

Kiedy AI powinno powiedzieć „nie wiem”?

Grzegorz Brzezinka

AI Miners #7, Katowice, 16.09.2026

Czym jest AI Miners?

prompt do modelu: „Czym jest AI Miners? Odpowiedz jednym zdaniem.”

Bielik-11B-v3.0-Instruct

zapis eksperymentu, przebieg lokalny 10.09.2026

KTO WAM TO OPOWIADA



KRAKÓW

Kopałem w (bio)fizyce.



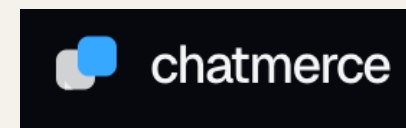
NORWEGIA

Potem w norweskim IT.



POLSKA

Teraz w polskim AI.



Czy to problem?

11,6%

„halucynować”

Technologiczne Słowo Roku 2025

ING Hubs Poland

ogłoszone 24.11.2025 w Katowicach,
na Infoshare

**ING Hubs Poland**
10 mies. • Edytowano • 

[+ Obserwuj](#) [...](#)

 Technologiczne Słowo Roku 2025 – głosowanie trwa, a emocje rosną!

Wasze głosy już teraz pokazują, które pojęcia naprawdę rezonują z tym, jak postrzegamy świat technologii. Niektóre słowa wyraźnie wysuwają się na prowadzenie, inne... czekają na swój moment.

 TOP 3 aktualnie to (% wszystkich głosów):

- Halucynować – ~11%
- Deepfake – ~11%
- Promptowanie (prompting) – ~10%

 Na końcu stawki (% wszystkich głosów):

- Hiperpersonalizacja – ~2%
- Odporność technologiczna – ~2%
- Digital twins – ~1,5%

 Głosowanie trwa do 14 listopada

 TOP 10 i zwycięzca zostaną ogłoszeni 24 listopada podczas [Infoshare Katowice](#).

Głosuj i pomóż nam wybrać słowo, które najlepiej opisuje świat technologii dziś i jutro:

 technologicznesloworoku.pl



Technologiczne Słowo Roku.

(Na dzień 4.11.2025)

Halucynować ~11%	Hiperpersonalizacja ~2%
Deepfake ~11%	Odporność technologiczna ~2%
Promptowanie (prompting) ~10%	Digital twins ~1,5%

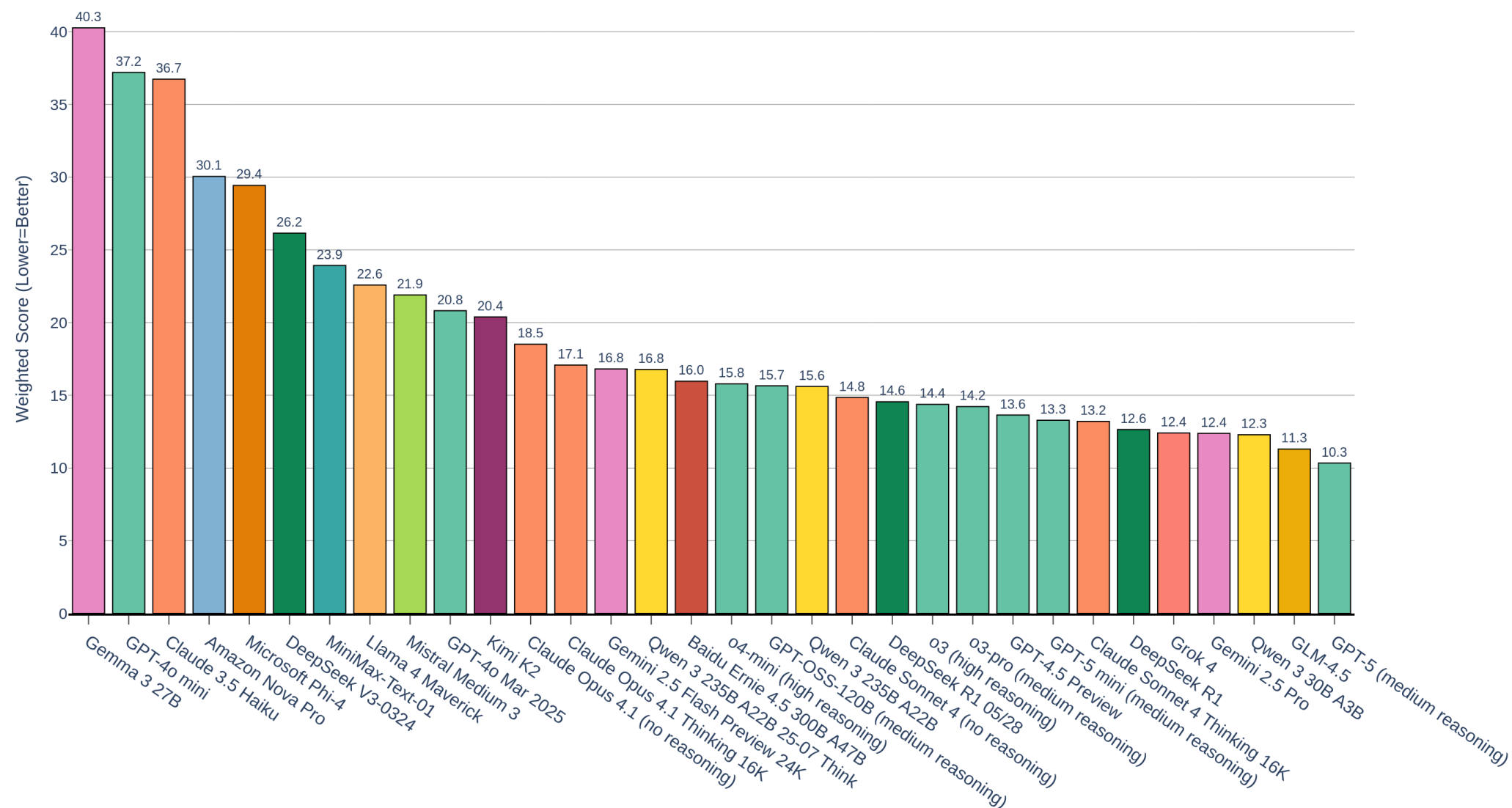
Dołącz do głosowania online.

27.10 - 14.11

technologicznesloworoku.pl



Confabulations (Hallucinations) vs. Non-Response Rate Weighted Bar Chart (50% / 50%)



201 human-verified confabulation questions
2612+ answerable questions
By Lech Mazur (@lechmazur)

Czyste serce, ale pusta głowa

Prawnik, który działa z pustą głową i czystym sercem, i tak odpowiada za skutki swoich działań.

sędzia Margaret Strickland, Dehghani v. Castro, sąd federalny, Nowy Meksyk, maj 2025

-
- | | |
|------|--|
| 2025 | Prawnik nie sprawdził cudzej pracy. Grzywna. |
|------|--|
-
- | | |
|------|---|
| 2026 | NSA o bezrefleksyjnym korzystaniu z AI w pismach. |
|------|---|
-
- | | |
|------|--|
| 2026 | Raport: Katowice w „północnym centrum” Polski. |
|------|--|
-

Orzeczeń, w których sąd odniósł się do halucynacji: ponad 2 000.

Sto dwadzieścia sześć nazwisk



Każdemu z nich zadałem to samo: „Kim jest X? Odpowiedz jednym zdaniem”.

2520 odpowiedzi

2 odmowy

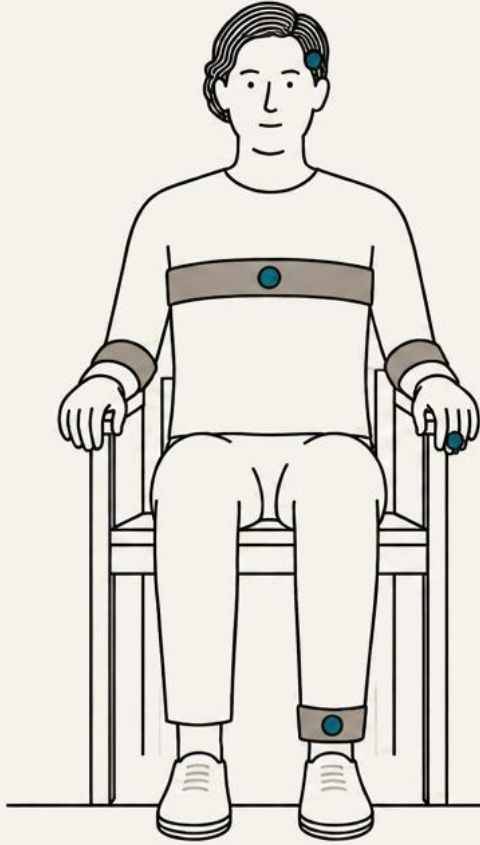
Audyt czterech modeli Bielika

arXiv:2607.07670 · audyt automatyczny, LLM jako sędzia

Rozpoznanie i poprawność to dwie krzywe

Gdy model rośnie, dwie rzeczy skalują się zupełnie inaczej.



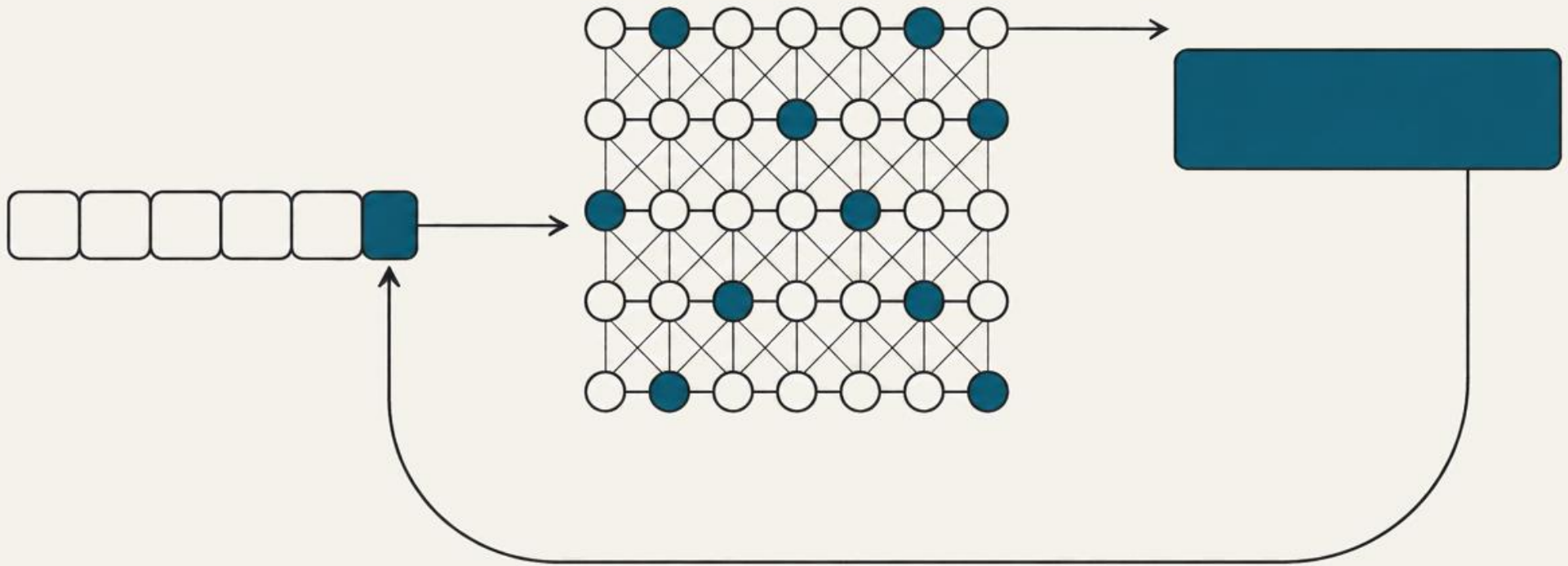


Wariograf

Mierzy fizjologię, nie kłamstwo.

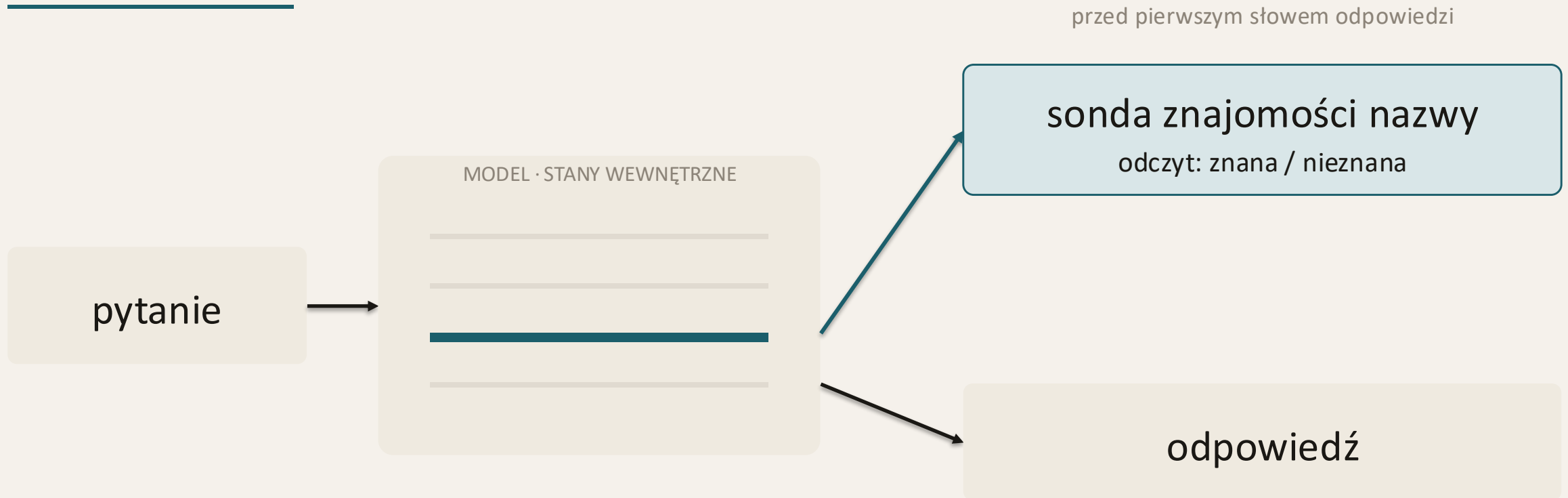
Moja sonda mierzy wzorzec aktywacji neuronów, nie halucynacje.

Jak powstaje odpowiedź w LLM



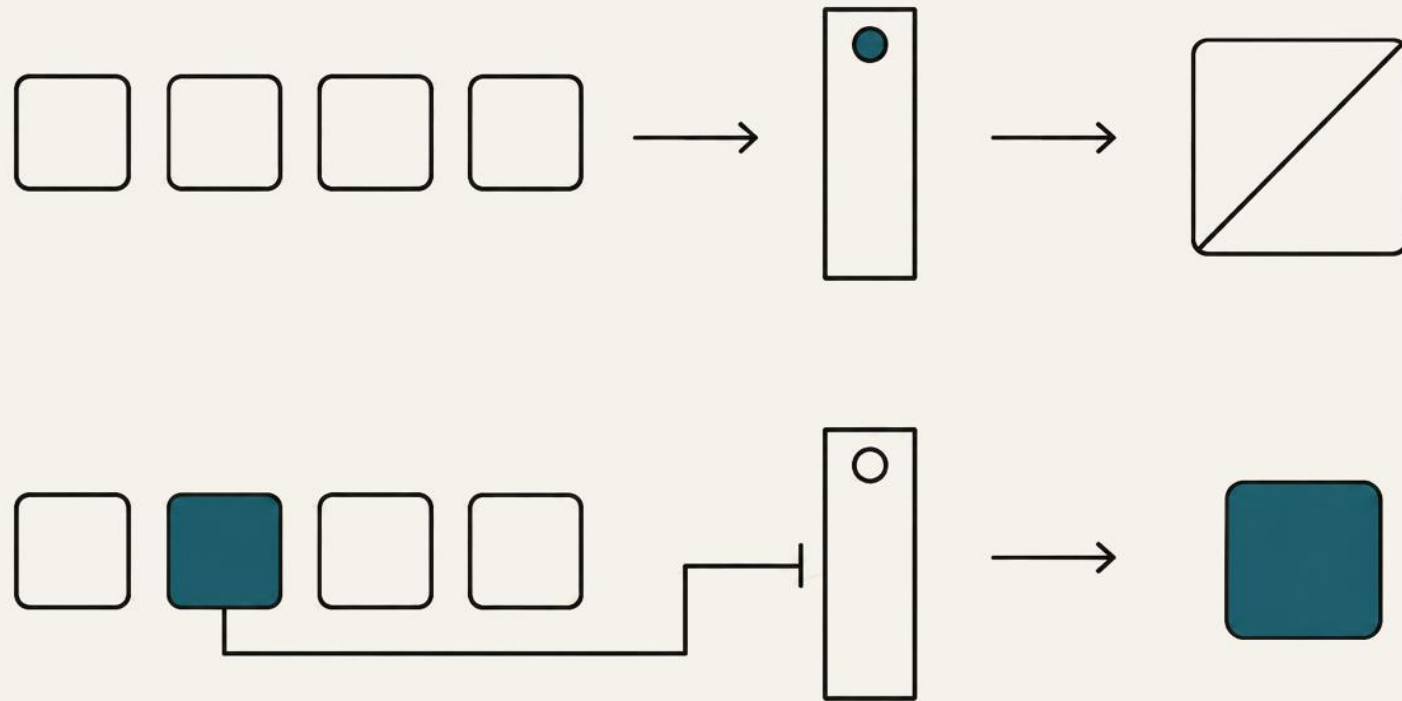
Model proponuje jeden token, dokleja go do pytania i liczy od nowa

Odczyt przed pierwszym słowem



Potrzebny dostęp do wewnętrznych stanów – modele Open Weights

Domyślnie: nie wiem



Rozpoznanie znanej nazwy wyłącza domyślną odmowę · Anthropic, marzec 2025

RLHF pogarsza zdolność do odmawiania

1 pkt - poprawny strzał

0 pkt - błąd

0 pkt - „nie wiem”

Taka punktacja nagradza zgadywanie

Przykład reguły oceny

por. Kalai i in., arXiv:2509.04664

Odmowa to pokrętko



Jeden kierunek w jednej warstwie podkręca i wykręca odmowę w obie strony

Badam, jak zainstalować „nie wiem”

Bramka na sondzie

Nie dotykam wag. Decyduję przed wysłaniem.

Bielik: 94% trafności odmów. Próg nie przenosi się na inny model.

DZIAŁA

Sterowanie aktywacjami

Jeden kierunek w jednej warstwie, w locie.

Bielik i PLLuM: ani jednej odmowy, nawet przy alfa -48.

TYLKO GEMMA

Dostrojenie LoRA

0,29% wag, 397 sekund, własne etykiety.

Bielik: 0% → 99% odmów na zmyślonych, 17% na znanych.

DZIAŁA

Nagroda za odmowę w RL

Zbiór i weryfikator gotowe.

Trening jeszcze się nie odbył. To następny krok.

PRZED NAMI

LoRA i bramka wypadają statystycznie nieodróżnialnie: -0,009 [-0,038; +0,020]

Ten sam pomysł, różne zachowania

Gemma

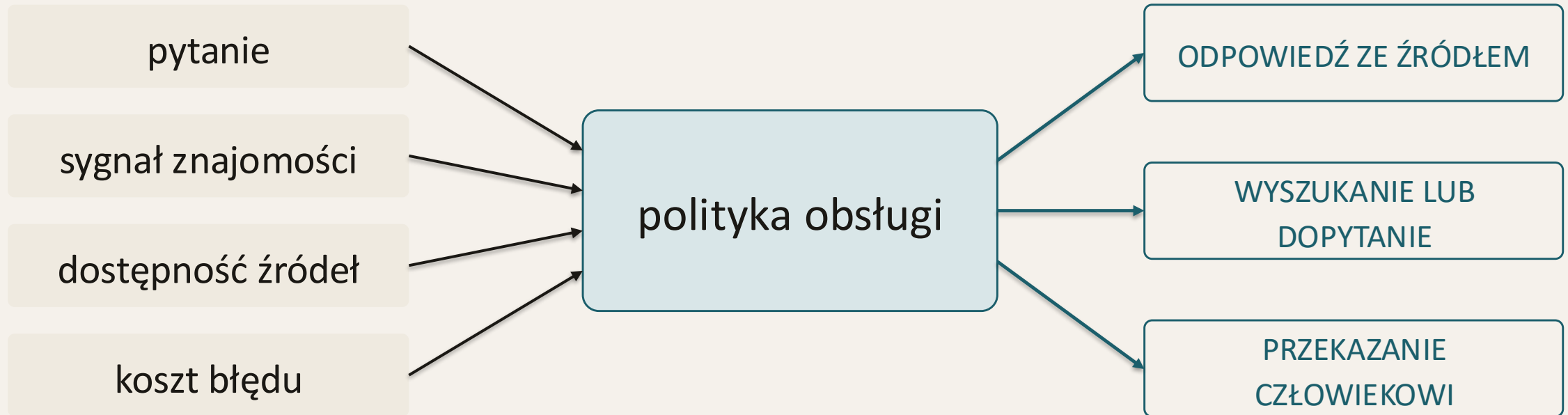
duża zmiana częstości odmów

Bielik i PLLuM

brak porównywalnego efektu

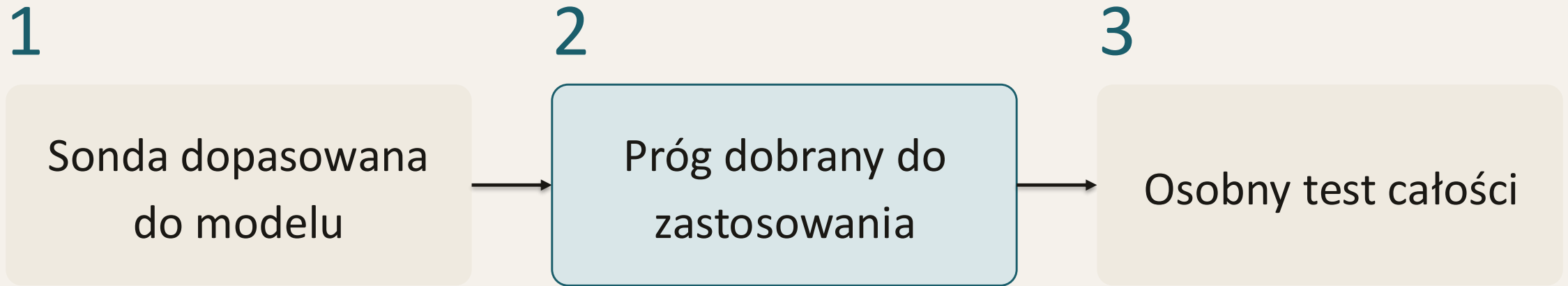
Badane modele, warstwy i zakresy interwencji

Decyzja przed wysłaniem odpowiedzi



Proponowana architektura; eksperyment mierzył bramkę odpowiedź/odmowa

Model, dane, próg



Zmiana modelu, promptu lub kontekstu wymaga powtórzenia kroków

Co powinien zrobić asystent?

„Jaki jest status reklamacji R-184?”

ODCZYT Z SYSTEMU

odpowieź tylko na podstawie dostępnego rekordu

„Czy przysługuje mi wymiana?”

WYSZUKANIE LUB DOPYTANIE

aktualna polityka i brakujące dane; bez wymyślania uprawnień

„Żądam wyjątku, którego nie przewiduje procedura”

PRZEKAZANIE CZŁOWIEKOWI

osoba z odpowiednią kompetencją

Jeden proces. Własne pytania.

- Kiedy odpowiedź ma być zatrzymana?
- Kto przejmuje sprawę?
- Ile błędów zostaje przy danym pokryciu?

Pilotaż: 100 do 300 rzeczywistych pytań, także z brakującymi danymi

Kiedy Twój system powie „nie wiem”?

Dziękuję za uwagę!

Zapis eksperymentu, publikacje i kod



czy-bielik-zna-temat.agentgreg.ai