Mapa pojęć: model, token, kontekst, prompt, temperatura

> Zanim zaczniemy mówić o agentach i o tym, co robią źle albo dobrze, warto ustawić wspólny słownik pięciu pojęć, które w dalszych stronach książki pojawiają się bez tłumaczenia. To słowniczek wejściowy, nie wykład o architekturze transformera.

Skróty użyte na tej stronie (2)

Skrót	Co znaczy
AI	Artificial Intelligence: sztuczna inteligencja, czyli programy naśladujące ludzkie rozumowanie
CLI	Command Line Interface: praca przez wpisywanie komend w terminalu

Z życia wzięte

Z ŻYCIA WZIĘTE

Tomek wkleja do czatu cały plik logów z serwera log01, żeby pokazać błąd z ostatniej godziny. Odpowiedź przychodzi wolniej niż zwykle, a licznik zużytych tokenów rośnie znacznie szybciej niż przy wcześniejszych pytaniach tego dnia. Tłumaczymy mu, że model rozlicza się w tokenach za wszystko, co wysyłamy, więc przefiltrowany grepem wycinek kosztowałby ułamek tego zapytania. Tomek zapamiętuje, że krótszy kontekst to nie tylko wygoda, to też realne koszty przy każdym kolejnym pytaniu.

Model, wersja i wariant

> **Pojęcie: model językowy.** Program, który na wejściu dostaje ciąg tekstu i zwraca rozkład prawdopodobieństwa kolejnego fragmentu tego tekstu — w praktyce dopisuje najbardziej prawdopodobny dalszy ciąg, token po tokenie. Model sam z siebie nie ma pamięci między wywołaniami, nie ma zegara i nie wie, która jest godzina, chyba że ktoś mu to powie w treści zapytania.

Modele dostawców takich jak Anthropic, OpenAI czy Google wychodzą w rodzinach i wersjach — na przykład Claude Sonnet 5 czy Claude Opus 5.5. Numer wersji zmienia się z każdym większym treningiem i zwykle oznacza ogólny wzrost jakości na całym froncie zadań, nie jedną konkretną poprawkę.

Osobną osią od wersji jest wariant, czyli tryb pracy tego samego modelu. Najważniejszy dziś podział to model rozumujący (reasoning) kontra szybki. Model w trybie rozumowania generuje dodatkowy, wewnętrzny ciąg rozważań przed właściwą odpowiedzią — kosztuje to więcej czasu i tokenów, ale poprawia wyniki na zadaniach wieloetapowych, matematycznych i takich, gdzie łatwo się pomylić po drodze. Model szybki pomija ten krok i odpowiada od razu. U Anthropic ten wybór sprowadza się dziś w dużej mierze do parametru głębokości rozumowania (`effort`) w obrębie tego samego modelu, a nie do osobnego produktu — inni dostawcy (rodzina modeli rozumujących OpenAI) historycznie trzymali to jako osobne linie modeli. To rozróżnienie bywa mylone z numerem wersji, a to dwie zupełnie inne osie: wersja mówi „jak dobry model w ogóle", wariant mówi „ile czasu ten konkretny model dostał na zastanowienie się nad tą konkretną odpowiedzią".

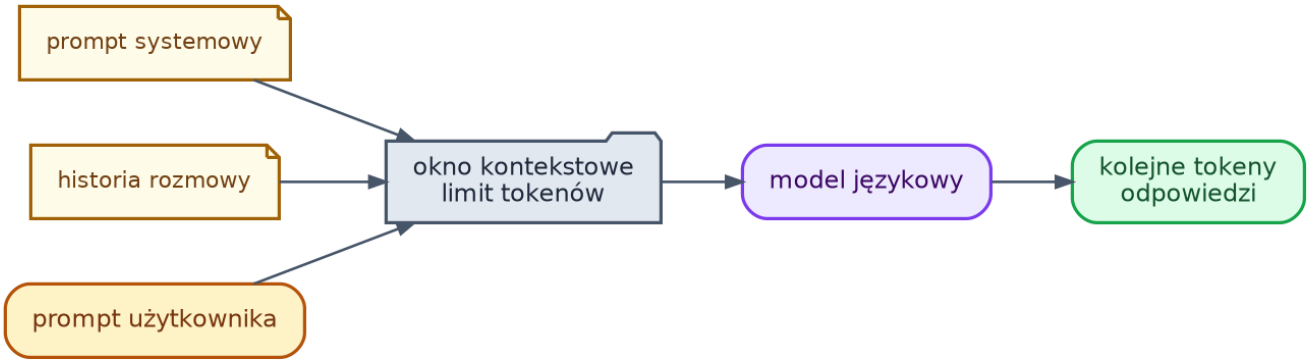
Token

> **Pojęcie: token.** Najmniejsza jednostka tekstu, jaką model faktycznie widzi i którą się rozlicza — nie znak, nie słowo, tylko fragment wyznaczony przez tokenizer danego modelu. Krótkie, częste angielskie słowa bywają jednym tokenem, dłuższe albo rzadsze słowa rozpadają się na kilka. Dla języka polskiego to ma konkretną, bolesną konsekwencję: tokenizery są trenowane głównie na tekstach angielskich, więc polskie końcówki fleksyjne i znaki diakrytyczne rozbijają się na więcej tokenów niż odpowiadający im tekst angielski o tej samej treści. Ten sam raport napisany po polsku bywa wyraźnie droższy w tokenach niż jego angielski odpowiednik.

Rozliczenie modeli w chmurze odbywa się właśnie w tokenach, osobno za wejście (to, co wysyłamy — instrukcja, historia, wklejone pliki, wyniki narzędzi) i za wyjście (to, co model wygenerował). Stąd dwa praktyczne wnioski: po pierwsze, wklejenie całego pliku logów zamiast jego przefiltrowanego wycinka kosztuje realne pieniądze przy każdym kolejnym obrocie pętli agenta, bo ten sam tekst jedzie w każdym zapytaniu na nowo. Po drugie, długość odpowiedzi też ma cenę — model poproszony o zwięzłość generuje mniej tokenów wyjściowych, nie tylko krótszy tekst do przeczytania.

Okno kontekstowe

> **Pojęcie: okno kontekstowe (context window).** Maksymalna liczba tokenów — wejściowych i wyjściowych razem — jaką model może obsłużyć w jednym wywołaniu. Dzisiejsze modele frontowe mają okna liczone w setkach tysięcy tokenów, część dostawców reklamuje miliony, ale nominalny limit i realnie użyteczny limit to dwie różne liczby — do tego wracamy osobno na stronie Kontekst i pamięć: co model wie, a czego nie, bo jakość odpowiedzi potrafi spadać dużo wcześniej, niż okno się fizycznie zapełni.



Co trafia do okna kontekstowego, zanim model zacznie generować odpowiedź (kliknięcie otwiera pełny rozmiar)

Ważne jest to, że okno kontekstowe nie jest tym samym co pamięć modelu. Model nie „zapamiętuje" niczego z poprzedniej rozmowy — jeżeli coś ma być dostępne w kolejnym wywołaniu, musi zostać dosłownie wklejone do okna od nowa, przez klienta albo przez samego agenta.

Prompt systemowy kontra prompt użytkownika

> **Pojęcie: prompt systemowy.** Instrukcja ustawiana przez twórcę aplikacji albo administratora agenta, niewidoczna domyślnie dla osoby piszącej z drugiej strony — określa rolę modelu, zasady, styl odpowiedzi i, w agentach, pełną listę dostępnych narzędzi razem z ich opisami. Prompt systemowy stoi zawsze na początku kontekstu i obowiązuje przez całą

sesję.

Pojęcie: prompt użytkownika. Treść, którą faktycznie wpisuje osoba prowadząca rozmowę — pytanie, polecenie, wklejony fragment. W przeciwieństwie do promptu systemowego zmienia się przy każdej turze rozmowy.

Rozróżnienie ma znaczenie praktyczne przy diagnozowaniu dziwnych zachowań: jeżeli agent zachowuje się nietypowo mimo dobrze sformułowanego polecenia, przyczyna często leży w prompcie systemowym albo w opisach narzędzi, których osoba pisząca zapytanie w ogóle nie widzi i nie może w locie poprawić.

Temperatura i determinizm

Pojęcie: temperatura. Parametr sterujący tym, jak bardzo model trzyma się najbardziej prawdopodobnego kolejnego tokena, a jak bardzo dopuszcza mniej prawdopodobne alternatywy. Temperatura bliska zeru daje odpowiedzi bardziej powtarzalne i ostrożne, wysoka temperatura — bardziej zróżnicowane i czasem bardziej pomysłowe, ale też bardziej podatne na zboczenie z tematu.

Warto rozbroić tu częste założenie: temperatura ustawiona na zero nie oznacza pełnej powtarzalności. Sposób, w jaki nowoczesne modele wykonują obliczenia równolegle na wielu kartach graficznych, wprowadza niewielkie różnice numeryczne między uruchomieniami tego samego zapytania, a modele typu mixture-of-experts potrafią przy okazji trafić w inną ścieżkę obliczeń. Dla zadań administracyjnych — audytu, dokumentacji, jednorazowego skryptu — to nie ma znaczenia. Ma znaczenie tam, gdzie ktoś buduje testy regresyjne i zakłada identyczny wynik przy identycznym wejściu; do tego wracamy na stronie Jak sprawdzić, czy to w ogóle działa.

Halucynacja jako zjawisko, nie usterka

Pojęcie: halucynacja. Sytuacja, w której model generuje treść brzmiącą pewnie i spójnie, a nieprawdziwą albo niemającą odpowiednika w rzeczywistości — nieistniejący parametr komendy, przełącznik CLI z innej wersji oprogramowania, cytat z dokumentu, którego nikt nie napisał. Kluczowe jest to, żeby nie traktować tego jak błędu programistycznego do

> zgłoszenia i naprawienia, tylko jako bezpośrednią konsekwencję tego, czym model jest: generatorem najbardziej prawdopodobnego dalszego ciągu tekstu, a nie bazą faktów z wbudowaną weryfikacją. Model nie ma osobnego mechanizmu, który sprawdzałby prawdziwość własnej odpowiedzi przed jej wysłaniem, chyba że dostanie do tego narzędzie — wyszukiwarkę, dokumentację, dostęp do naszego środowiska — i faktycznie z niego skorzysta.

Z tego samego powodu obniżenie temperatury albo zmiana wersji modelu zmniejsza częstość halucynacji, ale nigdy jej nie zeruje — to nie jest przełącznik do wyłączenia, tylko ryzyko do ograniczania przez uziemienie odpowiedzi w realnych źródłach (dokumentacja, wynik narzędzia, nasz własny plik), o czym więcej na stronie Kontekst i pamięć: co model wie, a czego nie.

Jak te pojęcia się łączą

Pojedyncze wywołanie modelu wygląda tak: prompt systemowy plus historia rozmowy plus prompt użytkownika trafiają razem do okna kontekstowego, liczonego w tokenach; model, w wybranej wersji i wariancie, z ustawioną temperaturą, generuje kolejne tokeny aż uzna odpowiedź za skończoną. Każdy z tych elementów da się zepsuć osobno — złe okno (za mało miejsca albo za dużo szumu), zła temperatura (niepowtarzalność tam, gdzie akurat jest potrzebna), pomylony prompt systemowy z użytkownika, i wreszcie halucynacja jako ryzyko wpisane w samą naturę generowania tekstu.

Częste błędy

Mylenie tokena ze słowem albo ze znakiem — przy szacowaniu kosztu albo długości promptu prowadzi to do niedoszacowania, zwłaszcza dla tekstu polskiego.

Zakładanie, że temperatura zero daje matematycznie identyczny wynik przy każdym uruchomieniu — bliski prawdy w praktyce, nieprawdziwy w ścisłym sensie i istotny przy budowaniu testów regresyjnych.

Traktowanie halucynacji jak buga do zgłoszenia dostawcy zamiast jak ryzyka do ograniczenia przez weryfikację i narzędzia.

Mylenie wersji modelu z wariantem — upgrade z jednej wersji na drugą i przełączenie trybu rozumowania w tej samej wersji to dwie osobne decyzje, z osobnym wpływem na koszt i czas odpowiedzi.

□□ Ćwiczenia

> Zadania sprawdzające.

1. Mamy dwa zdania o tej samej treści: jedno po polsku, jedno po angielsku. Które prawdopodobnie zużyje więcej tokenów i dlaczego?

Odpowiedź

Zdanie polskie, z dwóch powodów naraz: końcówki fleksyjne i znaki diakrytyczne rozbijają słowa na więcej fragmentów niż w tokenizerze trenowanym głównie na angielskim, a same znaki spoza podstawowego alfabetu bywają kodowane mniej efektywnie. Ta sama treść kosztuje więcej w polskim niż w angielskim.

2. Agent w trybie rozumowania (reasoning) i ten sam agent w trybie szybkim dostają identyczne zadanie. Co się zmienia poza czasem odpowiedzi?

Odpowiedź

Liczba tokenów wygenerowanych, zanim padnie właściwa odpowiedź — tryb rozumowania produkuje dodatkowy, wewnętrzny ciąg rozważań, co podnosi koszt tego konkretnego wywołania. W zamian zyskuje się zwykle wyższą trafność na zadaniach wieloetapowych, ale to korzyść statystyczna, nie gwarancja dla pojedynczego przypadku.

3. Ktoś twierdzi, że skoro ustawił temperaturę na zero, to jego test regresyjny promptu jest w pełni deterministyczny i nie wymaga uruchamiania więcej niż raz. Co pomija w tym rozumowaniu?

Odpowiedź

Pomija to, że temperatura zero ogranicza losowość wyboru tokena, ale nie eliminuje niedeterminizmu numerycznego wynikającego z równoległych obliczeń na sprzęcie i, w części modeli, z routingu w architekturze mixture-of-experts. Rzetelny test regresyjny przy zadaniach o wysokiej stawce nadal wymaga kilku powtórzeń, nie jednego uruchomienia.

Źródła: dokumentacja Claude Platform, prompt engineering i parametr `effort`: https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices — dokumentacja Claude Platform, ogólnie: https://platform.claude.com — Building Effective AI Agents, Anthropic: https://www.anthropic.com/engineering/building-effective-agents

---

Revision #34
Created 2026-09-25 10:33:14 UTC by Tomasz Konieczny
Updated 2026-10-26 14:10:08 UTC by Tomasz Konieczny