

Modele AI, Hermes Agent i PCSS

Czym to jest i jak to sobie uruchomicie · 15 min

Max Małecki · Tech Lead @ Insly

Model językowy w jednym zdaniu

Przewiduje kolejny fragment tekstu na podstawie tego, co dostał. Nic więcej.

- Wytrenowany raz, więc jego wiedza jest **zamrożona na dzień treningu**
- Nie ma dostępu do niczego, czego mu **sami nie dacie** w rozmowie
- Poprzednią rozmowę pamięta tylko wtedy, gdy mu ją wkleicie
- I nie sprawdza, czy to, co napisał, jest prawdą

Cała reszta dzisiejszego dnia to dokładanie mu kontekstu i narzędzi.

W czym jest naprawdę dobry

- Streszcza i wyciąga konkret z długiego dokumentu
- Daje **pięć wersji zamiast jednej**, w minutę
- Bałagan na wejściu potrafi zamienić w tabelę na wyjściu
- Porówna dwa regulaminy albo dwie wersje tego samego pomysłu
- A kiedy wiecie dokąd idziecie, ale nie od czego zacząć, podpowie pierwszy krok

Na tym realnie zaoszczędzicie czas przez najbliższe dwa dni.

Czego sam z siebie nie gwarantuje

- Nie gwarantuje prawdy. Brzmi tak samo pewnie, kiedy wie i kiedy zmyśla
- Jego świat kończy się na dniu treningu, więc aktualność też odpada
- Kontekstu, którego mu nie daliście, po prostu nie zna
- Odpowiedzialność za decyzję zostaje po Waszej stronie. Zawsze

To szybki współpracownik bez dostępu do źródeł. Tak go traktujcie.

Hermes Agent to program, nie strona

Instalujecie go **u siebie na laptopie**. Nie logujecie się do niczego w przeglądarce.

- Czyta i zmienia pliki, uruchamia komendy. Na **Waszych uprawnieniach**
- Jest na macOS z Apple Silicon, na Windowsa i na Linuksa
- **Intel Mac nie jest oficjalnie wspierany**

Wszyscy mamy to samo narzędzie, więc każda instrukcja i każdy mentor mówi o tym samym ekranie. A jeśli macie Maka na Intelu, podnieście rękę teraz, nie za godzinę.

Podpięcie PCSS — dokładne wartości

Pole	Wartość
Name	PCSS
Provider ID	pcss
Endpoint URL	https://llm.hpc.psnc.pl/v1
Default model	gpt-oss_120b
Reasoning	Medium
API Key	Wasz osobisty klucz z panelu PCSS

Test, że żyje: nowy czat, wpiszcie `Cześć, żyjesz?`

Trzy błędy, które zobaczycie

Kod	Zwykle znaczy
401	zły albo brakujący klucz API
404	zły adres albo zła nazwa modelu
429	przekroczony limit zapytań, poczekajcie chwilę

Czwarty przypadek nie ma kodu i nazywa się Intel Mac. Tu od razu wołajcie mentora, sami tego nie naprawicie.

Dwie minuty własnego debugowania, potem mentor. Naprawdę.

Co to jest PCSS i skąd klucz

Poznańskie Centrum Superkomputerowo-Sieciowe. Akademicka infrastruktura, która wystawia modele przez API.

- Modele stoją **w Polsce**, więc dane nie wychodzą z kraju
- Nie płacicie kartą, macie **przydzieloną kwotę tokenów**
- API mówi **dialektem OpenAI**, więc działa z każdym normalnym klientem
- Klucz generujecie w panelu PCSS raz i wklejacie do Hermesa

Po Campie dostęp zostaje z Wami na **miesiąc**.

Jakie modele macie

`gpt-oss_120b` to domyślny wybór na tym warsztacie, najmocniejszy i uniwersalny. Zaczynajcie od niego i nie kombinujcie, dopóki działa.

Ale lista modeli jest **przypisana do konta**, nie publiczna:

```
curl https://llm.hpc.psnc.pl/v1/models \
-H "Authorization: Bearer $PCSS_API_KEY"
```

Nie zgadujcie nazw. Skopiujcie je stąd. Zła nazwa to `404`.

Kwota tokenów to Wasz budżet

Nie ma rachunku za request. Jest przydzielona kwota, na konto, i **nie odnowi się w trakcie Campu.**

- Debugujcie na mniejszym modelu, przełączcie na duży, gdy przepływ już działa
- Długi kontekst to dużo tokenów wejścia przy **każdym** wywołaniu
- Wektoryzujcie dokumenty **raz**, nie przy każdym pytaniu

To ostatnie jest najdroższą pomyłką, jaką da się tu popełnić.

Otwórz future-lab.t13.dev

Instrukcja jest tam. Agenda, konfiguracja Hermes i Q&A, gdy utkniecie.

