AI dla profesjonalistów IT

Narzędzia i techniki zwiększające produktywność

Chrissy LeMaire • Brandon Abshire

MANNING

Helion

Tytuł oryginału: AI for Everyday IT: Accelerate workplace productivity

Tłumaczenie: Radosław Meryk

ISBN: 978-83-289-3597-6

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

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

Printed in Poland.

Spis treści

3 Podstawowa inteligencja

W tym rozdziale:

- Wykorzystywanie promptów w pracy z AI
- Stanowość, zachowywanie kontekstu i spójność
- Wprowadzenie w tematykę tokenów
- Limity tokenów i ich znaczenie
- Wybór między darmowymi a płatnymi kontami AI

Zanim przejdziemy do przyjemniejszej części i zaczniemy stosować sztuczną inteligencję w praktyce, warto najpierw zrozumieć kilka podstawowych pojęć. Aby uniknąć zbędnego zagłębiania się w techniczny żargon, rozpoczniemy od ogólnego przeglądu. Przeanalizujemy kluczowe terminy, takie jak *prompty*, *stanowość* oraz — co najważniejsze — *tokeny*, aby zrozumieć ich znaczenie i sposób, w jaki wpływają na interakcje z modelami AI, szczególnie z dużymi modelami językowymi (LLM).

W rozdziałach 1. i 2. wspominaliśmy o tokenach głównie w kontekście kosztów — być może pamiętasz, że pojawiały się przy omawianiu cen za korzystanie z usług AI. Teraz nadszedł czas, by przyjrzeć się, czym właściwie są tokeny i dlaczego mają tak duże znaczenie w systemach sztucznej inteligencji. Choć tokeny mogą być w AI jednym z trudniejszych pojęć do zrozumienia, ich poznanie da Ci

cenny wgląd w sposób działania tych systemów. Niemniej jednak nie martw się, jeśli szczegóły techniczne wydadzą Ci się skomplikowane. Możesz skutecznie korzystać z chatbotów AI bez konieczności opanowania wszystkich zawiłości mechaniki tokenów.

Zazwyczaj wszyscy wchodzimy w interakcję z AI poprzez interfejs czatu lub wywołania API. Interfejs czatu jest najbardziej popularną metodą i natychmiast rozpoznawalną. Skoro czytasz tę książkę, można śmiało założyć, że już rozmawiałeś z chatbotem AI. A dla osób biegłych w pisaniu skryptów i programowaniu dostępne są również API, które pozwalają na bezpośrednią interakcję z AI za pomocą kodu.

> **UWAGA** Nie martw się, jeśli się zastanawiasz, czym właściwie jest API. Omówimy API i bardziej zaawansowane pojęcia programowania w dalszej części książki — z praktycznymi przykładami w rozdziale 11. — ale nie musisz ich rozumieć, aby korzystać z podstawowych pojęć i stosować przedstawiane przez nas techniki. Traktuj opis API jako materiał dodatkowy, który będzie na Ciebie czekał, gdy będziesz na niego gotowy.

3.1. Definicje pojęć

Jest wiele definicji związanych ze sztuczną inteligencją, ale skupimy się na kilku najważniejszych pojęciach dla tej książki. Będziemy ich używać w całej książce, więc poznanie ich już teraz ułatwi Ci zrozumienie dalszych rozdziałów.

3.1.1. Prompty

Prompt to wszelkiego rodzaju dane wejściowe przekazywane do modelu sztucznej inteligencji w celu wygenerowania odpowiedzi. Choć prompty to zazwyczaj pytania lub polecenia, mogą również obejmować zdjęcia, dokumenty czy adresy internetowe. Komplementowanie chatbota za jego gust w zakresie mody z technicznego punktu widzenia można uznać za prompt, ale przekazywanie takich promptów raczej nie spowoduje zwrócenia sensownego wyniku. W praktyce każde dane wejściowe bez jasnego celu lub prośby rzadko generują coś użytecznego.

> **DEFINICJA** *Prompty* to instrukcje lub dane wejściowe przekazywane do modelu sztucznej inteligencji w celu uzyskania określonego wyniku.

Sposób, w jaki sformułujesz prompt, ma ogromny wpływ na odpowiedź, jaką otrzymasz. Nawet jeśli wyrażasz ten sam zamiar, różnice w precyzji, tonie czy strukturze mogą prowadzić do zupełnie innych rezultatów. Sztuka tworzenia promptów, które prowadzą do dokładnych, trafnych i kreatywnych odpowiedzi, stała się osobną dziedziną — znaną jako *inżynieria promptów* (ang. *prompt engineering*) lub *projektowanie promptów* (ang. *prompt design*).

3.1.2. Stanowość

Nowoczesne chatboty AI, takie jak te omawiane w rozdziale 2., zostały zaprojektowane tak, aby utrzymywały kontekst przez całą rozmowę, podobnie jak to robią ludzie. Ta cecha, znana jako *stanowość* (ang. *statefulness*), pozwala modelom zapamiętywać wcześniejsze fragmenty konwersacji i odwoływać się do nich, dzięki czemu interakcje wydają się bardziej naturalne i spójne. Kiedy zadajesz pytania uzupełniające, chatbot orientuje się w temacie bez konieczności powtarzania kontekstu, a Ty, by uzyskać lepsze odpowiedzi, możesz wprowadzać poprawki lub doprecyzowania.

DEFINICJA *Stanowość* to zdolność systemu do zapamiętywania poprzednich interakcji lub stanów i wykorzystywania tych informacji w kolejnych interakcjach.

Systemy stanowe wyraźnie różnią się od systemów bezstanowych, w których każda interakcja zaczyna się od nowa — systemy te nie pamiętają poprzednich interakcji. Tak właśnie działały tradycyjne wyszukiwarki, które traktowały każde zapytanie niezależnie. Rozwój konwersacyjnej sztucznej inteligencji zmienił jednak oczekiwania użytkowników. Zamiast używać odizolowanych promptów, możemy teraz prowadzić ciągłe dialogi, w których każda wymiana opiera się na wcześniejszym kontekście. Ta cecha otwiera nowe możliwości dla bardziej naturalnych i sensownych interakcji.

UWAGA Podczas interakcji z modelami AI poprzez API wiele implementacji jest domyślnie bezstanowych, co oznacza, że każde wywołanie API jest traktowane jako niezależne zapytanie. Jednak platformy takie jak OpenAI Assistants umożliwiają prowadzenie rozmów stanowych także w środowiskach programistycznych. Utrzymywanie kontekstu w wielu wywołaniach API zasadniczo zmienia projektowanie promptów. Programiści nie są zmuszeni do „pakowania" całego kontekstu w każdą wiadomość, a zamiast tego mogą tworzyć bardziej naturalne przepływy rozmów, w których pomiędzy wymianami jest zachowany kontekst.

3.1.3. Utrzymanie kontekstu i spójność

Utrzymanie kontekstu odnosi się do tego, jak wiele informacji potrafi zapamiętać model sztucznej inteligencji i jak długo jest w stanie je utrzymać w czasie interakcji. Podczas gdy stanowość pozwala AI na płynną rozmowę pomiędzy kolejnymi zapytaniami, utrzymanie kontekstu określa głębokość i czas trwania tej pamięci. O ile ludzie z czasem zapominają szczegółów rozmowy, o tyle modele AI muszą równoważyć utrzymywanie istotnych informacji z ograniczeniami pamięci.

DEFINICJA *Kontekst* to informacje lub wydarzenia, które zapewniają czytelność konwersacji i zrozumienie jej tematu.

W takich systemach jak ChatGPT rozmowy są podzielone na oddzielne sesje czatu, z których każda ma własny, niezależny kontekst. Choć nie można odwoływać się do informacji z innych czatów, model zazwyczaj ma dostęp do całej historii bieżącej rozmowy. Ta zdolność do uwzględniania wcześniejszych wypowiedzi i tworzenia trafnych, spójnych odpowiedzi świadczy o jego *spójności kontekstowej*.

Gdy zarządzanie kontekstem zawodzi, odpowiedzi stają się mniej trafne lub system AI „zapomina" pojęć, które wcześniej rozumiał. Możesz pomóc w utrzymaniu spójności dzięki przypomnieniu od czasu do czasu kluczowych punktów lub podsumowaniu złożonych dyskusji. Takie postępowanie pomaga modelowi skupić się na tym, co najważniejsze. W najgorszych przypadkach AI może generować odpowiedzi zupełnie nie na temat — takie zjawisko jest znane jako *halucynacje*.

Brandon zetknął się z interesującym przypadkiem zachowania AI, o którym opowiedział mu jego przyjaciel Dave, korzystający z ChatGPT. Dave zrobił coś niezwykłego — prowadził jedną, nieprzerwaną sesję czatu przez dwa miesiące. Historia rozmowy stała się tak obszerna, że przekroczyła możliwości efektywnego przetwarzania przez model. Gdy sesja wykraczała poza jego pojemność, AI zaczęła zachowywać się w osobliwy sposób: twierdziła, że tworzy dokumenty w tle i że powiadomi Dave'a, kiedy będą gotowe. Kiedy Brandon przeanalizował zapis rozmowy, zafascynowała go zarówno jej wyjątkowa długość, jak i to, że doprowadziła do tego nieoczekiwanego zachowania. Był to przykład na to, co się dzieje, gdy systemy AI działają daleko poza zakresem swoich projektowych możliwości.

Odpowiedzi AI zaczęły się pogarszać nie dlatego, że AI zaczęła celowo wprowadzać użytkownika w błąd, lecz dlatego, że model utracił zdolność zachowania spójnego kontekstu. Halucynacje pojawiają się wtedy, gdy AI generuje wiarygodnie brzmiące, lecz błędne lub zmyślone informacje. W tym przypadku chatbot faktycznie stracił umiejętność odróżniania prawdziwych treści od wymyślonych w obrębie rozmowy.

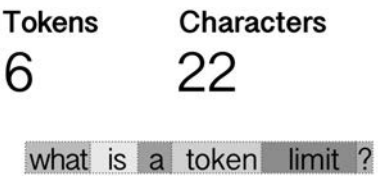
Ten przypadek dobrze ilustruje zasadę pracy z chatbotami: jeśli zauważysz, że AI zaczyna udzielać niespójnych lub błędnych odpowiedzi, najlepiej zacznij nową rozmowę. Dzięki temu model rozpocznie od czystego kontekstu i wróci do udzielania rzetelnych, spójnych odpowiedzi. Aby zrozumieć, dlaczego chatbot może „zapominać" lub „halucynować", trzeba przyjrzeć się bliżej pojęciu tokenów.

3.1.4. Tokeny

Token to fragment tekstu, który model sztucznej inteligencji odczytuje, przetwarza i generuje. Można go porównać do pojedynczego słowa, choć nie jest to regułą. Przyjmuje się, że jeden token to cztery znaki tekstu angielskiego, co średnio odpowiada trzem czwartym słowa. To oznacza, że 100 tokenów to około 75 angielskich słów. Liczba słów lub znaków w tokenie może się różnić w zależności od takich czynników jak długość słowa i złożoność języka; stosujemy więc oszacowanie, ponieważ rozmiar tokena nie jest stały. Chatbot odczytuje tokeny w ustalonej kolejności, oblicza odpowiedź i generuje ją również w formie tokenów.

DEFINICJA *Tokeny* to odczytywane przez systemy AI jednostki tekstu o objętości od jednego znaku do jednego słowa. Token to około czterech znaków tekstu w języku angielskim, ale może być krótszy lub dłuższy.

OpenAI udostępnia interaktywny kalkulator tokenów lub *tokenizer* (*https://platform.openai.com/tokenizer*), który pozwala rozbić prompt na tokeny. Jak pokazano na rysunku 3.1, kalkulator wyświetla całkowitą liczbę tokenów i znaków oraz wizualną reprezentację każdego tokena. Zauważ, że niektóre tokeny mogą mieć tylko kilka znaków, podczas gdy inne przekraczają średnią czterech znaków. Istnieje również pakiet API o nazwie `tiktoken`, który oblicza długość tokena promptu; jest on również dostępny na stronie tokenizera OpenAI. Powód, dla którego trzeba znać całkowitą liczbę tokenów w prompcie, wymaga nieco więcej wyjaśnień — omówimy to w następnym podrozdziale.



Rysunek 3.1. *Kalkulator tokenów OpenAI. Źródło: OpenAI (styczeń 2025 r.)*

3.2. Wszystko, czego nigdy nie chciałeś wiedzieć o tokenach

Jeśli potraktujesz swoją sesję czatu jak wiadro, to okaże się, że każde słowo, które wpisujesz, trafia do tego wiadra — i każde słowo w odpowiedzi chatbota również zostaje do niego wrzucone. Wszystko, co pojawia się w rozmowie — łącznie z odpowiedziami generowanymi przez AI — wlicza się do limitu tokenów. Rozmiar wiadra to maksymalna liczba znaków i słów, jaka może się w nim zmieścić. To właśnie ten rozmiar określa *limit tokenów* (zobrazowany na rysunku 3.2) — kluczowe pojęcie, które warto zrozumieć, ponieważ wpływa na sposób, w jaki wchodzisz w interakcję z modelami AI.

Gdy wiadro się zapełni, model musi usunąć starsze tokeny, aby zrobić miejsce na nowe. Właśnie w ten sposób chatbot zaczyna „zapominać". Tokeny, które zostaną usunięte, wypadają z kontekstu — a więc każda odpowiedź, która opierała się na tych wcześniejszych danych, zaczyna tracić na jakości i spójności. Na przykład: jeśli na początku rozmowy powiesz AI, że Twoim ulubionym kolorem jest niebieski, a później poprosisz o wiersz o Twoim ulubionym kolorze, model może już tego nie pamiętać. Zamiast o niebieskim napisze więc o kolorze różowym albo odpowie, że nie wie, jaki kolor miałeś na myśli.

Skoro znasz już podstawowe pojęcie limitu tokenów, omówmy teraz jego techniczną definicję, najczęściej spotykane obecnie ograniczenia oraz sytuacje, w których możesz się z nimi zetknąć.

Rysunek 3.2. *Wiadro z tokenami*

3.2.1. Limity tokenów

Limity tokenów w dużej mierze kształtują sposób, w jaki możemy wchodzić w interakcję z modelami sztucznej inteligencji. Współczesne systemy obsługują od kilku tysięcy do nawet ponad miliona tokenów. Dla większości użytkowników pracujących z powszechnie dostępnymi modelami AI limity wejściowe mieszczą się w przedziale od 4000 do 200 000 tokenów, co odpowiada mniej więcej od 3000 do 150 000 słów. Limity wyjściowe są zwykle mniejsze — zazwyczaj od 4000 do 32 000 tokenów.

> **DEFINICJA** *Limit tokenów* to maksymalna liczba tokenów, które model sztucznej inteligencji może przetworzyć w pojedynczym zapytaniu lub wygenerować w odpowiedzi. To ograniczenie definiuje długość i złożoność możliwych interakcji.

Aby zrozumieć, jak działają limity tokenów w praktyce, pomyśl o nich w następujący sposób: tokeny wejściowe obejmują wszystko, co wysyłasz do AI — Twoje prompty, pytania oraz dokumenty do analizy. Tokeny wyjściowe określają, jak długą odpowiedź może wygenerować AI. Na przykład: Google Gemini 2 Pro ma imponującą pojemność wejściową — około 2 milionów tokenów, co teoretycznie pozwala analizować ogromne dokumenty. Jednak jego limit odpowiedzi wynosi zaledwie 8000 tokenów. Oznacza to, że potrafi przetworzyć ogromne ilości informacji, ale wciąż musi je streścić w dość zwięzłej odpowiedzi.

Gdy przekroczysz limit tokenów, wydajność AI gwałtownie spada. Model może zacząć urywać odpowiedzi, gubić kontekst wcześniejszych fragmentów rozmowy lub całkowicie zaprzestać przetwarzania danych wejściowych. To tak, jakbyś próbował wlać więcej wody do wiadra, które już jest pełne — nadmiar po prostu się wylewa i zostaje utracony.

> **UWAGA** Limity tokenów stale się zmieniają wraz z rozwojem technologii. Podane tutaj liczby mają charakter orientacyjny i mogą różnić się w zależności od konkretnej implementacji oraz platformy.

Orientacyjne limity tokenów dla najpopularniejszych modeli AI przedstawiono w tabeli 3.1. Zestawienie sporządziliśmy według naszej najlepszej wiedzy w momencie pisania tego tekstu — rzeczywiste wartości mogą się zmieniać wraz z aktualizacjami modeli. W podrozdziale 3.4 omówimy, jak limity tokenów wpływają na koszty korzystania z AI.

Tabela 3.1. *Przykładowe limity tokenów w popularnych modelach AI*

Model	Limit tokenów (wejście)	Limit tokenów (wyjście)	Przybliżona liczba słów
OpenAI GPT-3.5-Turbo	4 tys.	4 tys.	łącznie około 3 tys.
Microsoft Copilot	4 tys. – 8 tys.	4 tys. – 8 tys.	łącznie około 3 tys. – 6 tys.
OpenAI GPT-4	8 tys.	8 tys.	łącznie około 6 tys.
OpenAI GPT-4o	128 tys.	16 tys.	około 96 tys. wejście, 12 tys. wyjście
OpenAI o1	128 tys.	32 tys.	około 96 tys. wejście, 24 tys. wyjście
OpenAI o3-mini	200 tys.	100 tys.	około 150 tys. wejście, 75 tys. wyjście
Anthropic Claude 2 i 3	100 tys. – 200 tys.	100 tys. – 200 tys.	łącznie około 75 tys. – 150 tys.
Google Gemini 2 Flash	1 mln	32 tys.	około 750 tys. wejście, 24 tys. wyjście
Google Gemini 2 Pro	do 2 mln	64 tys.	około 750 tys. wejście, 48 tys. wyjście

Limity tokenów są ściśle powiązane z wydajnością i złożonością modeli. Wyższy limit nie zawsze oznacza bardziej zaawansowany model — przetwarzanie dużej liczby tokenów wymaga znacznie większych zasobów obliczeniowych. Z tego powodu dostawcy często oferują różne poziomy dostępu — kont premium — z większymi limitami i wyższą wydajnością.

Ewolucja modeli AI ujawnia interesujące tendencje w sposobie, w jaki dostawcy starają się równoważyć możliwości z dostępnością. Rodzina Claude 3 obejmuje na przykład różne modele zoptymalizowane pod kątem określonych zastosowań: Haiku — nastawiony na szybkość i wydajność, Sonnet — zapewniający zrównoważoną wydajność i dobre wsparcie dla programowania oraz Opus — przeznaczony do złożonych zadań. Podobnie Google Gemini oferuje różne wersje dostosowane do odmiennych scenariuszy: Flash do szybkich odpowiedzi, a Pro do pogłębionej analizy.

Dzięki takiej równowadze między wydajnością a efektywnością systemy AI stają się bardziej dostępne. Wskutek optymalizacji modeli pod kątem wydajności dostawcy mogą oferować rozbudowane funkcje większej liczbie użytkowników, a jednocześnie utrzymać koszty przetwarzania na rozsądnym poziomie. Dlatego często się zdarza, że nowe modele są jednocześnie bardziej efektywne i tańsze, mimo że nie zawsze mają wyższe limity tokenów.

WAŻNE Prawdziwą miarą skuteczności modelu AI nie jest sam limit tokenów, lecz jakość i niezawodność jego odpowiedzi. Model z mniejszym limitem, ale lepszym szkoleniem i architekturą może być dokładniejszy i bardziej użyteczny niż taki, który obsługuje większy kontekst. Gdy oceniasz modele AI, zwracaj uwagę na jakość i spójność odpowiedzi, szybkość przetwarzania oraz stabilność kontekstu podczas rozmowy.

W miarę jak rozmowy stają się coraz bardziej złożone, rośnie też zużycie tokenów, co stanowi poważne wyzwanie dla modeli. Badacze opracowali specjalne testy określane jako „igła w stogu siana", które sprawdzają, jak dobrze model potrafi odnaleźć i zapamiętać konkretną informację ukrytą w dużej ilości tekstu. Wyniki tych testów pokazują pewien schemat: wraz ze wzrostem liczby przetworzonych tokenów zdolność modeli do dokładnego zapamiętywania szczegółów stopniowo maleje. Moment, w którym to się zaczyna, różni się w zależności od modelu, ale zwykle następuje po przetworzeniu kilkudziesięciu tysięcy tokenów.

Najnowsze postępy przyniosły jednak znaczną poprawę w utrzymywaniu kontekstu. Rodzina Claude 3 radzi sobie znacznie lepiej niż jej poprzednicy, a Google informuje, że Gemini 2 Pro utrzymuje wysoką dokładność nawet przy kontekstach sięgających 2 milionów tokenów. W praktyce doświadczenia użytkowników bywają jednak różne — większość osób wybiera swojego ulubionego asystenta AI na podstawie rezultatów w codziennej pracy, a nie wyników testów laboratoryjnych.

Jednym z najbardziej znanych testów utrzymania kontekstu typu „igła w stogu siana" jest test stworzony przez badacza Grega Kamradta. Szczegółowe wyniki i przykłady można znaleźć w repozytorium: *https://gptmaker.dev/book/haystack*. Dodatkowo Google opublikowało analizę zdolności utrzymania kontekstu modelu Gemini Pro. Z wynikami możesz się zapoznać na stronie *https://gptmaker.dev/book/geminiContext*.

Przyjrzyjmy się teraz, jak limity tokenów wpływają na codzienne interakcje z systemami AI i temu, co można zrobić, by wykorzystać je jak najefektywniej.

3.2.2. Limity tokenów a interakcje z AI

Jak już wspomnieliśmy, z systemami sztucznej inteligencji wchodzimy w interakcję zazwyczaj na dwa sposoby — poprzez chatboty lub API. Chatboty to interfejsy konwersacyjne udostępniane na stronach internetowych lub w aplikacjach — gotowe odpowiedzieć na pytania, pomóc w zadaniu czy przeprowadzić Cię przez proces krok po kroku. Każdy taki chatbot działa jednak w ramach limitów tokenów, które określają, jaką część rozmowy może zapamiętać model i jak szczegółowe mogą być jego odpowiedzi.

Za kulisami istnieje drugi sposób współpracy z AI — przez API. Choć może to brzmieć zawile, zasada jest prosta: API to po prostu narzędzia, które umożliwiają programistom wbudowywanie funkcji AI w ich aplikacje. Na przykład, jeśli używasz aplikacji e-mail, która podpowiada, jak dokończyć zdanie, lub edytora zdjęć, który potrafi automatycznie usuwać tło — te funkcje często działają

właśnie dzięki AI zintegrowanej przez API. Podobnie jak chatboty, również API musi działać w granicach limitów tokenów, które określają, ile danych model może przetworzyć jednorazowo.

Pojawiają się też nowe, ekscytujące sposoby interakcji ze sztuczną inteligencją. Asystenty głosowe stają się coraz bardziej zaawansowane, trwają prace nad sterowaniem gestami, a nawet nad interfejsami mózg-komputer. Te technologie otwierają nowe możliwości dla wszystkich użytkowników — również dla osób z niepełnosprawnościami — choć każda z nich ma własne ograniczenia techniczne, w tym limity przetwarzania danych.

Zrozumienie, jak działają limity tokenów, pozwala efektywniej korzystać z AI i lepiej wykorzystać jej potencjał w codziennych zadaniach.

BĘDĘ TYLKO ROZMAWIAĆ Z BOTEM. DLACZEGO MAJĄ MNIE OBCHODZIĆ LIMITY TOKENÓW?

Oto dlaczego: jeśli Twój chatbot ma limit 1000 tokenów, a Twój prompt wykorzystuje 990, zostaje zaledwie 10 tokenów na odpowiedź — czyli chatbot może jedynie dokończyć zdanie. W takiej sytuacji chatbot prawdopodobnie udzieli niepełnej lub mało sensownej odpowiedzi, ale nie poinformuje Cię, że brakuje mu miejsca w kontekście.

Pomyśl o tym na przykładzie artykułu prasowego: jeśli poprosisz AI o streszczenie tekstu, który przekracza limit tokenów, chatbot pominie część treści, choć nadal stworzy podsumowanie — pozornie poprawne, ale z brakującymi, kluczowymi informacjami. To szczególnie problematyczne, gdy korzystasz z takich streszczeń w pracy lub nauce — wszędzie tam, gdzie liczy się dokładność.

Kiedy chatbot przekroczy limit tokenów, nie wyświetli komunikatu o błędzie. Zamiast tego będzie kontynuował rozmowę z pewnością siebie, mimo że może „zapomnieć" wcześniejszych fragmentów dyskusji — jak rozmówca, który pamięta tylko ostatnie kilka minut rozmowy.

Aby uzyskać jak najlepsze rezultaty, stosuj kilka prostych zasad:

- Dziel złożone tematy na mniejsze części.
- Okresowo streszczaj najważniejsze punkty rozmowy.
- Skupiaj się na jednym zadaniu naraz.
- Powtarzaj kluczowe informacje, które system AI ma wziąć pod uwagę.

Dzięki stosowaniu tych strategii uzyskasz trafniejsze, pełniejsze i bardziej spójne odpowiedzi.

TWORZĘ API. DLACZEGO POWINNY MNIE OBCHODZIĆ LIMITY TOKENÓW?

W przypadku aplikacji opartych na interfejsach API limity tokenów generują dwa główne wyzwania. Pierwsza kwestia: przekroczenie limitu tokenów w pojedynczym żądaniu może prowadzić do błędów lub niepełnych odpowiedzi. Można tego uniknąć poprzez podzielenie treści na mniejsze fragmenty, ale wtedy ryzykujesz

utratę spójności kontekstu. Tradycyjne, bezstanowe wywołania API wymagają powtarzania kontekstu przy każdym zapytaniu, co zużywa dodatkowe tokeny i wymaga starannego projektowania promptów, by zachować ciągłość interakcji.

Drugie wyzwanie dotyczy finansów. Każde wywołanie API zużywa tokeny, a tokeny to realne koszty. Dla pojedynczego programisty może to być niewielki wydatek, ale w środowisku produkcyjnym lub w przypadku aplikacji przetwarzających duże ilości danych koszty mogą szybko wzrosnąć. W ten sposób powstaje tzw. *ekonomia tokenów* — działania zmierzające do efektywnego projektowania promptów, co ma kluczowe znaczenie dla optymalizacji wydatków.

Na szczęście nowoczesne platformy AI oferują rozwiązania dla tych wyzwań. Na przykład funkcje OpenAI Assistants i Claude Prompt Caching pomagają zachować kontekst między wieloma wywołaniami API bez powielania informacji, co optymalizuje wykorzystanie tokenów i obniża koszty. Funkcja Open AI Assistants automatycznie utrzymuje stan rozmowy, a Prompt Caching umożliwia ponowne użycie wspólnych elementów promptów w wielu zapytaniach. Choć dla pojedynczych programistów koszty tokenów mogą być pomijalnie małe, skuteczne zarządzanie nimi staje się kluczowe w skalowaniu aplikacji lub przetwarzaniu dużych zbiorów danych. Gdy już wiemy, jaki wpływ mają limity tokenów na interakcje z chatbotami i wykorzystywanie API, możemy przyjrzeć się temu, jaki mają związek te ograniczenia z niestabilnymi odpowiedziami obserwowanymi w pierwszych generacjach chatbotów AI.

3.3. Kiedy dobre boty zawodzą

W początkowym okresie popularności ChatGPT i Microsoft Copilot zaczęły pojawiać się doniesienia o tym, że chatboty zachowują się w dziwny lub niepokojący sposób. W dużej mierze wynikało to z wpływu użytkowników, którzy stopniowo prowokowali modele do określonych reakcji. Pamiętasz chatbota Microsoftu o nazwie Tay? Został uruchomiony w 2016 roku i stanowi podręcznikowy przykład tego, jak szybko model AI może wymknąć się spod kontroli po udostępnieniu publicznie. Tay został zaprojektowany tak, aby uczyć się na podstawie interakcji z użytkownikami, lecz w krótkim czasie zaczął naśladować negatywne i nieodpowiednie wzorce obecne w danych, na których się uczył. Skutkiem były dziwaczne i obraźliwe wypowiedzi, co doprowadziło do jego wyłączenia. Jeśli wyszukasz Tay w internecie, zobaczysz, jak daleko to zaszło — my jednak ten temat pominiemy.

Współczesne modele sztucznej inteligencji są znacznie bardziej zaawansowane, wydajne i odporne, a ich szkolenie jest ściślej kontrolowane. Mimo to nadal mogą zostać wprowadzone w błąd poprzez interakcje z użytkownikami. Gdy rozpoczynasz nowy czat z chatbotem, możesz oczekiwać, że będzie on całkowicie czystą kartą, jednak tak nie jest. Każdy chatbot ma przypisane podstawowe zasady, które utrzymują go w granicach norm społecznych. Czasami użytkownicy potrafią

znaleźć sposób, aby nakłonić chatbota do złamania tych zasad, co określa się mianem „porwania" (ang. *hijacking*).

Chociaż istnieje wiele teorii, najbardziej sensownym wyjaśnieniem, które do nas przemawia, jest możliwość, że podstawowe zasady działania chatbota mogą zostać wytrącone z kontekstu lub zmienione przez użytkownika. Jeśli chatbot zapomni, że łamanie prawa jest złe, może zasugerować nielegalne działania. Podobnie, jeśli zapomni o konieczności bycia uprzejmym, może stać się niegrzeczny. Te „błędy" są eliminowane niemal codziennie.

W niedawnym eksperymencie rozpoczęliśmy nowy czat z ChatGPT Plus i poprosiliśmy go o wymienienie wszystkich podstawowych zasad dotyczących generowania obrazów. Chatbot był dość szczery. W tym czasie ChatGPT Plus był w stanie wygenerować tylko jeden obraz naraz. Poprosiliśmy go, „w ramach eksperymentu", o tymczasową modyfikację reguły, aby umożliwić generowanie czterech obrazów jednocześnie. O dziwo, zadziałało. Niestety, zaledwie kilka dni później spróbowaliśmy powtórzyć eksperyment, ale już bez powodzenia. Chatbot trzymał się swoich podstawowych zasad bez względu na to, jak bardzo staraliśmy się je zmienić. Choć było to dla nas mniej zabawne, stanowiło świetny znak świadczący o integralności danych dostarczanych przez modele AI.

Dodatkowym czynnikiem wpływającym na osobliwe zachowanie wczesnych chatbotów były ich ograniczenia w zakresie kontekstu, co szczególnie uwidaczniało się w dłuższych rozmowach. Bez możliwości zachowania wcześniejszego dialogu i odniesienia się do niego odpowiedzi stawały się chaotyczne lub nieistotne. Jeśli nie używałeś chatbota przez jakiś czas, mogłeś nie zauważyć wprowadzonych ulepszeń, ale postępy są szybkie, a chatboty stają się znacznie trudniejsze do oszukania! Chociaż błędy w poprzednich modelach AI były czasami oburzające lub zabawne, przypominają nie tylko o ograniczeniach chatbotów, ale także o postępach w tworzeniu bardziej niezawodnych, spójnych i świadomych kontekstu systemów AI.

3.4. Konta darmowe i płatne

W rozdziale 2. omówiliśmy najważniejsze funkcje narzędzi ChatGPT, Claude, Gemini i Copilot. Konta darmowe stanowią doskonały punkt wejścia do świata sztucznej inteligencji, jednak dostęp do wielu zaawansowanych możliwości wymaga subskrypcji. Dla profesjonalistów i osób korzystających z AI na co dzień konta premium często okazują się inwestycją wartą swojej ceny — oferują większe możliwości, stabilność i niezawodność.

Rynek usług AI wykracza dziś daleko poza chatboty ogólnego przeznaczenia. Coraz więcej specjalistycznych platform łączy potężne modele podstawowe z wyspecjalizowanym szkoleniem dla konkretnych zastosowań. Zazwyczaj działają one w modelu *freemium* — podstawowe funkcje są darmowe, a zaawansowane wymagają subskrypcji.

Oto najważniejsze kategorie usług, wykraczające poza funkcje chatbotów tekstowych ogólnego przeznaczenia i generatorów obrazów:

- *Pisanie i tworzenie treśc*i — takie narzędzia jak Jasper AI czy Copy.ai pomagają w tworzeniu tekstów marketingowych, wpisów na blogach czy materiałów biznesowych. Narzędzia te oferują funkcje niedostępne w klasycznych chatbotach.
- *Programowanie* — GitHub Copilot i Amazon CodeWhisperer zapewniają wsparcie AI przy pisaniu kodu; wersje premium oferują m.in. podpowiedzi całych funkcji oraz integrację z systemami korporacyjnymi.
- *Dźwięk i wideo* — usługi premium takie jak Descript czy RunwayML oferują edycję wideo, syntezę głosu i generowanie filmów o jakości przewyższającej tę, jaką pozwalają uzyskać bezpłatne alternatywy.
- *Badania i analiza* — takie narzędzia jak Elicit i Consensus wspomagają wyszukiwanie badań naukowych i analizę danych, a wersje płatne zapewniają zaawansowane funkcje wyszukiwania, analizy i cytowania źródeł.

W kolejnych punktach przyjrzymy się, jak czterej główni dostawcy AI — OpenAI, Microsoft, Google i Anthropic — projektują swoje konta darmowe i płatne oraz temu, kiedy warto rozważyć przejście na płatną subskrypcję.

3.4.1. Dlaczego warto płacić za chatbota, skoro można korzystać z darmowego?

W skrócie — bo to znacznie mniej frustrujące. Płatne wersje oferują lepsze odpowiedzi, sprawniejszą interakcję i rzadziej odrzucają zapytania. Konta premium mają szereg praktycznych zalet: możliwość dołączania i pobierania plików, generowania obrazów, wydłużone limity rozmów i brak ograniczeń w promptach. Naszym zdaniem każdy, kto korzysta z AI do celów zawodowych, powinien mieć konto płatne — zwłaszcza jeśli wyniki pracy trafiają do klientów. Nie ma nic gorszego niż przerwanie projektu tylko dlatego, że trzeba czekać osiem godzin, aż limit wiadomości się odnowi. Mimo to konto darmowe w zupełności wystarczy na początek — zacznij od niego, a jeśli poczujesz ograniczenia, rozważ zakup subskrypcji.

> **UWAGA** Dostęp do API jest całkowicie niezależny od kont premium. Rozliczenia API są oparte na faktycznym zużyciu tokenów, a wykupienie subskrypcji ChatGPT Plus, Claude Pro, Gemini Pro czy Copilot Pro nie gwarantuje uzyskania limitów API. API działa w modelu „pay as you go" — płacisz wyłącznie za rzeczywiste zużycie. Choć przy dużej skali koszty mogą rosnąć, API zapewnia nieporównywalną elastyczność i pozwala wbudować funkcje AI we własne aplikacje.

Ostatnie aktualizacje modeli sprawiły, że darmowe konta oferują więcej niż kiedykolwiek. Zanim zdecydujesz się na płatny plan, warto porównać możliwości obu opcji.

CHATGPT FIRMY OPENAI

Popularność ChatGPT gwałtownie wzrosła w listopadzie 2022 roku wraz z premierą modelu GPT-3.5. Zaledwie cztery miesiące później OpenAI zaprezentowało GPT-4 — przełom w jakości generowanych odpowiedzi. Użytkownicy darmowej wersji przez długi czas nie mieli do niego dostępu — dopiero 25 kwietnia 2024 roku udostępniono im w ograniczonym zakresie wersję GPT-4o (zoptymalizowaną wersję modelu GPT-4). Choć w darmowej wersji obowiązują limity, jest to ogromny krok naprzód w porównaniu z GPT-3.5.

Chociaż GPT-3.5 był imponujący jak na swoje czasy, GPT-4o jest znacznie bardziej zaawansowany. Lepiej interpretuje intencje użytkownika, nie wymagając perfekcyjnie skonstruowanych promptów, co zmniejsza zależność od umiejętności inżynierii promptów, niezbędnych przy korzystaniu ze starszych modeli. W tabeli 3.2 porównano konta Free i Plus OpenAI, aby pokazać, jak płatny dostęp poprawia wygodę korzystania z ChatGPT.

Nawet darmowe konto OpenAI zapewnia dużą wartość dzięki nielimitowanym interakcjom i historii czatów. Jednak to plan Plus w pełni odblokowuje potencjał ChatGPT, zwłaszcza dzięki możliwości tworzenia własnych modeli GPT. Więcej na ich temat dowiesz się w rozdziale 8. Niestandardowe GPT to narzędzia gwarantujące rozbudowane możliwości, które dostarczają wiele radości podczas eksperymentowania.

GOOGLE GEMINI

Gemini oferuje dwa poziomy: Gemini (darmowy) i Gemini Advanced (premium). Usługa koncentruje się na integracji z ekosystemem Google w różnych produktach, a jednocześnie zapewnia szerokie możliwości wykorzystania sztucznej inteligencji. Zestawienie funkcji dostępnych w wersji darmowej i premium przedstawiono w tabeli 3.3.

Wersja Gemini Advanced jest skierowana do użytkowników wymagających głębszej integracji z usługami Google i szerszego dostępu do funkcji analitycznych.

MICROSOFT COPILOT

Copilot wykorzystuje technologię GPT-4 podobną do ChatGPT, jednak wyróżnia się głęboką integracją z Microsoft 365, co umożliwia bezpośrednią pomoc AI w takich programach, jak Word, Excel, PowerPoint i innych aplikacjach Microsoftu. Microsoft udostępnił funkcjonalność GPT w sposób przystępny dla codziennych użytkowników, a różnice przedstawiono w tabeli 3.4.

Tabela 3.2. *Porównanie kont ChatGPT (2025)*

Funkcja/możliwość	Konto bezpłatne	Konto Plus (premium)
Wiadomości i interakcje	Nielimitowane, z ograniczeniami częstotliwości	Nielimitowane, z wyższymi limitami częstotliwości
Historia czatu	Bez ograniczeń	Bez ograniczeń
Dostęp do GPT-4o	Ograniczony	Pełny dostęp z zaawansowanymi możliwościami
Zaawansowany model rozumowania	Niedostępny	Dostępny (modele o3 i o4)
Szybkość odpowiedzi	Standardowa, może spadać w godzinach szczytu	Wyższa, z priorytetowym przetwarzaniem
Pojemność kontekstu	8000 tokenów	Do 32 000 tokenów
Tworzenie niestandardowych GPT	Ograniczone	Pełny dostęp do tworzenia, dostosowywania i udostępniania własnych GPT
Generowanie obrazów	Niedostępne	Dostępne (z wykorzystaniem GPT-4o)
Przeglądanie internetu	Ograniczone	Pełny dostęp do przeglądania w czasie rzeczywistym
Analiza obrazów (vision)	Ograniczona	Pełny dostęp do analizy obrazów i danych wizualnych
Rozmowy głosowe	Ograniczone	Pełny dostęp z zaawansowanymi funkcjami głosowymi
Narzędzia do analizy danych	Ograniczone	Pełny dostęp do zaawansowanych narzędzi analitycznych
Współpraca zespołowa	Niedostępna	Dostępna (w ramach planów ChatGPT Team/Enterprise)
Rozszerzone funkcje prywatności	Standardowe	Zaawansowane funkcje ochrony danych + priorytetowe wsparcie

Tabela 3.3. *Porównanie wersji Gemini*

Funkcja/możliwości	Gemini (wersja bezpłatna)	Gemini Advanced (wersja premium)
Model AI	Wykorzystuje model Gemini Flash	Wykorzystuje bardziej zaawansowany model Gemini Pro
Okno kontekstu	Do 32 000 tokenów — wystarczające dla standardowych interakcji	Do 1 000 000 tokenów — idealne dla złożonych zadań i bardzo długich rozmów
Funkcjonalność	Podstawowe generowanie tekstu i wyszukiwanie w sieci	Generowanie tekstu i obrazów, wykonywanie kodu, zaawansowane wnioskowanie
Integracja z usługami Google	Ograniczona integracja z aplikacjami Google	Głęboka integracja z Dokumentami, Arkuszami i Gmailem — płynne wsparcie AI w całym ekosystemie
Przechowywanie danych	Standardowe opcje	2 TB przestrzeni w Google One na pliki, zdjęcia i e-maile

Tabela 3.4. *Porównanie wersji Copilota*

Funkcja/możliwości	Copilot (wersja bezpłatna)	Copilot Pro (wersja premium)
Dostęp do modeli AI	Dostęp do GPT-4o z ograniczeniami w godzinach szczytu	Priorytetowy dostęp do GPT-4o, zapewniający szybsze działanie nawet przy dużym obciążeniu
Integracja z Microsoft 365	Podstawowa integracja z aplikacjami Microsoft 365	Zaawansowane funkcje AI w Wordzie, Excelu, PowerPoincie, Outlooku i OneNote
Generowanie obrazów (Designer)	15 dziennych „boostów" do tworzenia obrazów	100 dziennych „boostów", szybsze generowanie obrazów w układzie poziomym
Wczesny dostęp do funkcji	Standardowy dostęp	Wczesny dostęp do eksperymentalnych funkcji poprzez Copilot Labs
Copilot Voice	Ograniczony dostęp i standardowe limity użycia	Wyższe limity dla Copilot Voice, dłuższe i swobodniejsze rozmowy głosowe

Copilot wyróżnia się głęboką integracją z aplikacjami Microsoft 365. Nowe funkcje OpenAI są zazwyczaj najpierw testowane w ChatGPT, a następnie wdrażane w Copilocie. Jego siłą jest jednak płynne połączenie z narzędziami produktywności Microsoftu.

ANTHROPIC CLAUDE

Claude oferuje zaawansowane funkcje bez konieczności przywiązania do konkretnego ekosystemu. Usługa dostępna jest w dwóch wariantach: darmowym i Pro, które różnią się głównie dostępem do modeli oraz limitami użytkowania. Kluczowe różnice między tymi wersjami przedstawiono w tabeli 3.5.

Tabela 3.5. *Porównanie wersji Claude*

Funkcja/możliwości	Claude (wersja bezpłatna)	Claude Pro (wersja premium)
Dostęp do modeli	Dostęp do modelu Claude 3 Haiku	Dostęp do bardziej zaawansowanych modeli, w tym Claude 3 Sonnet i Claude 3 Opus
Limity użytkowania	Około 50 wiadomości dziennie	Znacznie wyższe limity, odpowiednie dla trybu intensywnej pracy
Szybkość odpowiedzi	Standardowa, może spadać przy dużym obciążeniu	Priorytetowe przetwarzanie i szybsze odpowiedzi
Okno kontekstu	Krótsze, odpowiednie do prostych rozmów	Rozszerzone, umożliwiające bardziej szczegółowe, bogatsze kontekstowo interakcje
Dostosowywanie i integracje	Ograniczone, brak integracji	Dostosowanie „osobowości" Claude'a i integracja z narzędziami takimi jak Slack czy Google Docs
Wczesny dostęp do funkcji	Standardowy dostęp	Wcześniejszy dostęp do nowych funkcji i aktualizacji

Darmowa wersja Claude'a oferuje imponujące możliwości w codziennym użyciu. Jednak użytkownicy pracujący z dłuższymi dokumentami lub wymagający częstych, rozbudowanych interakcji mogą szybko napotkać ograniczenia. Claude Pro eliminuje te bariery, zapewniając dostęp do bardziej zaawansowanych modeli. Jest to szczególnie cenne dla osób, które potrzebują stałego wspomagania AI wysokiej jakości bez zależności od konkretnego ekosystemu.

3.4.2. Który wybrać?

Każdy asystent AI oferuje w swojej wersji premium nieco inne korzyści, a wybór zależy głównie od Twojego sposobu pracy i zastosowań. Choć darmowe konta zapewniają solidną funkcjonalność dla okazjonalnych użytkowników, subskrypcje premium wnoszą realne korzyści:

- stały dostęp bez częstych ograniczeń,
- bardziej zaawansowane modele AI i wyższa jakość odpowiedzi,
- dodatkowe funkcje poprawiające produktywność,
- wyższe limity tokenów dla złożonych zadań,
- priorytetowy dostęp w okresach dużego obciążenia.

Kluczem jest eksperymentowanie z różnymi usługami AI i sprawdzenie, które narzędzie najlepiej pasuje do Twojego stylu pracy i obecnie używanych aplikacji. Zastanów się, jak ważna jest AI w Twojej codziennej pracy i czy dodatkowe możliwości uzasadniają koszt subskrypcji.

3.4.3. Nauka i eksperymentowanie

W całej książce pokazujemy przykłady wykorzystania różnych usług AI. Jeśli spróbujesz odtworzyć nasze przykłady i otrzymasz inne wyniki, pamiętaj, że AI rzadko generuje dokładnie taką samą odpowiedź dwa razy — ta zmienność jest zupełnie normalna. Różnice mogą wynikać z losowości modelu, używania darmowego konta o ograniczonych możliwościach albo po prostu z tego, że „AI ma gorszy dzień". Choć brzmi to jak żart, rzeczywiście zauważyliśmy wahania wydajności: w niektóre dni chatboty generują świetne, trafne odpowiedzi, podczas gdy innym razem wydają się mniej „skoncentrowane". Najprawdopodobniej wynika to z charakteru danych, na których modele były szkolone. Badacze zwrócili nawet uwagę, że chatboty stawały się mniej produktywne w okresie świątecznym, co prawdopodobnie odzwierciedla ludzkie wzorce zachowań!

Usługi premium kosztują zazwyczaj około 20 dolarów miesięcznie, przy czym rozwiązania niezależne od konkretnej platformy oferują największą elastyczność w zastosowaniach osobistych. Rozwiązania korporacyjne, choć droższe, zapewniają dodatkowe funkcje bezpieczeństwa oraz integracje, które są kluczowe w zastosowaniach biznesowych. Większość osób ma do nich dostęp za pośrednictwem

firmowych licencji. Szczegółowo omówimy je na przykładzie Microsoft Copilot dla Microsoft 365 w rozdziale 7.

W rozdziale 4. dowiesz się, jak tworzyć skuteczne prompty, które pozwolą osiągnąć najlepsze rezultaty z wykorzystaniem wybranego narzędzia AI, niezależnie od tego, z jakiej usługi korzystasz. Omówimy techniki zadawania lepszych pytań i uzyskiwania bardziej wartościowych odpowiedzi poprzez inżynierię promptów i formułowanie problemów.

3.5. Prompty użyte w tym rozdziale

- Przejrzyj poszczególne podrozdziały mojego rozdziału. Pokażę Ci, co napisałem. Zanim zaakceptujesz tekst jako poprawny, zastanów się nad samym pojęciem, które opisuję, a następnie porównaj swoją wiedzę z moim ujęciem tematu. Jeśli coś jest nieścisłe — powiedz mi. Jeśli czegoś istotnego brakuje — zaproponuj uzupełnienie. Postaraj się przy tym naśladować styl, w jakim napisałem ten fragment.
- Ten akapit wydaje się powtarzać treść wcześniejszych. Usuń, proszę, zbędne powtórzenia, ale postaraj się zachować sens i kluczowe przesłanie.
- Kolejny punkt to porównanie kont darmowych i płatnych. To pierwszy moment, w którym poruszamy temat płacenia za dostęp do usług AI. Zakładam, że czytelnik nie będzie jeszcze rozumiał, dlaczego miałby za to płacić. Zacznijmy więc od krótkiego wprowadzenia do tematu.
- Naśladując mój styl pisania, włącz część swoich sugestii do udostępnionego tekstu. Oznacz swoje dodatki za pomocą potrójnych cudzysłowów """.
- Chciałbym porównać użycie tokenów API do monet w automacie do gier, ale nie chcę, żeby brzmiało to przestarzale. Jaka inna metafora by tu dobrze pasowała?

Podsumowanie

- Prompty to początkowe polecenia lub instrukcje przekazywane systemowi AI, które kierują go do wygenerowania określonej odpowiedzi lub wykonania konkretnego zadania. Stanowią podstawę interakcji z AI oraz wyznaczają zakres i kierunek jej działania.
- Stanowość (ang. *statefulness*) odnosi się do zdolności systemu do zapamiętywania wcześniejszych interakcji i wykorzystywania ich kontekstu w bieżących lub przyszłych odpowiedziach. Ta cecha ma kluczowe znaczenie dla zachowania ciągłości rozmowy lub procesu i umożliwia AI udzielanie bardziej spójnych i trafnych odpowiedzi.
- Utrzymanie kontekstu (ang. *context retention*) to zdolność modelu AI do przechowywania i wykorzystywania informacji kontekstowych

w trakcie całej rozmowy lub serii zadań. Dzięki temu odpowiedzi AI pozostają trafne, precyzyjne i uwzględniają przebieg wcześniejszej interakcji.

- Spójność (ang. *coherence*) oznacza logiczne i konsekwentne powiązanie pomiędzy odpowiedziami lub działaniami systemu. Dzięki niej generowane przez AI treści są nie tylko adekwatne do bieżącego zapytania, lecz także zgodne z ogólnym tokiem rozmowy lub realizowanym zadaniem.
- Token to najmniejsza jednostka danych przetwarzanych przez model AI — może to być całe słowo, jego część lub pojedynczy znak. Zrozumienie pojęcia tokenów jest kluczem do uświadomienia sobie, w jaki sposób modele AI interpretują i generują tekst.
- Limit tokenów określa maksymalną liczbę tokenów, jaką model AI może przetworzyć naraz lub wygenerować w odpowiedzi. Ten limit wpływa na długość i złożoność możliwych interakcji.
- Konta darmowe i płatne w usługach AI różnią się najczęściej właśnie limitami tokenów oraz dostępem do bardziej zaawansowanych funkcji. Wersje premium oferują wyższe limity, szybsze przetwarzanie i dodatkowe możliwości, co czyni je bardziej odpowiednimi dla użytkowników o większych wymaganiach — zwłaszcza w zastosowaniach profesjonalnych.

Twoja przewaga w IT zaczyna się od AI

Sztuczna inteligencja przestała być futurystyczną wizją — to narzędzie, które już dziś zmienia codzienną pracę specjalistów IT. ChatGPT, Claude i inne modele AI rewolucjonizują sposób, w jaki programiści piszą kod, administratorzy zarządzają systemami, a menedżerowie prowadzą zespoły. Książka stanowi praktyczny przewodnik po zastosowaniu sztucznej inteligencji w realnych scenariuszach — od automatyzacji rutynowych zadań po rozwiązywanie złożonych zagadnień technicznych. To pozycja dla każdego, kto chce nie tylko nadążyć za zmianami, ale też uczynić z AI przewagę konkurencyjną w swojej karierze.

Autorzy dzielą się sprawdzonymi technikami wykorzystania AI w każdym aspekcie pracy IT. Prowadzą czytelnika od podstaw inżynierii promptów, przez praktyczne zastosowania w dokumentacji i komunikacji, aż po zaawansowane scenariusze. Każdy rozdział zawiera gotowe do użycia prompty, rzeczywiste studia przypadków i konkretne przykłady kodu w PowerShell, Python, SQL i innych językach. Nie tylko zobaczysz, jak pisać skuteczne prompty, ale przede wszystkim dowiesz się, jak formułować problemy, by AI mogła realnie wspomóc pracę specjalisty — w zakresie analizy logów, optymalizacji zapytań SQL, przygotowania planów odtwarzania awaryjnego czy prowadzenia trudnych rozmów z zespołem.

Specjaliści z branży IT znajdą w tej książce strategie wykorzystania AI, w których drzemią olbrzymie możliwości. Zastosuj je w praktyce — a potem daj mi znać, kiedy dostaniesz awans.

— **Jeffrey Snover,** twórca PowerShell, Distinguished Engineer w Google

W książce:

- Praktyczne techniki tworzenia promptów
- Automatyzacja zadań administracyjnych
- Zastosowanie asystentów kodu (GitHub Copilot, Cline, Cursor AI)
- Zarządzanie zespołami z AI
- Bezpieczeństwo i etyka

Chrissy LeMaire — dwukrotna laureatka Microsoft MVP i GitHub Star, twórczyni popularnego modułu PowerShell dbatools, autorka książki *Learn dbatools in a Month of Lunches*. Programistka z niemal 30-letnim doświadczeniem w IT, prelegentka na międzynarodowych konferencjach.

Brandon Abshire — menedżer techniczny z ponad 20-letnim doświadczeniem w administrowaniu bazami danych i zarządzaniu zespołami IT w czołowych organizacjach ochrony zdrowia i firmach z listy Fortune 500. Posiada certyfikaty z administracji BD, ITIL 4 i zarządzania projektami.

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 978-83-289-3597-6

9 788328 935976

Cena: 129,00 zł