

• MODELE AI

Z jakimi modelami AI działa SumizAI?

SumizAI nie sprzedaje Ci tokenów. Podłączasz klucz API, który już masz, do dostawcy, którego już wybrałeś, a aplikacja rozmawia z nim w Twoim imieniu. **Obsługiwanych jest siedmiu dostawców**, w tym jeden działający w całości na Twojej własnej maszynie.

Aktualizacja: 2026-08-14 8 pytań 3 min czytania

Q. Jakich dostawców mogę podłączyć?

Anthropic, OpenAI, Gemini, Groq, OpenRouter, lokalną Ollamę oraz własny serwer.

Pod maską są dwa klienci, a nie siedem. Anthropic ma własny, a pozostała szóstka mówi protokołem OpenAI, więc pokrywa je jedna implementacja. Dlatego w ogóle działa dołożenie własnego endpointu — jeśli mówi tym protokołem, jest już obsługiwany.

Możesz podłączyć więcej niż jednego i wybierać, którego użyć w danej rozmowie, z etykietą kosztu przy modelu w nagłówku.

Q. Dlaczego muszę przynieść własny klucz?

Bo alternatywą jest marża. Gdyby aplikacja dostarczała model, cena musiałaby zawierać narzut na każdy wygenerowany token, a dolar miesięcznie byłby niemożliwy.

Oznacza to też, że relacja handlowa dotycząca drogiej części jest między Tobą a Twoim dostawcą. Widzisz własne zużycie, na własnym panelu, po własnej stawce — łącznie z kredytami czy rabatem, które już masz.

I dlatego lokalna Ollama jest pełnoprawną opcją. Skoro granicą jest Twój klucz, granicą jest też Twoja decyzja, żeby nie mieć żadnego zewnętrznego dostawcy.

Q. Skąd bierze się lista modeli?

Od dostawcy, odpytanego Twoim kluczem, z pamięcią podręczną na piętnaście minut.

Kiedyś katalog był w kodzie i po cichu bywał błędny: proponował model, który dostawca już wycofał, a dowiadywałeś się o tym, gdy notatka nie chciała się wygenerować. Teraz niepodłączony dostawca pokazuje pustą listę, a podłączony — to, co faktycznie ma.

Samo wklejenie klucza wystarczy, żeby zacząć. Aplikacja pyta dostawcę tym kluczem i bierze najmocniejszy model, jaki oferuje, a zły klucz kończy się natychmiast przyczyną od dostawcy, zamiast zapisem, który wywróci się później.

Q. Jak lista jest uporządkowana?

Modele darmowe nad płatnymi, a w każdej grupie od najmocniejszego — nasza ocena jakości, potem cena promptu, potem okno kontekstu, potem alfabet. Interfejs pokazuje obie grupy i nie sortuje ich ponownie.

Istnieje to z powodu OpenRouter, gdzie lista liczy setki modeli. Przy tej skali użytecznym pierwszym pytaniem jest „co mam za darmo”, a dopiero potem „co najlepszego warto kupić”.

Jeśli podłączysz tylko jednego dostawcę, ledwo to zauważysz. Jeśli podłączysz OpenRouter, jest to różnica między używalną a nieużywalną listą.

Q. Czy wszystko korzysta z tego samego modelu?

Nie, i jest to decyzja kosztowa. Jest model główny i pomocniczy.

Model główny odpowiada Tobie i — przy przebudowie spisu treści — układa rozdziały. Pomocniczy wykonuje mniejsze zadania o ustalonej strukturze: pisze tytuł i podsumowanie oraz ocenia, czy nowa notatka duplikuje istniejącą.

Ocena duplikatów biegnie przy każdej notatce wobec kilku kandydatów, więc przeniesienie jej na tańszy model jest różnicą między funkcją, którą zostawiasz włączoną, a taką, którą wyłączasz.

Q. Czy mogę sprawdzić, który dostawca jest u mnie najszybszy?

Tak — w ustawieniach jest test szybkości, który mierzy realny czas odpowiedzi każdego podłączonego dostawcy jego własnym modelem.

Publikowane benchmarki to średnie z cudzych warunków sieciowych. Twoje opóźnienie do danego dostawcy z miejsca, w którym faktycznie siedzisz, to inna liczba — i to ona decyduje, czy aplikacja wydaje się szybka.

Warto powtarzać test od czasu do czasu. Dostawcy się zmieniają, Twoje łącze też.

Q. Co się dzieje, kiedy dostawca zawiedzie?

Dostajesz konkretną przyczynę i ponowienie w miejscu, a nie ogólny błąd.

Aplikacja rozróżnia zły klucz, przekroczony limit i niedostępność dostawcy i mówi, co się stało. Twoje pytanie zostało zapisane przed wywołaniem, więc nadal tam jest — nieudana tura jest oznaczona i ponawiasz ją bez przepisywania.

Jest szczegół specyficzny dla aplikacji desktopowej: wywołania wychodzą z Twojej maszyny, więc dostawca widzi nagłówek `Origin`. Anthropic odrzuca takie żądania bez jawnego nagłówka zgody, a odrzucenie *wygląda* jak błąd uwierzytelnienia. Aplikacja wysyła właściwy nagłówek; warto o tym wiedzieć, bo jeśli budujesz coś podobnego samodzielnie, ten komunikat wyśle Cię na polowanie na problem z kluczem, którego nie masz.

Q. Czy mój materiał służy do trenowania czegokolwiek?

Nie przez SumizAI. Nie ma tu żadnego procesu treningowego; Twoje notatki to pliki na Twoim dysku.

Co Twój dostawca robi z ruchem API, reguluje Twoja umowa z tym dostawcą — i to jeszcze jeden argument za tym, żeby klucz był Twój: możesz te warunki przeczytać, wybrać odpowiednio i zmienić dostawcę bez zmiany aplikacji do notatek.

Jeśli chcesz zamknąć to pytanie całkowicie, podłącz lokalną Ollamę. Nic nie opuszcza maszyny.