

12 zasad, żeby zużywać mniej tokenów

Prosta checklista, która pomaga pisać krócej, jaśniej i skuteczniej.

Praca z AI coraz częściej przypomina zarządzanie budżetem. Tyle że zamiast pieniędzy bardzo często przepalamy coś mniej widocznego: **tokeny**.

Tokeny wpływają na koszt, szybkość działania, długość kontekstu i jakość odpowiedzi. Im więcej niepotrzebnych informacji trafia do modelu, tym więcej musi on przeczytać, przetworzyć i uwzględnić przy odpowiedzi.

Dobra wiadomość: w większości przypadków nie trzeba znać technicznych szczegółów, żeby zużywać mniej tokenów. Wystarczy zmienić kilka codziennych nawyków.

Ta checklista pomoże Ci:

- pisać lepsze prompty,
- ograniczyć niepotrzebny kontekst,
- szybciej dochodzić do dobrych odpowiedzi,
- zmniejszyć liczbę poprawek,
- korzystać z AI bardziej świadomie.

Zasada główna:

Pisz krócej, konkretniej i bardziej świadomie. Model nie potrzebuje wszystkiego. Potrzebuje właściwych informacji.

1

Nie wrzucaj całych plików bez potrzeby

Wysyłaj tylko fragment, który faktycznie analizujesz - reszta to zbędny koszt.

Jednym z najczęstszych błędów jest wrzucanie do AI całych dokumentów, prezentacji, raportów lub długich maili, mimo że pytanie dotyczy tylko małego fragmentu.

Model musi wtedy przetworzyć znacznie więcej treści, niż faktycznie potrzebuje. To zwiększa zużycie tokenów, wydłuża odpowiedź i może pogorszyć jej jakość, bo najważniejsze informacje giną w szumie.

Zamiast wysłać cały plik, **wybierz konkretny fragment**: akapit, tabelę, sekcję, kilka slajdów albo dokładny wycinek danych.

Słabszy prompt

„Przeanalizuj ten cały dokument i powiedz, co poprawić.”

Lepszy prompt

„Przeanalizuj tylko sekcję „Oferta” z poniższego tekstu. Sprawdź, czy argumentacja jest jasna, czy język jest sprzedażowy i co można skrócić bez utraty sensu.”

Zapamiętaj

Im mniej zbędnych danych, tym większa szansa na konkretną, szybką i tańszą odpowiedź.

2

Dobierz model do zadania

Do prostych zadań używaj tańszych modeli, a mocniejsze zostaw do trudnych analiz.

Nie każde zadanie wymaga najmocniejszego modelu. To tak, jakby do każdej notatki, korekty literówki lub prostego streszczenia angażować cały zespół ekspertów.

Mocniejsze modele warto zostawić do zadań, które wymagają rozumowania, strategii, analizy wielu zmiennych, pracy na trudnym materiale albo podejmowania decyzji.

Do prostych zadań często wystarczy tańszy i szybszy model.

W praktyce

- Proste zadania: korekta, skrócenie tekstu, zmiana tonu, formatowanie, proste maile.
- Trudniejsze zadania: strategia, analiza danych, porównanie wariantów, decyzje biznesowe.

Zapamiętaj

Najdroższy model nie zawsze jest najlepszym wyborem. Najlepszy model to ten, który pasuje do zadania.

3

Pisz dokładnie, czego oczekujesz

Podaj cel, kontekst, odbiorcę, format i ograniczenia. Mniej zgadywania oznacza mniej poprawek.

AI zużywa więcej tokenów, gdy musi zgadywać, czego właściwie chcesz. Ogólny prompt zwykle prowadzi do ogólnej odpowiedzi.

Lepiej poświęcić kilkanaście sekund na napisanie dobrego polecenia niż później prowadzić długą serię poprawek.

Dobry prompt powinien zawierać pięć elementów: **cel, kontekst, odbiorcę, format i ograniczenia.**

Słabszy prompt

„Napisz post o AI.”

Lepszy prompt

“Napisz post na LinkedIn o tym, że tokeny w LLM-ach stają się nowym kosztem pracy z AI. Odbiorcy: marketerzy i founderzy. Ton: praktyczny, lekko prowokacyjny. Format: hook, 5 punktów, pytanie na końcu.”

Zapamiętaj

Dobry prompt to nie długi prompt. Dobry prompt to konkretny prompt.

4

Ogranicz kontekst

Usuń wszystko, co nie wpływa na odpowiedź. Każde dodatkowe słowo to więcej tokenów, które model musi ponownie przeczytać.

Wielu użytkowników zakłada, że im więcej kontekstu, tym lepiej. To prawda tylko częściowo. Dobry kontekst pomaga. Nadmiarowy kontekst przeszkadza.

Jeżeli wklejasz długą historię, stare wersje tekstu, nieaktualne założenia lub poboczne informacje, model musi je uwzględnić.

Warto usuwać stare wersje, powtórzenia, nieaktualne założenia, komentarze poboczne i fragmenty niezwiązane z pytaniem.

Słabszy prompt

„Robiliśmy wcześniej kilka wersji landing page’a, potem zmieniliśmy sekcję hero, ale nie wiem, czy to jeszcze działa...”

Lepszy prompt

„Oceń sekcję hero poniżej. Cel: zwiększyć zrozumienie wartości produktu w 5 sekund. Odbiorca: marketer B2B. Sprawdź nagłówek, podtytuł i CTA.”

Zapamiętaj

Kontekst ma pomagać modelowi, a nie zasypywać go informacjami.

5

Proś o konkretny format odpowiedzi

Lista, tabela, 3 punkty, maksymalna długość - format ogranicza rozwlekłość.

Jeżeli nie określisz formatu, model sam zdecyduje, jak długa i szczegółowa ma być odpowiedź.

Format działa jak ogranicznik. Pomaga modelowi odpowiedzieć krócej i bardziej użytecznie.

Możesz poprosić o listę 5 punktów, tabelę z plusami i minusami, checklistę, odpowiedź do 120 słów albo mail w 3 akapitach.

Słabszy prompt

„Co poprawić na tej stronie?”

Lepszy prompt

„Wypisz 7 najważniejszych poprawek UX. Dla każdej podaj: problem, dlaczego to przeszkadza użytkownikowi, rekomendację. Odpowiedź w tabeli.”

Zapamiętaj

Jeśli chcesz krótką odpowiedź, nie licz na to, że model sam się domyśli. Napisz, jak ma wyglądać wynik.

6

Zaczynaj nowy chat przy nowym temacie

Nowy wątek resetuje historię i usuwa koszt starego kontekstu.

Długie rozmowy są wygodne, ale mają ukryty koszt. Model często musi brać pod uwagę wcześniejszy kontekst rozmowy.

Jeżeli w jednym czacie mieszasz strategię, copywriting, błędy techniczne i analizę konkurencji, kontekst robi się ciężki.

Nowy temat warto zaczynać w nowym czacie. Dzięki temu model nie musi nosić historii, która nie ma już znaczenia.

W praktyce

- Zaczynaj nowy chat, gdy zmieniasz temat.
- Zaczynaj nowy chat, gdy poprzednia rozmowa jest bardzo długa.
- Zaczynaj nowy chat, gdy model zaczyna mieszać stare założenia.

! Zapamiętaj

Długi czat jest dobry dla ciągłości pracy. Nowy czat jest dobry dla świeżości, precyzji i oszczędności.

7

Edytuj prompt, zamiast pisać follow-upy

Mniej wiadomości w historii oznacza mniej tokenów przy każdej kolejnej odpowiedzi.

Kiedy odpowiedź nie jest dobra, najczęściej dopisujemy kolejną wiadomość: „bardziej konkretnie”, „krócej”, „dodaj przykład”.

Po kilku poprawkach rozmowa robi się długa, a model musi analizować coraz więcej historii.

Czasem lepiej wrócić do pierwszego prompta i go poprawić. Dzięki temu model dostaje od razu lepsze instrukcje.

Słabszy prompt

„Napisz post. Krócej. Bardziej ekspercko. Dodaj przykłady. Nie takim tonem.”

Lepszy prompt

„Napisz post na LinkedIn. Ton: ekspercki, ale prosty. Struktura: hook, krótka historia, 5 praktycznych wniosków, pytanie na końcu. Maksymalnie 1800 znaków. Dodaj 2 konkretne przykłady.”

Zapamiętaj

Nie każda poprawka wymaga nowej wiadomości. Czasem najtańsza iteracja to lepszy pierwszy prompt.

8

Grupuj podobne zadania

Zamiast pisać pięć osobnych próśb, połącz podobne zadania w jeden uporządkowany prompt.

Jeżeli masz kilka podobnych zadań, nie musisz pisać osobnej wiadomości do każdego z nich.

To szczególnie dobrze działa przy tekstach, analizach, wariantach reklam, meta opisach, nagłówkach, pomysłach kreatywnych i prostych transformacjach treści.

Model widzi całość zadania i może utrzymać spójność. Ty dostajesz komplet materiałów szybciej.

Słabszy prompt

„Napisz meta title. Teraz meta description. Teraz H1. Teraz 5 nagłówków reklamowych.”

Lepszy prompt

„Przygotuj zestaw SEO i reklamowy dla poniższej strony: meta title do 60 znaków, meta description do 155 znaków, H1, 5 nagłówków reklamowych i 5 CTA. Odpowiedź podaj w tabeli.”

Zapamiętaj

Jeśli zadania mają ten sam kontekst, połącz je. Jeśli mają różne cele, rozdziel je.

Ustaw własne instrukcje startowe dla AI

Nie powtarzaj za każdym razem zasad stylu, języka i formatu.
Zapisz je raz.

Jeżeli często prosisz AI o podobny styl odpowiedzi, nie warto za każdym razem wpisywać tych samych reguł.

Możesz przygotować własne instrukcje startowe: jakim językiem ma odpowiadać, jaki ton preferujesz, czego ma unikać i jak ma formatować odpowiedzi.

To skraca codzienne prompty i porządkuje odpowiedzi.

W praktyce

- Przygotuj jeden niezmienny styl przy swojej konwersacji z AI, przykładowo: "Odpowiadaj po polsku. Pisz konkretnie i praktycznie. Najpierw daj krótką odpowiedź, potem szczegóły. Preferuj listy punktowane i gotowe teksty do wklejenia. Unikaj długich wstępów."

Zapamiętaj

Stała instrukcja to Twój osobisty brief dla AI.

Zapisz własny zestaw reguł pracy z AI

Stwórz bibliotekę sprawdzonych promptów i instrukcji do powtarzalnych zadań.

Największe oszczędności pojawiają się wtedy, gdy przestajesz pisać wszystko od zera.

Jeśli często wykonujesz podobne zadania, warto stworzyć własną bibliotekę promptów.

To mogą być szablony do maili, analiz, postów, reklam, briefów, researchu, SEO, raportów albo pracy z danymi.

W praktyce

- analiza landing page'a
- post na LinkedIn
- brief kampanii
- audyt SEO

Zapamiętaj

Dobra biblioteka promptów działa jak system operacyjny Twojej pracy z AI.

Najważniejsza zasada

Nie chodzi o to, żeby pisać jak najmniej. Chodzi o to, żeby nie wysyłać AI rzeczy, których nie potrzebuje.

Największy koszt tokenów powstaje zwykle przez trzy rzeczy: zbyt dużo niepotrzebnego kontekstu, zbyt ogólne prompty i zbyt wiele poprawek wynikających z braku precyzji.

Przed wysłaniem prompta sprawdź: czy AI zna cel, czy podałeś tylko potrzebny kontekst, czy określiłeś format odpowiedzi, czy wybrałeś właściwy model i czy temat wymaga nowego czatu.

Prosty wzór dobrego prompta

Chcę osiągnąć: [cel]

Kontekst: [tylko najważniejsze informacje]

Odbiorca: [kto]

Twoje zadanie: [konkretna czynność]

Format odpowiedzi: [lista/tabela/mail/post]

Ograniczenia: [długość, ton, język]

! Zapamiętaj

Krótko, konkretnie, na temat. To najlepszy sposób, żeby uzyskać świetne odpowiedzi przy mniejszym zużyciu tokenów.



Dodatkowe 2 TIPY dla zaawansowanych

Stosuj wyraźne separatory danych

Używaj tagów, np. XML, markdown albo potrójnego cudzysłowu, aby oddzielić polecenie od treści. Model nie musi się domyślać, co jest instrukcją, a co materiałem do analizy.

Dlaczego to działa

- **Model szybciej rozpoznaje strukturę zadania.**
Jasne granice pomagają od razu zidentyfikować, co ma zrobić.
- **Mniej miesza instrukcje z materiałem źródłowym.**
Wyraźny podział redukuje ryzyko błędnej interpretacji treści.
- **To ogranicza błędy, poprawki i niepotrzebne iteracje.**
Lepsza organizacja to krótsze odpowiedzi i wyższa jakość.

W praktyce

- Oddziel zadanie od danych potrójnym cudzysłowem.
- Używaj sekcji: „Zadanie”, „Tekst”, „Format”.
- Przy dłuższych promptach stosuj znaczniki XML lub markdown.
 - `<instrukcja> Streszcz tekst </instrukcja>`
 - `<tekst> [tutaj wklejasz dane] </tekst>`
- Dzięki temu model szybciej rozumie, co ma zrobić.

Zapamiętaj

Im czytelniej oddzielisz polecenie od danych, tym mniej model musi zgadywać.

Wyeliminuj „puste” tokeny - zasada “Negative Prompting”

Zakaż modelowi pisanie wstępów, podziękowań i podsumowań. Niech generuje tylko czyste mięso - dokładnie to, czego potrzebujesz.

Dlaczego to działa

- **Mniej grzecznościowych formułek i ogólników.**
Model nie traci tokenów na wstępy typu „Jasne”, „Pewnie” albo końcowe podsumowania.
- **Więcej treści, którą da się od razu wykorzystać.**
Odpowiedź szybciej nadaje się do wklejenia do maila, dokumentu, prezentacji lub strony.
- **To daje realne oszczędności przy codziennej pracy.**
Przy wielu promptach dziennie nawet drobne skróty przekładają się na niższy koszt i szybsze działanie.

W praktyce

- Napisz: „Bez wstępu, bez podziękowań, bez podsumowania”.
- Poproś o samą finalną odpowiedź albo gotowy wynik.
- Przy analizie określ dokładnie strukturę odpowiedzi.
- Eliminuj każde zdanie, którego nie potrzebujesz.

Zapamiętaj

Każde zdanie, którego nie potrzebujesz, też kosztuje tokeny.

Dziękuję za pobranie checklisty

AI nie potrzebuje więcej tekstu. AI potrzebuje lepszego briefu.

Praca z AI będzie coraz bardziej przypominać pracę z budżetem. Nie zawsze wygrywa ten, kto ma dostęp do najmocniejszego modelu. Coraz częściej wygrywa ten, kto potrafi dobrze formułować zadania.

Ta checklista to prosty punkt startowy. Możesz wracać do niej za każdym razem, gdy AI odpowiada zbyt długo, zbyt ogólnie albo wymaga zbyt wielu poprawek.

Chcesz sprawdzić, ile tokenów zużywa Twój tekst?

iletokenow.pl

Kalkulator tokenów, optimizer promptów i praktyczne przykłady pracy z AI.

Mniej chaosu. Więcej konkretności. Lepsze odpowiedzi.