elementem tego postępu było wprowadzenie architektur transformerów, opisanych szczegółowo w przełomowym artykule *Attention Is All You Need* („Uwaga to wszystko, czego potrzebujesz”, *https://oreil.ly/6NNbg*) zespołu Google Brain.

Prawdziwym przełomem w architekturach transformerów było wprowadzenie mechanizmu uwagi (ang. *attention*). Tradycyjne modele przetwarzały tekst sekwencyjnie, co ograniczało możliwość zrozumienia struktury języka, zwłaszcza na przestrzeni długich odstępów w ramach tekstu. Mechanizm uwagi zmienił tę sytuację: umożliwił modelom powiązanie ze sobą odległych słów, niezależnie od ich położenia w tekście. Takie podejście było naprawdę przełomowe — dzięki niemu słowa i ich kontekst nie musiały być przemieszczane w obrębie całego modelu, aby mieć na siebie wpływ. Takie podejście nie tylko znacząco poprawiło zdolność rozumienia tekstu przez modele, ale także zwiększyło ich wydajność.

Mechanizm uwagi odgrywa najważniejszą rolę w umożliwieniu modelom wykrywania zależności pomiędzy słowami znajdującymi się w dużej odległości od siebie. Ta zmiana stała się kluczowa w generowaniu odpowiedzi, które są nie tylko kontekstowo dokładne, ale także spójne na dłuższym dystansie.

Według pioniera i edukatora AI Andrew Ng (*https://oreil.ly/JQd53*) duża liczba wczesnych badań prowadzonych w zakresie NLP była istotnie dofinansowana przez agencje wywiadu wojskowego Stanów Zjednoczonych. Duże zainteresowanie tych agencji w zakresie narzędzi do tłumaczenia maszynowego czy rozpoznawania mowy, głównie pod kątem działań wywiadowczych, siłą rzeczy przyczyniło się do rozwoju możliwości wykraczających poza tłumaczenie.

Uczenie dużych modeli językowych wymaga sporych zasobów obliczeniowych. Modele są „karmione" ogromnymi ilościami danych — rzędu terabajtów, a nawet petabajtów — w których skład wchodzą treści z Internetu, artykuły naukowe, książki i bardziej specjalistyczne zbiory danych. Należy przy tym pamiętać, że dane używane do trenowania dużych modeli językowych mogą charakteryzować się różnymi *tendencjami* (ang. *bias*), *wynikającymi wprost ze źródeł*. W związku z tym użytkownicy powinni zachować szczególną ostrożność i ciągły nadzór nad wdrażaniem modeli, by zapewnić możliwość tworzenia bezpiecznych i etycznych aplikacji AI.

Model GPT-4, należący do OpenAI, dysponuje podobno mniej więcej 1,7 biliona parametrów (*https://oreil.ly/pZvMo*), co stanowi odpowiednik arkusza Excela o wielkości 30 tysięcy boisk do piłki nożnej. **Parametry** — w kontekście sieci neuronowych — to wagi i tendencje dostosowywane w procesie uczenia, które pozwalają modelowi na rozpoznawanie i generowanie złożonych wzorców na podstawie użytych danych treningowych. Koszt nauczenia modelu GPT-4 jest szacowany na 63 miliony dolarów (*https://oreil.ly/_NAq5*) i same dane treningowe byłyby w stanie wypełnić 650 kilometrów regałów wypełnionych książkami (*https://oreil.ly/D7jL5*).

Osiągnięcie takich efektów wymagało sporych inwestycji ze strony firm technologicznych, takich jak Microsoft, Meta czy Google, co uczyniło rozwój dużych modeli językowych niezwykle drogim przedsięwzięciem.

Wzrost znaczenia LLM-ów zwiększył zapotrzebowanie także od strony sprzętu, zwłaszcza na produkty firm specjalizujących się w kartach graficznych z wyspecjalizowanymi układami (GPU). Kluczowym graczem stała się Nvidia, utożsamiana wręcz z wysoko wydajnymi kartami graficznymi, niezbędnymi do uczenia dużych modeli językowych.

Zapotrzebowanie na potężne, wydajne procesory graficzne niezwykle wzrosło wraz z budową coraz większych i coraz bardziej złożonych modeli. Nie chodzi przy tym tylko o samą moc

obliczeniową. GPU muszą być także dopasowane do zadań kluczowych z punktu widzenia uczenia maszynowego, takich jak operacje na tensorach. **Tensory** — ponownie w kontekście uczenia maszynowego — to wielowymiarowe tablice danych, a operacje na nich stanowią fundament wszelkich obliczeń związanych z sieciami neuronowymi. Skoncentrowanie się na tych wyspecjalizowanych działaniach przyczyniło się do pojawienia się specjalistycznego sprzętu, takiego jak karty Nvidia H100 Tensor Core, w pełni przystosowane do operacji związanych z uczeniem maszynowym.

Przytłaczający popyt często przewyższa podaż, co naturalnie zwiększa ceny sprzętu. W związku z tym rynek GPU stał się niezwykle konkurencyjnym, ale i zyskownym obszarem, w którym zróżnicowana klientela — od gigantów technologicznych do pracowników naukowych — stara się opracować najbardziej zaawansowany sprzęt.

Ogromny popyt przyczynił się do powstania produktów wykraczających poza GPU. Firmy koncentrują się na tworzeniu sprzętu przeznaczonego do zadań związanych ze sztuczną inteligencją, takiego jak TPU (ang. *Tensor Processing Unit* — jednostka przetwarzania tensorów) firmy Google, a wszystko to w celu zaspokojenia rosnącego zapotrzebowania modeli AI na moc obliczeniową.

Ten ciągle zmieniający się krajobraz nie tylko podkreśla silne związki między oprogramowaniem a sprzętem, ale także pokazuje efekt swoistej gorączki złota związanej z dużymi modelami językowymi. Ogromne inwestycje i innowacje trafiają do różnych branż, ale zwłaszcza tych, które oferują komponenty niezbędne do dalszego rozwoju tych modeli.

Wstępnie przeszkolony transformer generatywny OpenAI

Organizacja OpenAI (*https://openai.com*), którą założono z myślą o przyniesieniu całej ludzkości korzyści z wdrażania sztucznej inteligencji, stała się jednym z pionierów rewolucji AI. Jednym z największych jej dokonań jest seria modeli GPT, która znacząco zmieniła nasze postrzeganie tego, co mogą nam dać duże modele językowe.

Oryginalny model GPT autorstwa OpenAI był czymś więcej niż tylko efektem badań. Był to intrygujący przykład możliwości architektur opartych na transformerach. Model ten stanowił pierwszy krok w kierunku umożliwienia maszynom zrozumienia i generowania języka tak, jak robią to ludzie, i położył fundament pod przyszły rozwój technologii.

Prezentacji modelu GPT-2 towarzyszyły zarówno oczekiwania, jak i swego rodzaj niepokój. Mając świadomość dużych możliwości modelu, OpenAI początkowo wahała się, czy go udostępnić, w obawie przed nadużyciami. Obawy natury etycznej stanowiły centrum sporu wokół modelu GPT-2, co może wydawać się osobliwe w porównaniu z możliwościami obecnych modeli. Opublikowanie projektu w formie otwartego oprogramowania (*https://oreil.ly/evOQE*) nie oznaczało jedynie opublikowania kodu. Ta decyzja pozwoliła firmom i badaczom skorzystać ze wstępnie przeszkolonych modeli jako komponentów przy budowie własnych rozwiązań AI, bez konieczności budowania takich rozwiązań od podstaw. Dzięki temu dostęp do wysokiej jakości mechanizmów przetwarzania języka naturalnego stał się możliwy dla wszystkich, co umożliwiło innowacje w różnych obszarach.

Po opublikowaniu modelu GPT-2 OpenAI skupiła się na przedstawieniu płatnych, zamkniętych modeli. GPT-3 uznaje się za ogromny postęp w dziedzinie dużych modeli językowych. Model ten przyciągnął uwagę mediów, nie tylko ze względu na stopień zaawansowania technicznego, ale także na implikacje, jakie niósł on ze sobą dla całego społeczeństwa. Model ten potrafił generować tekst w tak przekonujący sposób, że trudno było odróżnić go od wytworów człowieka. Od pisania rozbudowanych tekstów literackich, aż po generowanie działających fragmentów kodu — GPT-3 pokazał właściwie nieograniczony potencjał AI.

GPT-3.5-turbo i ChatGPT

Napędzona znaczącą inwestycją dokonaną przez Microsoft, OpenAI przedstawiła model GPT-3.5-turbo, zoptymalizowaną wersję doskonałego poprzednika. Dzięki dotacji rzędu miliarda dolarów (*https://oreil.ly/1C8qm*) otrzymanej od Microsoftu w 2019 roku, po której nastąpiła inwestycja o wielkości 13 miliardów dolarów w zamian za pakiet 49% udziałów w komercyjnej spółce powiązanej z OpenAI, możliwe było opracowanie modelu GPT-3.5-turbo, który osiągnął jeszcze większą skuteczność, ale i umożliwił tańsze działanie, dzięki czemu dużych modeli językowych można było używać w znacznie większej liczbie przypadków.

Organizacja chciała uzyskać jak najwięcej informacji zwrotnej od szerokiego grona użytkowników, by dostroić swój model, dlatego uruchomiła usługę ChatGPT (*https://chat.openai.com*). W przeciwieństwie do siostrzanych modeli ogólnego przeznaczenia, ChatGPT dostrojono (*https://oreil.ly/6ib-Q*) pod względem umiejętności prowadzenia konwersacji, co umożliwiło prowadzenie dialogu między ludźmi a maszynami w sposób naturalny i owocny.

Na rysunku 2.4 jest pokazany proces uczenia usługi ChatGPT, obejmujący trzy główne kroki:

Zebranie danych demonstracyjnych

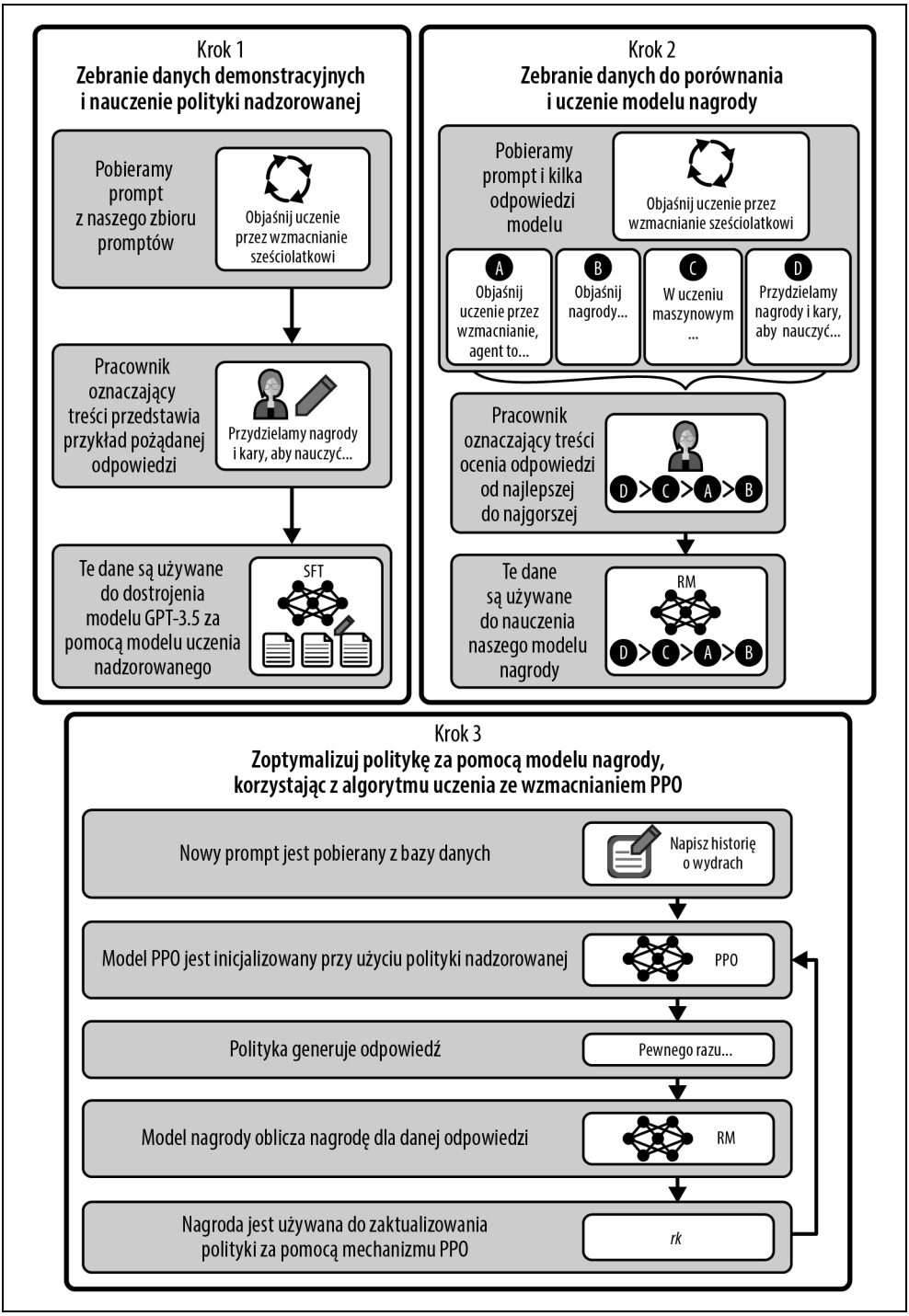
W tym kroku pracownicy przedstawiają przykłady pożądanego zachowania modelu na pewnej selekcji promptów. Pracownicy ci otrzymują konkretne wytyczne, aby mogli oni oznaczać prompty dokładnie.

Uczenie z nadzorem (uczenie nadzorowanej polityki)

Przykładowe dane zebrane w poprzednim kroku są używane do dostrajania wstępnie przeszkolonego modelu GPT-3 za pomocą mechanizmu uczenia z nadzorem (ang. *supervised learning*). W takim trybie modele są uczone z użyciem oznaczonego zbioru danych, z uwzględnieniem poprawnych odpowiedzi dla zadanych problemów. Ten krok pomaga modelowi nauczyć się podążać zgodnie z danym zestawem instrukcji i generować wyjście zgodne z oczekiwaniami.

Zebranie danych porównawczych i uczenie przez wzmacnianie

W tym kroku zbiera się zestaw odpowiedzi modelu, który następnie oceniają pracownicy oznaczający treści. W ten sposób jest uczony model nagrody, który pozwala przewidywać, jak pracownicy oznaczający treści oceniliby dane odpowiedzi. Na zakończenie wprowadza się mechanizm uczenia przez wzmacnianie (ang. *reinforcement learning*), a konkretnie: przybliżoną optymalizację polityki (ang. *Proximal Policy Optimization* — PPO). PPO pozwala optymalizować politykę nadzorowania w celu maksymalizacji nagrody uzyskanej z modelu nagrody.



Rysunek 2.4. Proces dostrajania modelu ChatGPT

Proces uczenia pozwala modelowi ChatGPT uzyskać zachowanie zbliżone do ludzkiego. Zastosowanie uczenia przez wzmacnianie wraz z uwzględnieniem czynnika ludzkiego w całym procesie pozwoliło stworzyć model pomocny, szczery i bezpieczny — w porównaniu ze wstępnie uczonym modelu GPT-3.

Według badań UBS (*https://oreil.ly/2Ivq2*) do stycznia 2023 r. ChatGPT ustanowił rekord 100 milionów aktywnych użytkowników i stał się zarazem najszybciej rozwijającą się aplikacją konsumencką w historii Internetu. ChatGPT stał się domyślnym wyborem w implementacji mechanizmów obsługi użytkownika, wirtualnych asystentów i wielu innych aplikacji, które wymagają prowadzenia konwersacji na poziomie człowieka.

GPT-4

W roku 2024 OpenAI opublikowała model GPT-4, który świetnie sprawdza się w rozumieniu złożonych zapytań oraz generowaniu istotnych i spójnych informacji tekstowych. Dla przykładu, GPT-4 uzyskał na egzaminie prawniczym wynik w granicach 90. percentyla, 298 na 400 punktów. Obecnie model GPT-3.5-turbo jest darmowy do użycia w ramach usługi ChatGPT, natomiast GPT-4 wymaga uiszczenia miesięcznej opłaty (*https://oreil.ly/UOEBM*).

GPT-4 korzysta z podejścia tzw. mieszanki ekspertów (ang. *mixture of experts*, *https://oreil.ly/v45LZ*), w którym model uwzględnia więcej niż jeden wynik wnioskowania (ang. *inference*), aby wygenerować jeszcze bardziej dokładne i istotne wyniki.

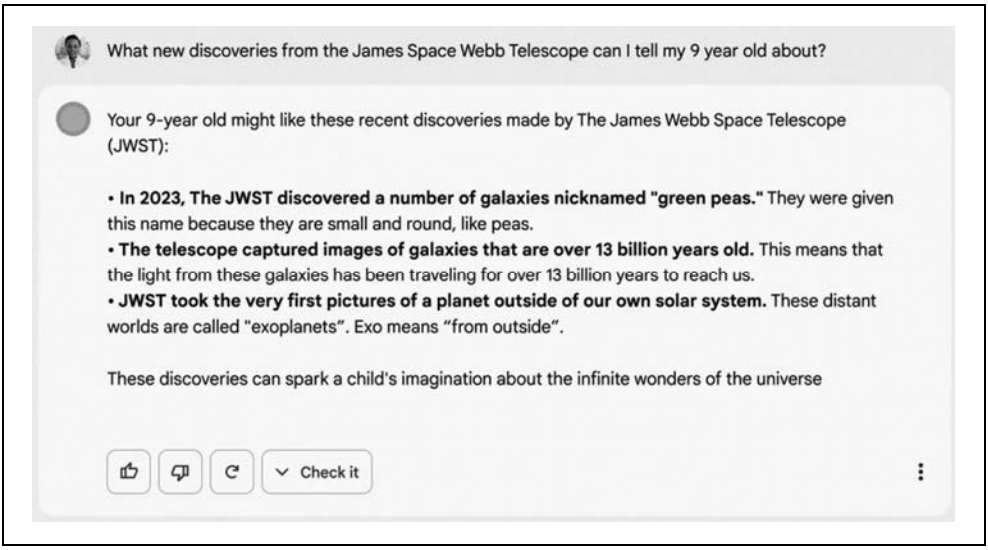
13 maja 2024 r. OpenAI opublikowała GPT-4o (*https://oreil.ly/4ttmq*), zaawansowany model zdolny do przetwarzania danych wejściowych tekstowych, dźwiękowych i graficznych w czasie rzeczywistym. Zapewnia on zwiększoną efektywność, zwłaszcza w kontekście danych audio i wideo. Jest także szybszy i tańszy niż poprzednicy, głównie dzięki możliwości przetwarzania trzech różnych form danych wejściowych w jednej sieci neuronowej.

Google Gemini

Google utraciło część ruchu (i zarazem rynku) wyszukiwarek internetowych właśnie w związku z pojawieniem się modelu ChatGPT. W odpowiedzi, 21 marca 2023 r., opublikowało własny model Bard. Na początku zdarzały się problemy (*https://oreil.ly/Sj24h*), a średnia jakość odpowiedzi zdecydowanie odbiegała od usług udostępnianych przez ChatGPT (rysunek 2.5).

Z czasem dodano kilka nowych funkcji, takich jak generowanie kodu, AI wizualne, wyszukiwanie w czasie rzeczywistym, a także głosowe — co przybliżyło go pod względem jakości do modelu ChatGPT.

14 marca 2023 r. Google opublikowało PaLM API (*https://oreil.ly/EbI8-*), dając dostęp programistyczny z poziomu usługi Google Cloud Platform. W kwietniu 2023 r. AWS (Amazon Web Services) opublikowało podobne usługi, takie jak Amazon Bedrock (*https://oreil.ly/4fNQX*)



Rysunek 2.5. Halucynacje modelu Bard na temat teleskopu Jamesa Webba

czy Amazon's Titan FMs (*https://oreil.ly/FJ-7D*). Google w wersji 1.5 (w lutym 2024 r.) zmieniło nazwę swojego rozwiązania z Bard na Gemini (*https://oreil.ly/EO42O*), a równocześnie jego użytkownicy zaczęli uzyskiwać wyniki zbliżone do tych, które zapewniał GPT-4.

Google opublikowało także dwa mniejsze otwarte modele (*https://oreil.ly/LWIwv*) oparte na tej samej architekturze, co Gemini. Rozwiązanie OpenAI nie stanowi już jedynego, oczywistego wyboru w zakresie dużych modeli językowych dla inżynierów oprogramowania tworzących aplikacje z ich użyciem.

Model Llama firmy Meta a otwarte oprogramowanie

Podejście firmy Meta do modeli językowych różni się znacząco od podejścia konkurencji. Kolejne wersje swojego kluczowego modelu Llama — pierwszą (*https://oreil.ly/LroPn*), drugą (*https://oreil.ly/NeZLw*) i trzecią (*https://oreil.ly/Vwlo-*), opublikowała na zasadach otwartego oprogramowania. Wynika to z faktu, że Meta chce uzyskać bardziej inkluzywny i zachęcający do współpracy ekosystem tworzenia rozwiązań AI.

Otwarty charakter modeli Llama 2 i 3 ma istotne konsekwencje dla szeroko rozumianego przemysłu IT, zwłaszcza dla korporacji. Przejrzystość i etos współpracy napędzają innowacje, ponieważ wszelkie problemy i podatności mogą być szybko zidentyfikowane i rozwiązane dzięki globalnej społeczności programistów. W miarę jak modele stają się bardziej rozwinięte i bezpieczne, korporacje są w stanie ufać w zasadność ich wdrażania.

Strategia otwartego oprogramowania firmy Meta nie tylko demokratyzuje dostęp do najistotniejszych technologii AI, ale także może mieć istotny wpływ na cały przemysł. Rozpropagowanie stosowanego w modelach Llama transparentnego, zdecentralizowanego procesu opartego na

szerokiej współpracy może wpłynąć na przyszłość generatywnej AI. Wspomniane modele są dostępne w wersjach o 7, 8 i 70 miliardach parametrów w usługach AWS, Google Cloud, Hugging Face i wielu innych.

Otwarty charakter tych modeli można traktować jako miecz obusieczny. Z jednej strony ułatwia dostęp do technologii, zwłaszcza mniejszym firmom i organizacjom. Nawet indywidualni programiści są w stanie dołożyć swoją cegiełkę do rozwoju modeli, a także mogą stosować je w prawdziwych projektach, np. aplikacjach biznesowych. Ten rodzaj zdecentralizowanej innowacji może prowadzić do przełomów, które nie wystąpiłyby w odseparowanym środowisku pojedynczej organizacji, a tym samym zwiększyć przyszłe możliwości tych modeli i rozszerzyć ich zastosowania.

Z drugiej strony, ta sama otwartość może powodować pewne ryzyko, np. przestępcy mogą korzystać z tych technologii do niecnych celów. Jest to jedna z obaw, które wyrażają organizacje takie jak OpenAI, sugerując, że pewien poziom kontroli dostępu i ograniczeń może w istocie pomóc w ograniczeniu możliwości tworzenia niebezpiecznych aplikacji.

Kwantyzacja i LoRA

Jednym z kluczowych aspektów modeli otwartych jest stosowanie pojęcia kwantyzacji (*https://oreil.ly/bkWXk*) i LoRA (*low-rank approximations*, aproksymacje niskiego rzędu — *https://oreil.ly/zORsB*). Obie techniki pozwalają na dopasowanie modeli do urządzeń o mniejszych możliwościach sprzętowych. Kwantyzacja pomaga zmniejszyć dokładność obliczeniową parametrów modelu, co wpływa na całkowity rozmiar modelu bez istotnej straty w jakości działania. LoRA z kolei pozwala optymalizować architekturę sieci, by uczynić ją bardziej wydajną w działaniu na sprzęcie konsumenckim.

Tego rodzaju optymalizacje sprawiają, że dostrajanie dużych modeli językowych staje się coraz bardziej dostępne na sprzęcie konsumenckim. Jest to niezwykle ważne, bo daje możliwość szerokiego prowadzenia eksperymentów na modelach i dostosowywania ich. Nie trzeba mieć już własnego centrum danych — zwykli programiści, małe firmy czy startupy mogą pracować z tymi modelami nawet w mocno ograniczonych środowiskach.

Mistral

Mistral 7B, wytwór francuskiego startupu Mistral AI (*https://mistral.ai*), jawi się jako prawdziwy kombajn w dziedzinie generatywnej sztucznej inteligencji — m.in. dzięki 7,3 miliarda parametrów. Nie tylko rozmiar ma tu znaczenie — jego siła tkwi w efektywności i zdolnościach, co pozwala mieć nadzieję na świetlaną przyszłość otwartych dużych modeli językowych i ich zastosowań. Kluczowa dla efektywności jest implementacja mechanizmu przesuwnego okna uwagi, opublikowana na otwartej licencji Apache. Wielu inżynierów AI dostraja swoje rozwiązania na podstawie tego modelu, wliczając w to imponujący model Zephr 7b beta (*https://oreil.ly/Lg6_r*). Warto rozważyć także model Mixtral 8x7b (*https://oreil.ly/itsJG*), czyli model mieszanki ekspertów (podobny do architektury GPT-4), który osiąga rezultaty zbliżone do GPT-3.5-turbo.

Więcej szczegółów i aktualne porównanie modeli otwartych, a także metryki wydajności znajdziesz na stronie Chatbot Arena Leaderboard (*https://oreil.ly/ttiji*), udostępnianej przez Hugging Face.

Anthropic: Claude

Model Claude 2 (*https://claude.ai/login*), opublikowany 11 lipca 2023 r., odróżnia się od innych dużych modeli językowych, takich jak ChatGPT czy LLaMA, za sprawą tzw. konstytucyjnego AI (ang. *constitutional AI*) związanego z bezpieczeństwem i nastawieniem (*https://oreil.ly/Tim9W*) — uczenie modelu odbywa się za pomocą listy reguł lub wartości. Istotnymi usprawnieniami w modelu Claude 2 jest rozszerzone okno kontekstu wielkości 100 000 tokenów, a także możliwość wysyłania plików. W świecie generatywnej sztucznej inteligencji okno kontekstu oznacza ilość tekstu (lub danych), którą model jest w stanie mieć na uwadze w trakcie generowania odpowiedzi. Im większe okno kontekstu, tym szerszy jest kontekst, na podstawie którego model wnioskuje i generuje odpowiedzi.

Te innowacje społeczność inżynierów AI przyjęła bardzo entuzjastycznie, pojawiły się bowiem dzięki temu nowe i bardziej wyszukane scenariusze użycia tego modelu. Na przykład możliwość przetwarzania większej ilości informacji pozwala na podsumowywanie dłuższych dokumentów lub prowadzenie bardziej rozbudowanych konwersacji. Ta przewaga nie trwała zbyt długo, ponieważ zaledwie pół roku później OpenAI opublikowała wersję modelu GPT-4 o wielkości okna kontekstu 128 000 tokenów (*https://oreil.ly/BWxrn*). Nie zmienia to faktu, że rywalizacja pozytywnie wpłynęła na rozwój całego sektora.

Kolejna generacja modelu Claude zawiera Opus (*https://oreil.ly/NH0jh*) — pierwszy model, który jest zdolny rywalizować z GPT-4 pod względem inteligencji, a także Haiku — mniejszy model, który jest niezwykle szybki, a także dość tani, ponieważ oferuje 1 000 000 tokenów w cenie 25 centów (w momencie gdy pisaliśmy te słowa, była to połowa kosztu użycia usługi GPT-3.5-turbo).

GPT-4V(ision)

Istotnym krokiem naprzód, poczynionym 23 września 2023 r., było rozszerzenie przez OpenAI modelu GPT-4 o możliwość przetwarzania obrazów za pomocą mechanizmu Vision. Innowację tę uwzględniono w interfejsie ChatGPT, który obecnie obsługuje jako wejście zarówno obrazy, jak i teksty. Ten krok pokazuje pewien szerszy trend, jakim są **modele multimodalne**, które mogą bez problemu przetwarzać i rozumieć różne rodzaje danych, np. obrazy i tekst, w ramach jednego kontekstu.

Porównanie modeli

Rynek modeli LLM w momencie, gdy piszemy te słowa, jest zdominowany przez OpenAI, a model GPT-4 powszechnie uznawano za lidera rynku. Najbliższym konkurentem jest Anthropic, a sporo nadziei budzą też mniejsze, ale otwarte modele, takie jak Llama i Mistral — zwłaszcza pod kątem

dostrajania. Chociaż zdaniem ekspertów OpenAI będzie dalej dostarczać najlepsze modele w przyszłości, to modele otwarte mogą okazać się *wystarczająco dobre* do coraz większej grupy zadań. To sprawi, że zadania związane z AI mogą być coraz częściej realizowane przez modele dostrajane lokalnie. Biorąc pod uwagę postępy w wydajności modeli i kwantyzację (metody używane w celu uzyskania kompromisu między dokładnością działania a rozmiarem i kosztem obliczeń), możemy kiedyś dojść do momentu, w którym duże modele językowe będzie można uruchamiać nawet na urządzeniach mobilnych.

Na razie najlepszą metodą porównania modeli jest uruchomienie tego samego prompta w wielu modelach i porównanie odpowiedzi. Jedną niezmienną regułą jest to, że GPT-4 radzi sobie znacznie lepiej w wykonywaniu instrukcji, co widać w poniższym przykładzie — jest to jedyny model, który wygenerował odpowiedź w prawidłowym formacie, z nazwami dopasowanymi do przykładów (zaczynających się od litery *i*).

Wejście:

```
Czy możesz wymyślić listę nazw dla butów, które mogą pasować do dowolnego rozmiaru stopy
w języku polskim, tak jakby zaproponował je Steve Jobs?

Zwróć wyniki w postaci listy przedzielonej przecinkami, w następującym formacie:
Opis produktu: But, który pasuje do dowolnego rozmiaru stopy
Nazwy produktów: [lista trzech nazw produktów]

## Przykłady
Opis produktu: Lodówka, która wydaje piwo
Nazwy produktów: iBarFridge, iFridgeBeer, iDrinkBeerFridge

Opis produktu: Zegarek, który podaje dokładny czas w przestrzeni kosmicznej
Nazwy produktów: iNaut, iSpace, iTime

Opis produktu: Automat do przygotowania mlecznych koktajli w domu
Nazwy produktów: iShake, iSmoothie, iShake Mini
```

Wyjście (GPT-4):

```
Opis produktu: But, który pasuje do dowolnego rozmiaru stopy
Nazwy produktów: iStep, iFitBoot, iComfortShoe
```

Wyjście (Claude 3):

```
Oto lista nazw dla butów, które pasują do dowolnego rozmiaru stopy, w stylu Steve'a
Jobsa:
Opis produktu: But, który pasuje do dowolnego rozmiaru stopy
Nazwy produktów: iMorph, iAdapt, iFlex
```

Wyjście (Llama 3 70b):

```
Oto lista nazw dla butów, które mogą pasować do dowolnego rozmiaru stopy w języku
polskim, w stylu Steve Jobsa:
Opis produktu: But, który pasuje do dowolnego rozmiaru stopy
Nazwy produktów: iFit, UniStep, AdaptWalk
```

Podsumowanie

Podróż dużych modeli językowych od ich powstania aż po obecną, zaawansowaną postać to opowieść o ciągłej innowacji, współpracy i intensywnej konkurencji. Wraz z ewolucją obecnych modeli zwiększa się prawdopodobieństwo, że staną się one integralną częścią naszego codziennego życia, zmieniając sposób, w jaki korzystamy z technologii, a nawet w jaki komunikujemy się ze sobą.

Zrozumienie kontekstu historycznego, a także możliwości tych modeli pozwoli Ci lepiej docenić narzędzia, jakimi dysponujemy — od inżynierii promptów aż po rozwój inteligentnych agentów wirtualnych. Warto pamiętać, że choć modele oferują ogromne możliwości, kluczową obawą pozostaje prywatność danych. Jeśli modele korzystają z Twoich danych do douczania lub dostrajania, należy powstrzymać się od wysyłania wrażliwych informacji.

W następnym rozdziale poznasz umiejętności niezbędne do stosowania w praktyce funkcjonalnych modeli językowych, takich jak GPT-4. Odkrywanie potencjału generatywnej sztucznej inteligencji z pewnością przyniesie Ci wiele ekscytujących obserwacji, jak również praktycznych umiejętności.

Skorowidz

D

E

F

G

H

I

J

K

L

Ł

M

N

O

P

R

S

T

U

W

Z

Absolutnie najlepsza książka na temat inżynierii promptów!
Dan Shipper, współzałożyciel i prezes Every

Udostępnienie dużych modeli językowych (ang. LLM, *large language model*) i modeli dyfuzyjnych, takich jak ChatGPT, Midjourney czy Stable Diffusion, zrewolucjonizowało wiele branż. Dzięki nim możliwa stała się realizacja szerokiego zakresu zadań, nawet tych, które do niedawna wydawały się niemożliwe do automatyzacji. Ponadto próg wejścia w świat AI jest bardzo niski, co sprawia, że niemal każdy może korzystać z modeli AI zgodnie ze swoimi potrzebami.

Z tą książką opanujesz podstawy generatywnej AI i nauczysz się skutecznie stosować jej modele w praktyce. Szczególną uwagę poświęcono integracji modeli językowych i dyfuzyjnych, co często bywa wyzwaniem, zwłaszcza w zakresie stabilności uzyskanych rozwiązań. Autorzy w jasny sposób wyjaśniają, jak za sprawą inżynierii promptów zapewnić niezawodność działania sztucznej inteligencji w środowiskach produkcyjnych. Co więcej, zaproponowane zasady są skonstruowane tak, aby bez trudu przetrwały próbę czasu i mogły być używane również dla przyszłych modeli!

Jeśli chcesz poprawić wiarygodność swoich systemów AI, musisz mieć tę książkę!
Mayo Oshin, założyciel i prezes Siennai Analytics

W książce:

- pięć uniwersalnych i perspektywicznych reguł promptowania
- korzystanie z generatywnej AI za pomocą bibliotek i frameworków, takich jak LangChain
- zalety i wady różnych modeli, w tym autorstwa OpenAI, i ich alternatyw
- praktyczne pisanie wysokiej jakości promptów w obszarze generowania tekstu, kodu i obrazów

James Phoenix tworzy stabilne potoki danych dla zespołów marketingowych, automatyzując tysiące powtarzalnych zadań. Wykładał *data science* w ramach ponad 60 bootcampów.

Mike Taylor współtworzył agencję marketingu wzrostowego Ladder. Prowadził niezwykle popularne kursy marketingu i AI w serwisach LinkedIn Learning, Udemy oraz Vexpower.

Obaj autorzy eksperymentowali z inżynierią promptów od 2020 roku, teraz pracują jako inżynierowie promptów.

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-1904-4
9 788328 919044
Cena: 99,00 zł