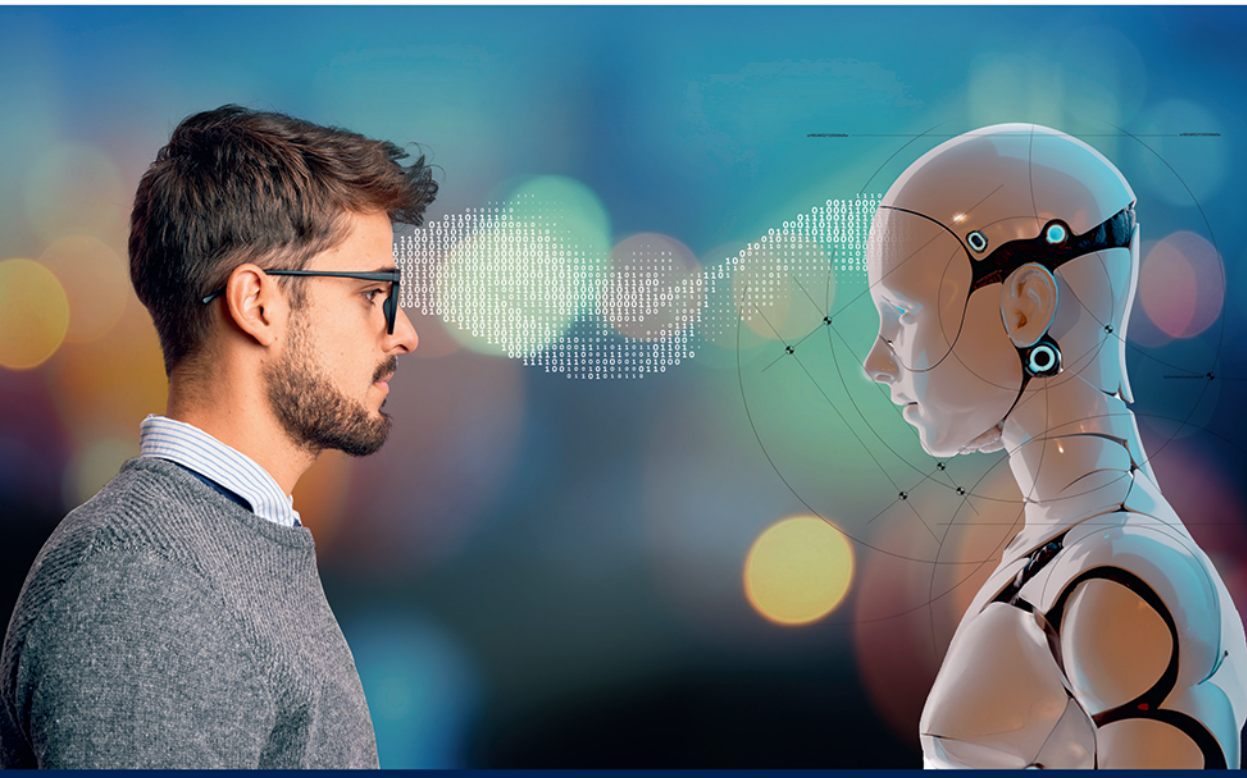
Andrzej Kacprzak

Prompt engineering i ChatGPT

Poradnik skutecznej komunikacji ze sztuczną inteligencją



Helion

Redaktor prowadzący: Tomasz Gojowy

Projekt okładki: Studio Gravite/Olsztyn
Obarek, Pokoński, Pazdrijowski, Zaprucki

Materiały graficzne na okładce zostały wykorzystane za zgodą AdobeStock.

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/popren
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Kolorowe wersje wybranych rysunków dostępne pod adresem
https://ftp.helion.pl/przyklady/popren.zip

ISBN: 978-83-289-1362-2

Printed in Poland.

SPIS TREŚCI

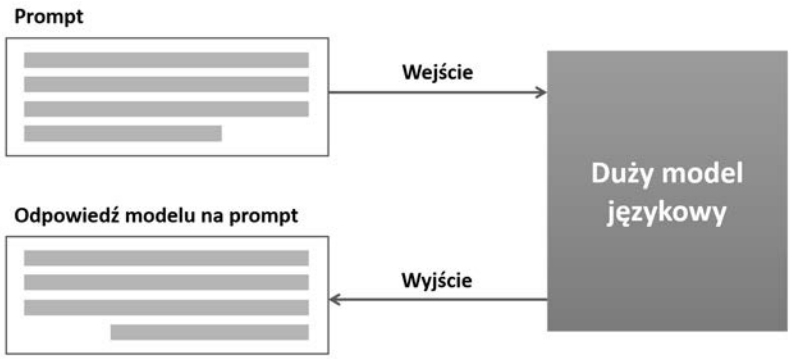
6

PROMPT ENGINEERING – OPTYMALIZACJA KOMUNIKACJI ZE SZTUCZNĄ INTELIGENCJĄ

6.1. Czym jest prompt engineering?

Jednym z kluczowych elementów skutecznego korzystania z ChatGPT jest umiejętne dostosowywanie fraz, poleceń i zapytań, zwanych „promptami". Proces ten jest znany jako „prompt engineering", czyli „inżynieria zapytań" lub „inżynieria poleceń".

Wiesz już, że sztuczna inteligencja to dziedzina nauki zajmująca się tworzeniem systemów, które mogą wykonywać zadania na poziomie zbliżonym do ludzkiego, a nawet lepiej. Do tego niezbędna jest skuteczna inżynieria zapytań, czyli technika, która polega na dostarczaniu modelom SI odpowiednich informacji i instrukcji, które kierują ich działaniem. Dzięki temu modele te mogą generować odpowiedzi, które uwzględniają kontekst i niuanse zadania, co zwiększa ich użyteczność w różnych obszarach. Niezależnie od tego, czy chodzi o tłumaczenie tekstu, analizę emocji zawartych w tekście, czy tworzenie nowych treści, właściwie skonstruowane i dostosowane prompty pozwalają lepiej dopasować wyjście modelu do potrzeb i oczekiwań użytkownika.



Prompt engineering to zestaw technik i metod projektowania, pisania i optymalizowania instrukcji (promptów) dla dużych modeli językowych. Celem jest uzyskanie od modelu dokładnych, konkretnych, precyzyjnych, powtarzalnych i poprawnych odpowiedzi. Przez precyzyjne formułowanie promptów użytkownicy mogą uzyskać od SI bardziej wartościowe odpowiedzi.

6.2. Strategie prompt engineering

Niektóre najważniejsze ogólne strategie prompt engineering to:

- **Ustalanie konkretnej struktury zdania**

  W kontekście prompt engineering jedno z kluczowych pytań dotyczy tego, czy lepiej jest używać pytań, poleceń, próśb czy innych typów promptów, aby najlepiej odpowiadały naszym potrzebom. To, jak skonstruujesz swój prompt — czy będzie to pytanie, polecenie, czy coś innego — może mieć duży wpływ na to, jak dobrze model zrozumie, czego od niego oczekujesz. Testowanie różnych sposobów zadawania pytań może pomóc znaleźć te, które dają najjaśniejsze i najbardziej trafne odpowiedzi sztucznej inteligencji.

  Można np. sprawdzać, jak model reaguje na bezpośrednie pytania w porównaniu z pytaniami warunkowymi lub jak radzi sobie z różnymi rodzajami pytań otwartych. Dostosowanie sposobu, w jaki formułujemy swoje prośby, może znacząco wpłynąć na to, jak dokładne i użyteczne informacje otrzymujemy w odpowiedzi.

- **Dostosowywanie do specyficznej dziedziny**

  Ta strategia umożliwia otrzymywanie odpowiedzi, które są nie tylko dokładniejsze, ale też bardziej adekwatne. Na przykład, gdy szukasz informacji technicznych czy marketingowych, użycie w prompcie odpowiednich terminów może prowadzić do otrzymania informacji, które będą dokładne i naprawdę przydatne.

  W praktyce chodzi o to, by umiejętnie wykorzystywać specjalistyczne słowa kluczowe, wyrażenia lub akronimy, które są typowe dla interesującej Cię dziedziny. Dzięki eksperymentom z dostosowaniem zapytań do konkretnych obszarów lub specjalności możesz znacząco poprawić jakość interakcji z modelem SI.

- **Kontrolowanie tonu i stylu**

  W praktyce projektowania promptów dla SI bardzo ważne jest, aby dobrze rozumieć i kontrolować ton oraz styl, w jakim formułuje się prośby. To, jak mówimy lub piszemy do SI, może mocno wpłynąć na to, co i jak nam odpowie. Przykładowo zastosowanie formalnego i oficjalnego języka w kontekście np. biznesowym skłania ChatGPT do udzielania odpowiedzi profesjonalnych i szczegółowych, podczas gdy luźniejszy ton rozmowy może skłonić model do udzielania odpowiedzi bardziej swobodnych i nieformalnych. Różnicowanie stylu i tonu komunikacji w zależności od tematu czy kontekstu pozwala dostosować odpowiedzi ChatGPT do własnych oczekiwań.

  To trochę jak eksperymentowanie z różnymi przepisami, aby znaleźć najlepszy smak. Próbując różnych sposobów zadawania pytań, możemy odkryć, które z nich są najskuteczniejsze w przypadku różnych tematów i w różnych sytuacjach.

- **Optymalizacja długości promptu**

  Optymalizacja długości promptu jest fundamentalnym aspektem efektywnego prompt engineeringu, odgrywającym kluczową rolę w zapewnieniu klarowności komunikacji z SI. Eksperymentuj z promptami o różnej długości, co pozwoli Ci znaleźć idealną równowagę, tak aby model otrzymał odpowiednią ilość informacji, a zarazem nie został przeciążony. Zmiana długości promptu umożliwia precyzyjne dopasowanie go do charakterystyki konkretnego zadania.

Zbyt długie prompty mogą prowadzić do zniekształcenia intencji zapytania i zwiększać ryzyko niejasności w jego interpretacji, co może skutkować generowaniem niecelnych lub nadmiernie rozbudowanych odpowiedzi. Natomiast zbyt krótkie prompty, choć minimalizują ryzyko przeciążenia modelu, mogą nie zapewniać wystarczającego kontekstu, przez co odpowiedzi mogą się okazać mniej trafne lub zbyt ogólnikowe.

Uwaga: **ChatGPT łączy wiedzę, którą zyskał podczas swojego treningu, z informacjami, które otrzymuje od użytkownika. Mimo że główna zasada jego funkcjonowania jest niezmienna, odpowiedzi, których udziela, za każdym razem mogą być różne, nawet jeśli zadane pytanie jest sformułowane w ten sam sposób.**

7

PODSTAWOWE TECHNIKI PROMPT ENGINEERING

Prompt engineering obejmuje zestaw metod mających na celu dostosowanie optymalnego zapytania dla modeli opartych na tekście (np. ChatGPT), generatorów obrazów (np. Midjourney, DALL-E) oraz generatorów kodu (np. GitHub Copilot), co wpływa na generowane przez nie wyniki. W tym rozdziale przyjrzymy się technikom tworzenia promptów pod kątem zastosowania ich w chatbotach SI, takich jak ChatGPT czy Gemini. Zaprezentuję również przypadki zastosowania inżynierii promptów w praktyce, ilustrując, w jaki sposób manipulacja zapytaniem może prowadzić do uzyskania lepszych rezultatów.

Omówimy praktyczne aspekty tworzenia prostych i skutecznych zapytań, pozwalających użytkownikom uzyskać pożądane odpowiedzi.

Dodatkowo należy podkreślić, że niniejszy rozdział pełni funkcję wstępu do technik prompt engineering. W następnym poznasz zaawansowane techniki umożliwiające efektywniejsze wykorzystanie potencjału nie tylko modeli językowych, ale również generatorów obrazów.

7.1. Informacje wprowadzające

OGRANICZENIA CZASOWE DLA DANYCH TRENINGOWYCH

Ograniczenia czasowe dla danych treningowych są kluczowe dla zrozumienia możliwości i ograniczeń każdego z modeli GPT. Chociaż są one wyjątkowo zaawansowane w generowaniu tekstu i rozumieniu zapytań, to ich wiedza jest zawsze ograniczona do tego, co było dostępne w trakcie ich treningu. To ma znaczący wpływ na ich zastosowanie szczególnie w kontekstach, które wymagają aktualnych informacji, takich jak wiadomości, bieżące wydarzenia i najnowsze odkrycia naukowe. Użytkownicy muszą być świadomi tych ograniczeń, aby odpowiednio wykorzystywać te modele.

GPT-3 zawiera informacje do września 2021 roku, zatem wszelkie wydarzenia i zmiany, jakie nastąpiły po tej dacie, jak też wszelkie nowsze informacje nie są uwzględnione w jego „wiedzy". Następcą GPT-3 jest model GPT-3.5, który uaktualniono o informacje do roku 2022. Najnowsza dostępna wersja to GPT-4 (w chwili gdy pisze te słowa), która ma rozszerzoną bazę wiedzy ogólnej obejmującą dane do kwietnia 2023 roku.

LICZBA TOKENÓW ZAPAMIĘTYWANYCH PRZEZ RÓŻNE WERSJE GPT

Ze względu na praktyczne i obliczeniowe ograniczenia ChatGPT musi operować z określoną liczbą tokenów, co wpływa na jego zastosowanie w różnych sytuacjach. Zrozumienie tych ograniczeń jest kluczowe w konwersacji z modelem, który będzie efektywnie wykorzystywał dostępne tokeny, zarazem zapewniając wartościowe wyniki.

Model	Rozmiar kontekstu tokenów
ChatGPT-3.5	4096
ChatGPT-4	8192

Rozmiar kontekstu tokenów w ChatGPT odnosi się do maksymalnej liczby tokenów (słów, znaków interpunkcyjnych, specjalnych znaków itp.), które model może wziąć pod uwagę przy generowaniu odpowiedzi. Token to podstawowa jednostka tekstu, którą model rozumie i przetwarza.

GPT — technologia stojąca za ChatGPT — działa przez przewidywanie następnego tokenu w zdaniu na podstawie wcześniejszych tokenów. Maksymalna długość kontekstu jest ograniczona ze względu na architekturę i ograniczenia obliczeniowe modelu. Oznacza to, że model jest w stanie „pamiętać" lub „rozważyć" tylko określoną liczbę ostatnich tokenów podanych w ramach pytania lub dialogu.

Przykładowo, jeśli rozmiar kontekstu tokenów wynosi 4096 tokenów, oznacza to, że generując swoją odpowiedź, model może wziąć pod uwagę do 4096 ostatnich tokenów. Jeśli podana przez użytkownika historia rozmowy (łącznie z pytaniem i poprzednimi odpowiedziami) przekracza ten limit, najstarsze tokeny zostaną „zapomniane", czyli nie będą brane pod uwagę przy generowaniu odpowiedzi.

Rozmiar kontekstu tokenów ma kluczowe znaczenie dla jakości i spójności odpowiedzi modelu, ponieważ wpływa on na to, ile informacji model może wykorzystać do zrozumienia kontekstu i wygenerowania odpowiedzi.

- **Długość konwersacji** — najważniejszym aspektem limitu tokenów jest jego wpływ na długość rozmów. Obejmuje to zarówno dane wejściowe, jak i wyjściowe. Oznacza to, że dłuższe rozmowy mogą przekraczać ustalony limit. W takiej sytuacji, kiedy limit zostanie przekroczony, model może nie być w stanie przetworzyć dalszych części konwersacji, co będzie skutkować koniecznością zakończenia lub ograniczenia rozmowy. Aby uniknąć takich problemów, użytkownicy powinni zarządzać długością zarówno swoich promptów, jak i otrzymywanych odpowiedzi — dzielić dłuższe dyskusje na mniejsze części lub stosować bardziej zwięzłe sformułowania, co pozwoli efektywnie wykorzystać dostępne tokeny i bez przeszkód kontynuować interakcję.
- **Zrozumienie kontekstu** — limit tokenów wpływa również na „pamięć" modelu językowego, co ma znaczenie dla jego zdolności do rozumienia kontekstu. Gdy rozmowa przekroczy ustalony limit, model może stracić fragmenty wcześniejszego kontekstu, co może skutkować mniej trafnymi lub niespójnymi odpowiedziami podczas konwersacji.

- **Zakres odpowiedzi** — ograniczenie liczby tokenów wpływa także na możliwość dostarczania przez model obszerniejszych i bardziej szczegółowych odpowiedzi. Gdy model zbliża się do limitu, musi formułować krótsze odpowiedzi, co może ograniczać ich szczegółowość.

HALUCYNACJE

W świecie sztucznej inteligencji istnieje zjawisko polegające na tym, że chatboty SI, takie jak ChatGPT lub Gemini, podają nieprawdziwe lub mylące informacje. Eksperci określają je mianem halucynacji. Termin ten przypisuje się Kevinowi Roose'owi, dziennikarzowi „The New York Times", który użył go w kontekście swoich doświadczeń z chatbotem Bing[1]. Zdarza się zatem, że modele językowe mogą w odpowiedzi na prompty podawać informacje, które są błędne, nieprawdziwe albo mylące. Chociaż może to wydawać się nieszkodliwe, to istotne jest, aby pamiętać, że ludzie często ufają temu, co generuje SI, a to może prowadzić do nieporozumień i błędnych przekonań.

ChatGPT, jako duży model językowy, wytrenowano na różnorodnym zbiorze danych tekstowych, który obejmował książki, artykuły, strony internetowe i inne formy tekstowe. Zbiór ten zawiera szeroki zakres tematów, od literatury po naukę, od codziennych rozmów po specjalistyczne teksty techniczne. Niektóre źródła, np. te naukowe, mogą być bardziej wiarygodne, ale trudno jest stwierdzić, kiedy model podaje informacje, które nie są prawdziwe. Odpowiedzi ChatGPT mogą zawierać wiarygodne informacje, ale równie dobrze może się tam znaleźć coś nieprawdziwego, co trudno jest zauważyć. Dlatego powinniśmy być bardzo ostrożni, kiedy korzystamy z odpowiedzi ChatGPT, zwłaszcza gdy chodzi o ważne lub delikatne tematy.

Przykłady halucynacji w odpowiedziach modeli językowych mogą obejmować:

- **wymyślanie faktów** — model może tworzyć fałszywe informacje lub twierdzenia, które nie mają żadnego uzasadnienia w rzeczywistości,
- **niezrozumiałe odpowiedzi** — model może generować odpowiedzi, które nie są sensowne w danym kontekście lub nie pasują do pytania,

[1] K. Roose, *A conversation with Bing's chatbot left me deeply unsettled*, „The New York Times", 2023.

- **błędne przewidywania** — model może sugerować przyszłe wydarzenia lub trendy, które są nieprawdziwe lub niepotwierdzone,
- **utrzymywanie fałszywych przekonań** — model może potwierdzać fałszywe przekonania lub teorie spiskowe, co może być niebezpieczne lub być źródłem dezinformacji.

Zawsze warto dokładnie czytać i weryfikować teksty wygenerowane przez ChatGPT (lub inne chatboty), aby sprawdzić ich dokładność, precyzję i wiarygodność.

OGRANICZNIKI

W swoich promptach używaj ograniczników, aby wyraźnie wskazywać odrębne części (sekcje) w danych wejściowych (prompcie) i wyjściowych (wygenerowanej odpowiedzi; tu należy poinformować ChatGPT, kiedy ma stosować dany ogranicznik). Ograniczniki mogą mieć dowolną postać np.:

‘ ’	„ ”	< >	<tag> </tag>	{START} {STOP}

By ułatwić poruszanie się po tekście i zwiększyć jego czytelność, przykłady promptów w kolejnych podrozdziałach wyraźnie oznaczono szarym kolorem tła.

Ta wizualna różnica ma na celu szybkie wskazanie czytelnikowi, które fragmenty tekstu są przykładami promptów prezentowanych w kontekście danego tematu.

Do generowania dalszych przykładów wykorzystano płatną wersję ChatGPT-4.

Warto zaznaczyć, że wszystkie techniki prompt engineering opisane w tym poradniku będą równie skuteczne zarówno w bezpłatnej wersji ChatGPT-3.5, jak i w innych modelach językowych (np. Gemini).

Twój przewodnik po świecie prompt engineeringu

Wszystkie znaki na niebie i ziemi wskazują wyraźnie: wkraczamy w erę, w której sztuczna inteligencja (SI) będzie wszechobecna. Wygra na tym ten, kto szybciej nauczy się z nią skutecznie porozumiewać. Nie czekaj zatem i już dziś opanuj sztukę tworzenia precyzyjnych i trafnych promptów, czyli instrukcji dla modeli językowych, takich jak ChatGPT.

Autor z zapałem dzieli się swoją fascynacją sztuczną inteligencją i prezentuje praktyczne metody jej zastosowania w różnych obszarach. Korzystając z własnego doświadczenia inżynierskiego, w przystępny sposób wyjaśnia zasady prompt engineeringu, pozwalające optymalnie formułować „podpowiedzi" dla SI.

Pracując z tym praktycznym przewodnikiem, nauczysz się między innymi:

- **Efektywnie komunikować się z SI**
- **Budować prompty, które przynoszą oczekiwane rezultaty**
- **Korzystać z prostych, a także bardziej zaawansowanych strategii i technik prompt engineeringu**
- **Zaprzęgać ChatGPT do rozwiązywania problemów z różnych dziedzin życia zarówno prywatnego, jak i zawodowego**

Odkryj fascynujący świat sztucznej inteligencji i poznaj możliwości, jakie otwiera przed Tobą dobrze skonstruowany prompt.

Dr inż. Andrzej Kacprzak

Zastępca kierownika Katedry Zaawansowanych Technologii na Politechnice Częstochowskiej, gdzie z zaangażowaniem łączy pracę naukową z dydaktyczną. Od lat pasjonat technologii przyszłości ze szczególnym uwzględnieniem sztucznej inteligencji, innowacyjnych węglowych ogniw paliwowych i odnawialnych źródeł energii. Z entuzjazmem eksploruje świat SI i prompt engineeringu, skupiając się na tworzeniu skutecznych instrukcji dla zaawansowanych modeli językowych.

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 978-83-289-1362-2

9 788328 913622

Cena: 59,00 zł