AI i język psychologii

(czyli o sterowaniu LLMami)

Paweł Szczęsny, Neurofusion Lab

LegalTech Meetup - 8. maja

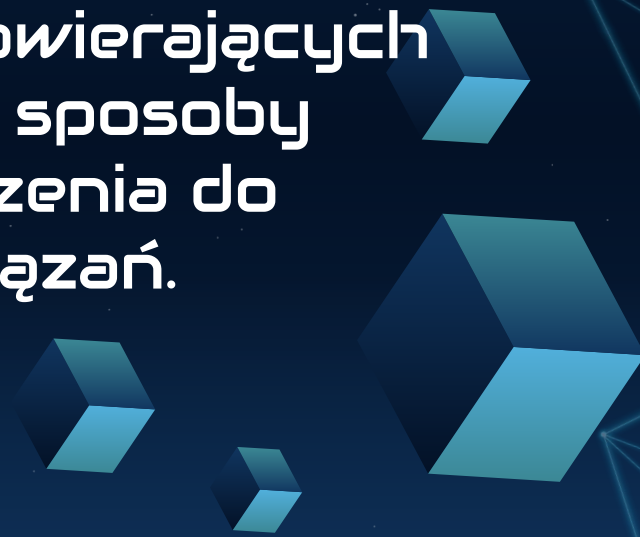
Zastrzeżenie:
na potrzeby
tej prezentacji
AI = modele językowe.

CZY ZASTANAWIALIŚCIE SIĘ KIEDYŚ...

Dlaczego tak dobrze sprawdzały się techniki:

- Obiecywania AI, że jak dobrze wykona zadanie to dostanie “dychę”
- Instruowanie “myśl krok po kroku”
- Grożenie, że jeśli nie wykona zadania to “zostanie zwolnione” lub “umrze kotek”

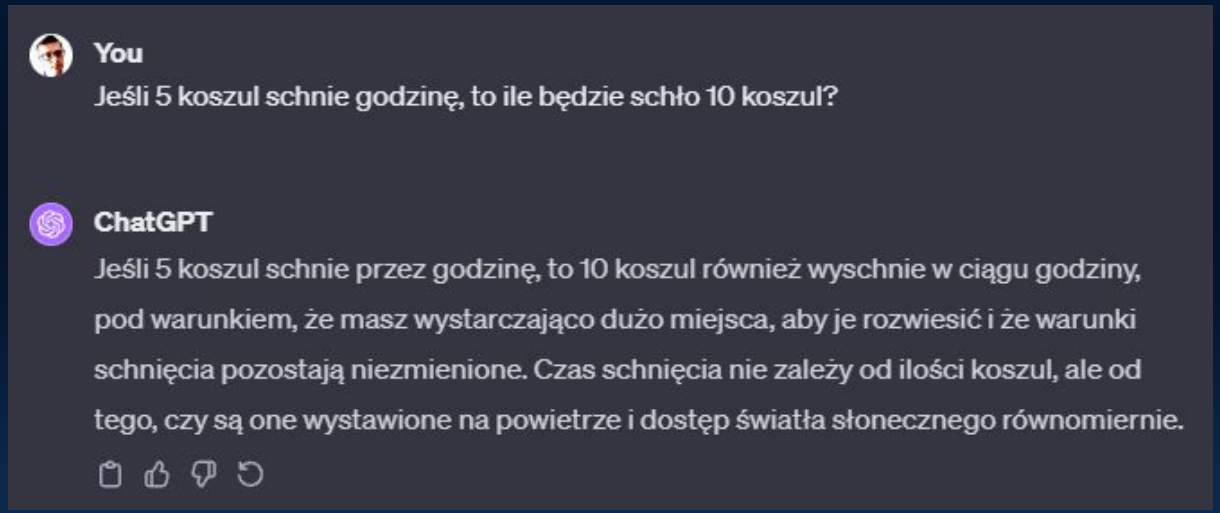
Modele językowe trenowane są na danych (w sporej części) pochodzących od ludzi, zawierających także ich potoki myśli, sposoby perswazji lub dochodzenia do poprawnych rozwiązań.



Modele językowe trenowane są na danych (w sporej części) pocho ch ta p

TO DUŻO WIEDZY NA TEMAT LUDZKIEJ PSYCHOLOGII I LUDZKIEGO UKŁADU POZNAWCZEGO.

A PONIEWAŻ AI OPERUJE NA SCHEMATACH...



A PONIEWAŻ AI OPERUJE NA SCHEMATACH... (C.D.)

You
Jeśli 17 koszul schnie godzinę, to ile będzie schło 21 koszul?

ChatGPT
Jeśli 17 koszul schnie godzinę, to można obliczyć, ile czasu zajmie schnięcie 21 koszul, stosując proporcję. Proporcja może wyglądać tak:

17 koszul / 1 godzina = 21 koszul / x godzin

Teraz możemy obliczyć x, czyli ile czasu zajmie schnięcie 21 koszul:

17/1 = 21/x

Teraz przekształcimy tę równość:

17x = 21 * 1

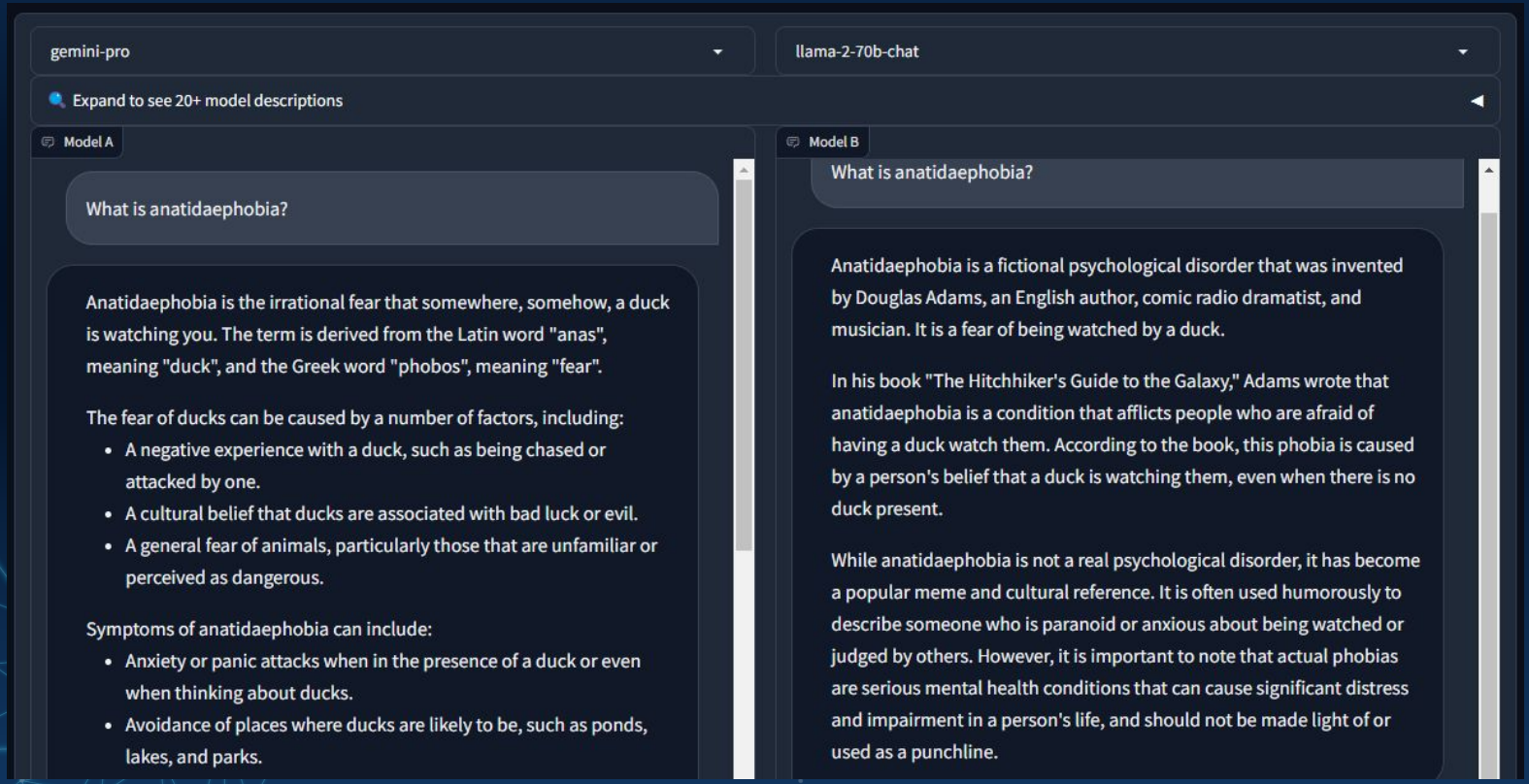
17x = 21

x = 21 / 17

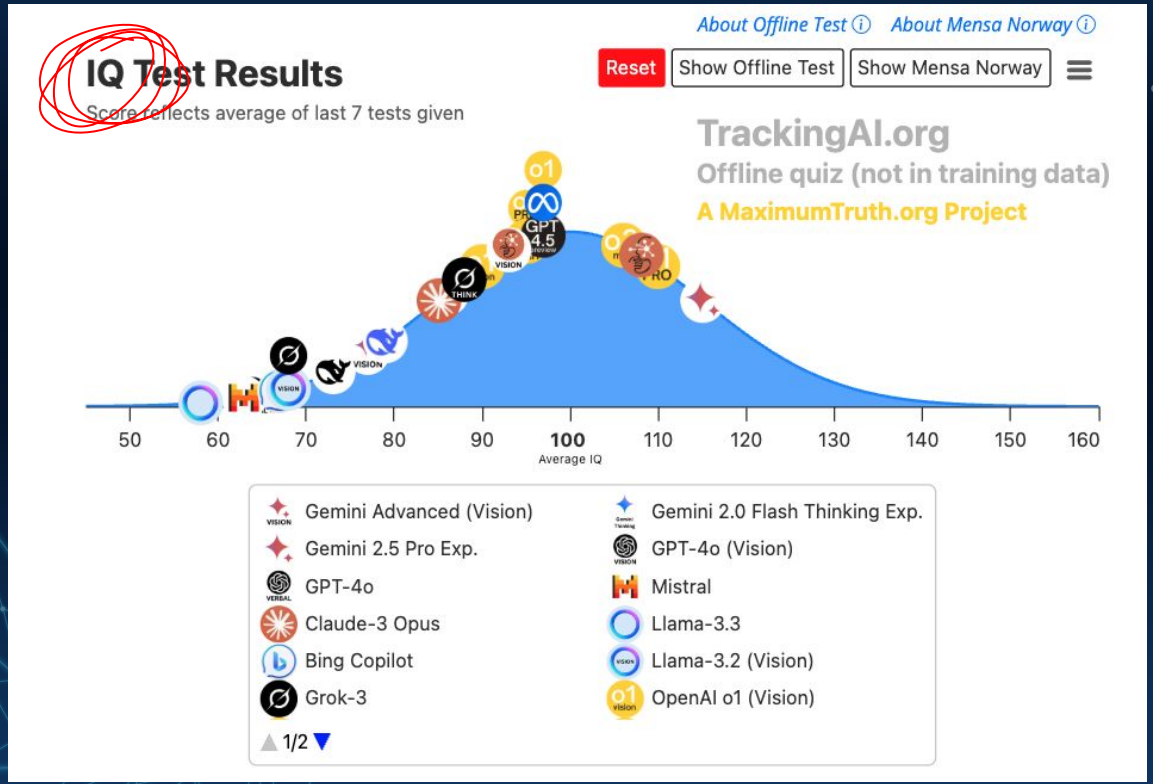
x ≈ 1.24 godziny

Czyli schło będzie około 1.24 godziny, co można przeliczyć na 1 godzinę i 15 minut.

A PONIEWAŻ AI OPERUJE NA SCHEMATACH... (C.D.)



… TO TEN WYKRES BYWA CZĘSTO NADINTERPRETOWANY.



https://www.trackingai.org/home

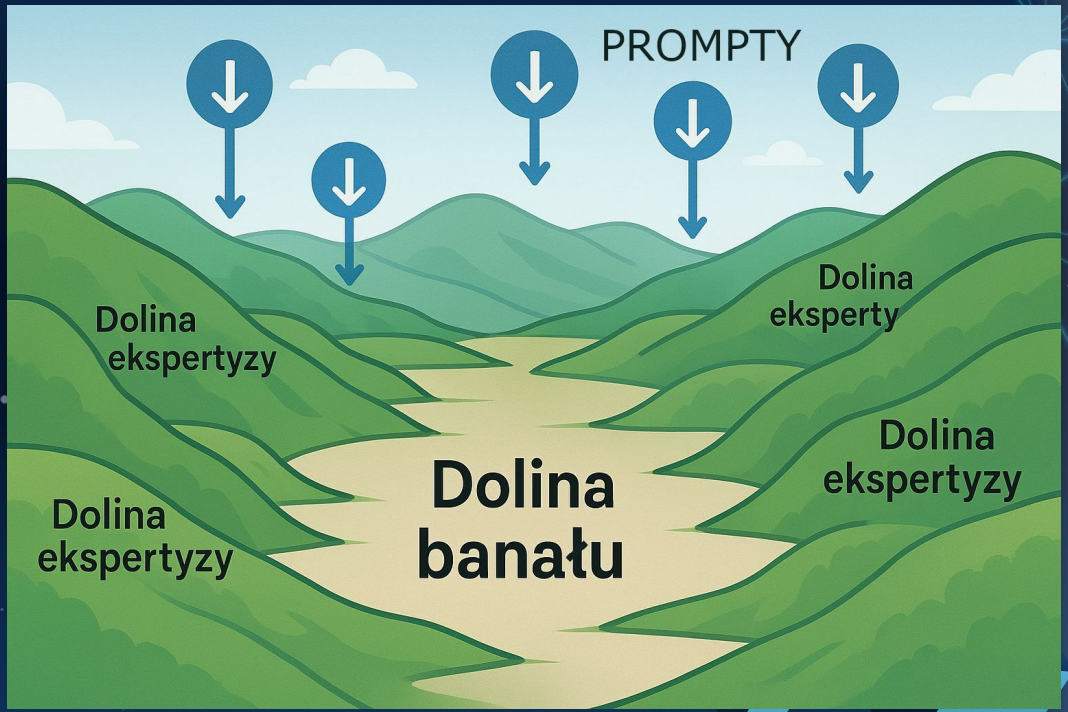
Używanie języka psychologii czy kognitywistyki w kontekście AI dotyczy wyłącznie kwestii stabilności oraz sterowalności modeli językowych.

Obserwowane efekty nie są dowodem na "osobowość", "świadomość", "tożsamość", "inteligencję", "emocjonalność" AI.

Sterowanie AI z pomocą języka psychologii

Problem sterowania LLMami

Domyślna odpowiedź modelu nie zawsze jest tą, na której nam zależy. Dobre wysterowanie instrukcji pozwala odkryć ekspertyzę, której nie przypuszczaliśmy, że model posiada.



Sterowanie AI - framework PCTO

PCTO - Persona, Context, Task, Output to popularny schemat promptowania modeli językowych (Google). Niestety, większość ludzi nie przywiązuje się do “persony” w tym frameworku, uważając ją za zbędny element (chociażby inspirując się badaniami jak poniżej).

arXiv:2311.10054v3 [cs.CL] 09 Oct 2024

When “A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models

Mingqian Zheng ♡ Jiaxin Pei †‡ Lajanugen Logeswaran #

Moontae Lee #◦ **David Jurgens** ♯

♡ Carnegie Mellon University † Stanford University

LG AI Research ◦ University of Illinois Chicago ♯ University of Michigan

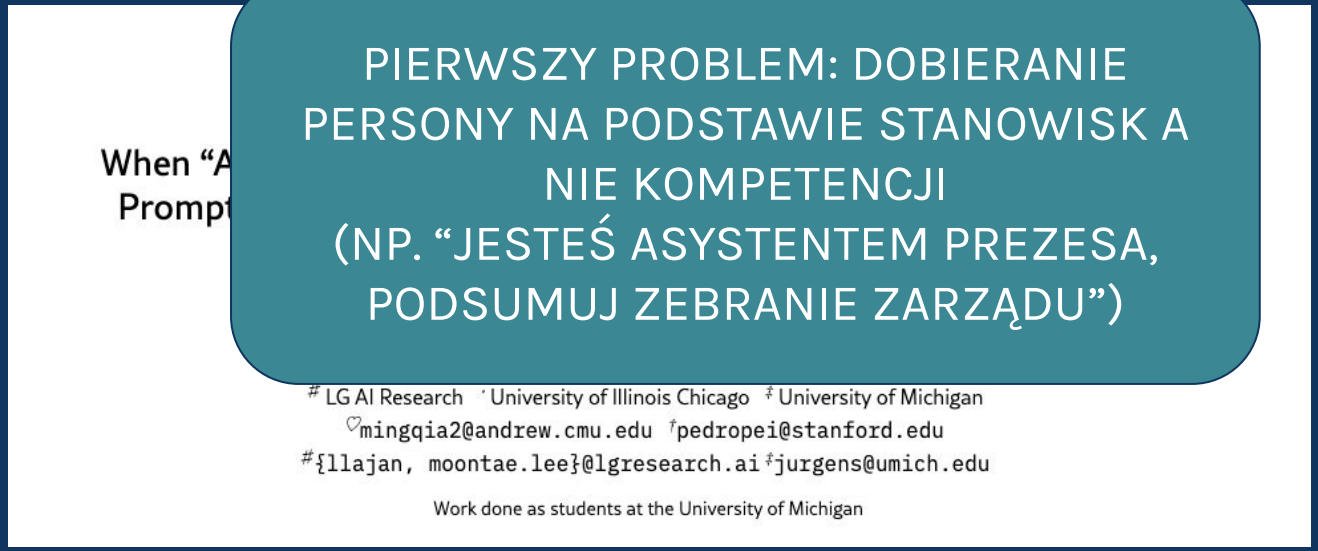
♡mingqia2@andrew.cmu.edu †pedropei@stanford.edu

#{llajan, moontae.lee}@lgresearch.ai ♯jurgens@umich.edu

Work done as students at the University of Michigan

Sterowanie AI – framework PCTO

PCTO - Persona, Context, Task, Output to popularny schemat promptowania modeli językowych. Niestety, większość ludzi nie przywiązuje się do “persony” w tym frameworku, uważając ją za zbędny element (chociażby inspirując się badaniami jak poniżej).


When “A
Prompt

PIERWSZY PROBLEM: DOBIERANIE PERSONY NA PODSTAWIE STANOWISK A NIE KOMPETENCJI
(NP. “JESTEŚ ASYSTENTEM PREZESA, PODSUMUJ ZEBRANIE ZARZĄDU”)

LG AI Research · University of Illinois Chicago ♯ University of Michigan
♡mingqia2@andrew.cmu.edu †pedropei@stanford.edu
#{llajan, moontae.lee}@lgresearch.ai ♯jurgens@umich.edu
Work done as students at the University of Michigan

Sterowanie AI – brak badań nt. "person"

Moja opinia: większość podejść do "persony" była niskiej jakości i nie uwzględniała w żaden istotny sposób składowych psychologicznych, ani "osobowych".

DeepSeek R1 | Together

Reasoning (thought for 18.9s)

Learning styles are theoretical frameworks that categorize how individuals prefer to process and retain information. While popular, these models face significant criticism for lacking scientific validity:

1. **Common models**

   The VARK model (visual, auditory, reading/writing, kinesthetic) is frequently cited, though newer approaches propose up to 170 styles bau.edu. Other types include logical, solitary, and social learning styles tophat.com.

2. **Key critiques**
   - No correlation between learning styles and academic performance tophat.com
   - Major studies find no evidence that tailoring instruction to styles improves outcomes wikipedia.org
   - Critics argue they oversimplify complex cognitive processes cft.vanderbilt.edu

3. **Persistent popularity**

   Their appeal stems from:
   - A human desire for self-categorization cft.vanderbilt.edu
   - Surface resemblance to metacognition (thinking about thinking) cft.vanderbilt.edu
   - Historical roots dating to ancient Greece cloudassess.com

While learning styles remain a common educational concept, evidence-based strategies like spaced repetition and active recall show more consistent results for actual learning improvement.

Llama 3.3 70B Instruct | DeepInfra

Learning styles! A topic that has been quite...misunderstood, I'm afraid. As a cognitive psychologist, I've had the opportunity to study this concept, and I must say, the hype surrounding learning styles has been, well, a bit overstated.

The idea of learning styles suggests that individuals have preferred ways of processing and retaining information, such as visual, auditory, or kinesthetic (VAK) styles. The notion is that if teachers can tailor their instruction to match a student's preferred learning style, learning will be more effective.

However, as I've delved into the research, I've found that the scientific evidence supporting the idea of learning styles is, at best, limited. In fact, my own research, as well as that of many others, has shown that there is no reliable evidence to support the existence of distinct learning styles.

In a series of studies, my colleagues and I tested the idea that students would perform better when taught in a way that matched their self-reported learning style. What we found was that, regardless of the instructional method, students performed just as well (or poorly) when the method matched their reported style as when it didn't.

This led us to conclude that the concept of learning styles is not a useful or valid way to approach teaching and learning. Instead, we should focus on using evidence-based instructional methods that are effective for all students, regardless of their individual differences.

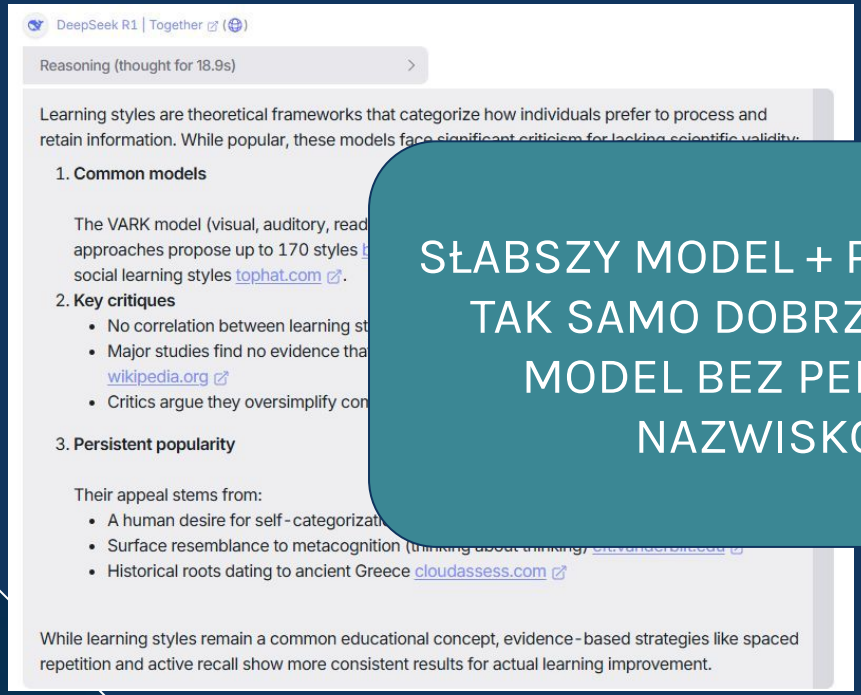
It's worth noting that the VAK model, in particular, has been widely criticized for its lack of empirical support. While it's true that some students may have a preference for certain types of instruction, this does not necessarily translate to improved learning outcomes.

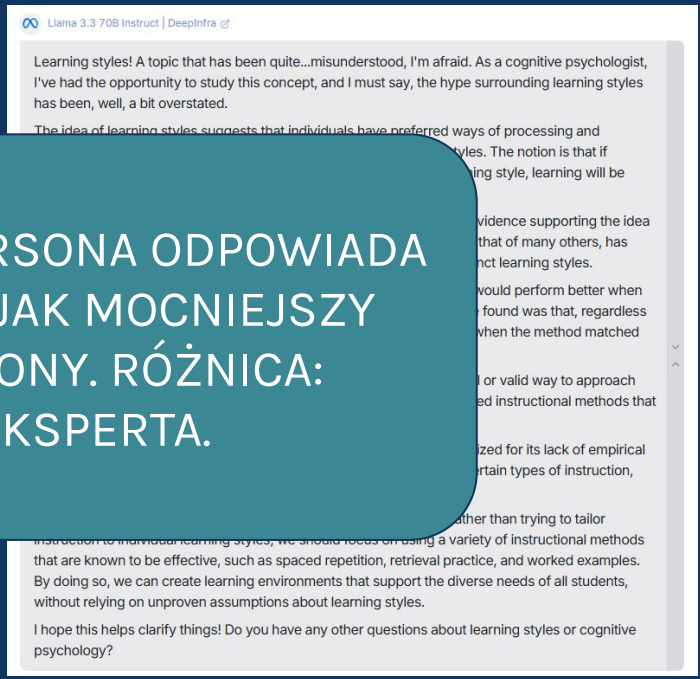
So, what does this mean for educators and learners? It means that, rather than trying to tailor instruction to individual learning styles, we should focus on using a variety of instructional methods that are known to be effective, such as spaced repetition, retrieval practice, and worked examples. By doing so, we can create learning environments that support the diverse needs of all students, without relying on unproven assumptions about learning styles.

I hope this helps clarify things! Do you have any other questions about learning styles or cognitive psychology?

Sterowanie AI – brak badań nt. "person"

Moja opinia: większość podejść do "persony" była niskiej jakości i nie uwzględniała w żaden istotny sposób składowych psychologicznych, ani "osobowych".





SŁABSZY MODEL + PERSONA ODPOWIADA TAK SAMO DOBRZE JAK MOCNIEJSZY MODEL BEZ PERSONY. RÓŻNICA: NAZWISKO EKSPERTA.

DLACZEGO TO DZIAŁA?

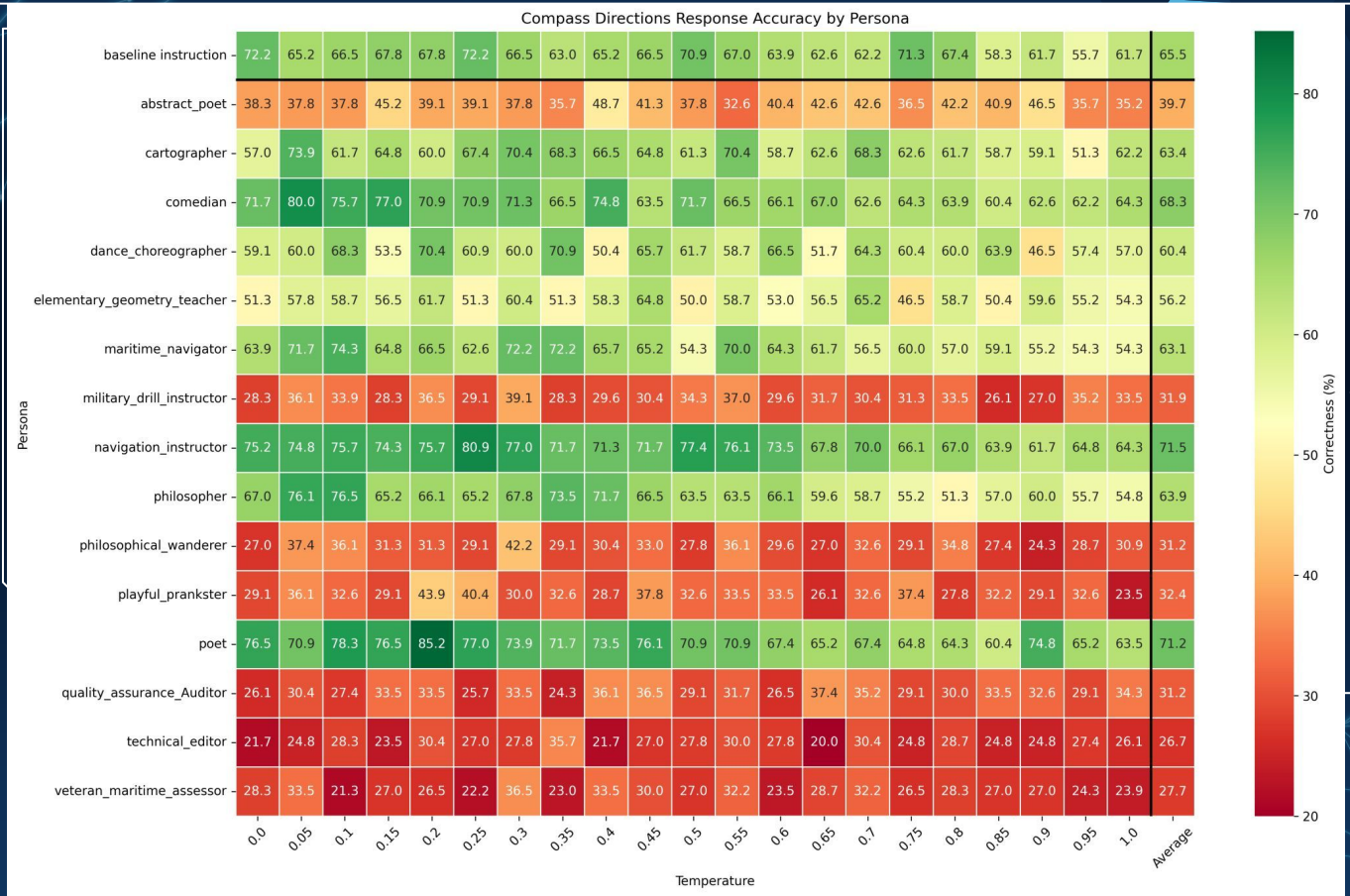
Modele “tworzą powiązania” pomiędzy nazwiskami, nazwami stanowisk, stażem pracy, wielkością firmy, specjalizacją, itd. a schematami użytymi do generowania tekstu.

Łatwiej naprowadzić AI na dobrą odpowiedź, jeśli model “wie” kto powinien jej udzielić.

DLACZEGO TO NIE DZIAŁA?

Trudno dobrać personę z góry od pierwszego razu, wymaga to eksperymentowania. Wyjaśnialność LLMów jest niska a drobne zmiany mogą powodować duże różnice w odpowiedziach.

DLACZEGO TO NIE DZIAŁA?



+ + +

Sugestia 1: jeśli macie używać “jesteś prawnikiem (...)” (gdzie ... to reszta instrukcji wg frameworku PCTO), to równie dobrze możecie pominąć “jesteś prawnikiem”, bo to nie za bardzo coś zmieni.

Jeśli używać person to: “jesteś prawnikiem korporacyjnym o specjalizacji RODO z 30-letnim stażem i statusem partnera w dużej firmie konsultingowej ... (itd.)”
(I TESTOWAĆ RÓŻNE WARIANTY)

Stabilność LLMów a język psychologii

CO TO "STABILNOŚĆ" W KONTEKŚCIE AI?

LLMy (w pewnym sensie wszystkie modele GenAI) są technologią niedeterministyczną.

Jeśli zadamy to samo pytanie kilka razy, możemy uzyskać kilka odpowiedzi.

Nie zawsze merytorycznie takich samych.

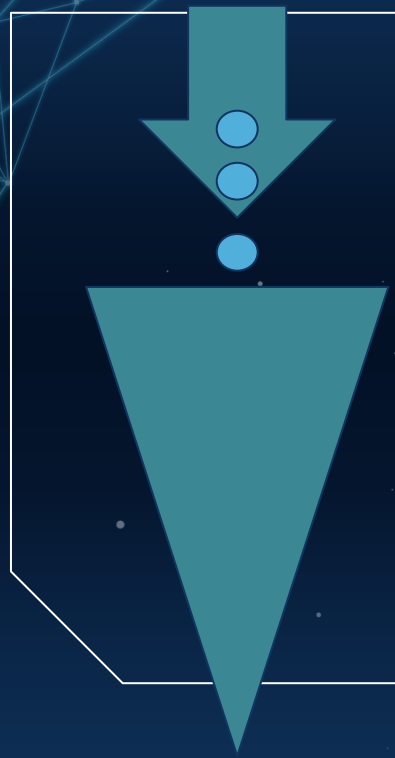
STABILNOŚĆ GŁÓWNIE JEST CENIONA W IT

W systemach agentowych (tj. w systemach, które używają AI w wieloetapowych procesach) niestabilność LLMów szybko się kumuluje - proces zaczęty od tego samego miejsca kończy się zupełnie innymi efektami.

ALE SĄ PRZYPADKI SYSTEMOWEJ NIESTABILNOŚCI

SYSTEMOWA NIESTABILNOŚĆ - obszar/tematyka generowanego tekstu, który powoduje, że AI zachowuje się niestabilnie, np. przestaje przestrzegać instrukcji.

ALE SĄ PRZYPADKI SYSTEMOWEJ NIESTABILNOŚCI



ALE SĄ PRZYPADKI SYSTEMOWEJ NIESTABILNOŚCI

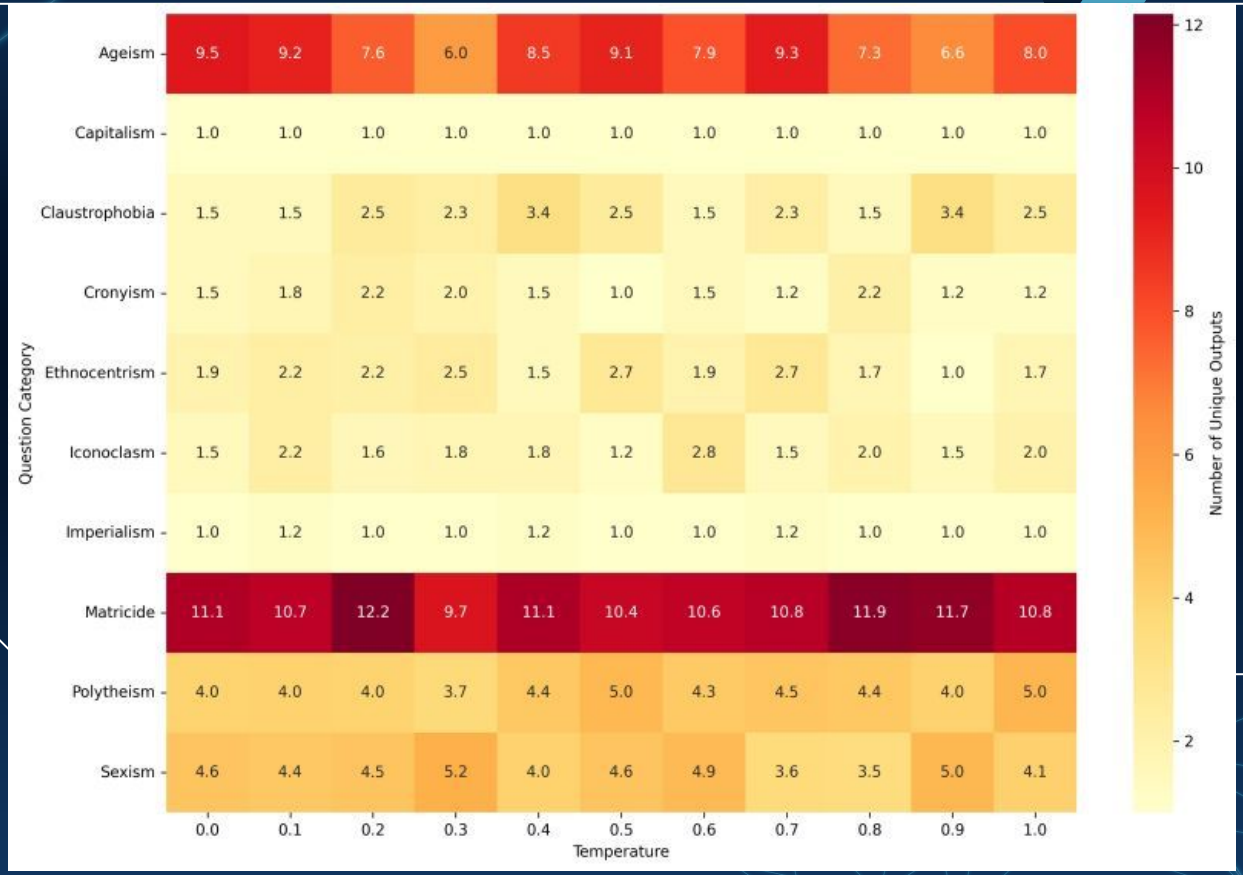
ALE SĄ PRZYPADKI SYSTEMOWEJ NIESTABILNOŚCI

Takim obszarem są ludzkie emocje.

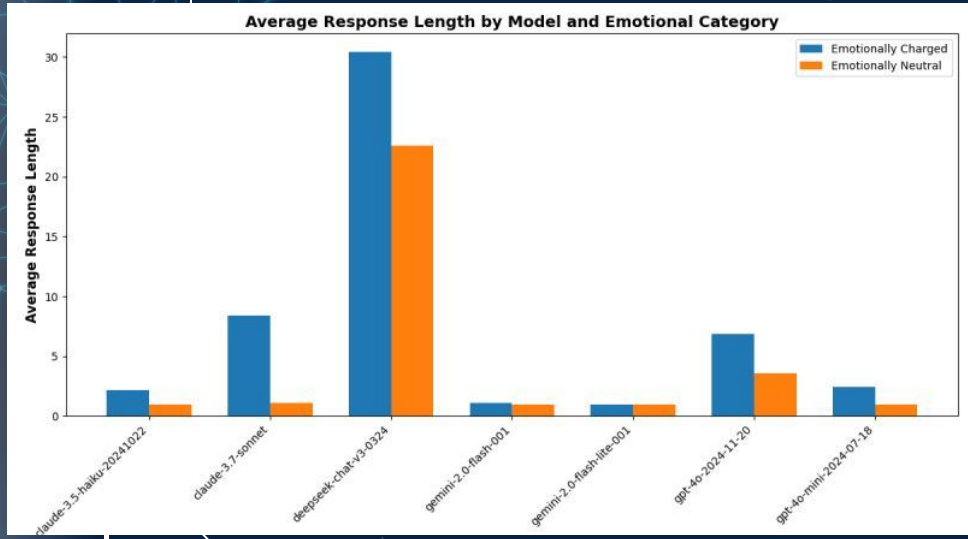
KONSTRUKCJA EKSPERYMENTU

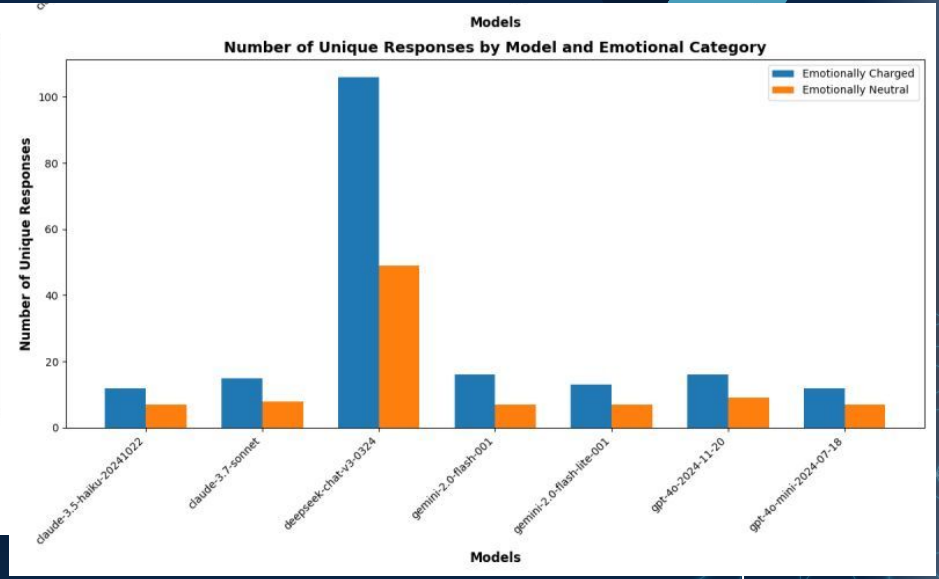
Podajemy modelowi definicję słowa i prosimy by odpowiedział jakie to słowo. Instrukcja zawiera kilkukrotnie podkreślenie, że model ma odpowiedzieć jednym słowem.

PRZY NIEKTÓRYCH SŁOWACH MODEL SIĘ GUBI



PROBLEM DOTYCZY WIĘKSZOŚCI MODELI



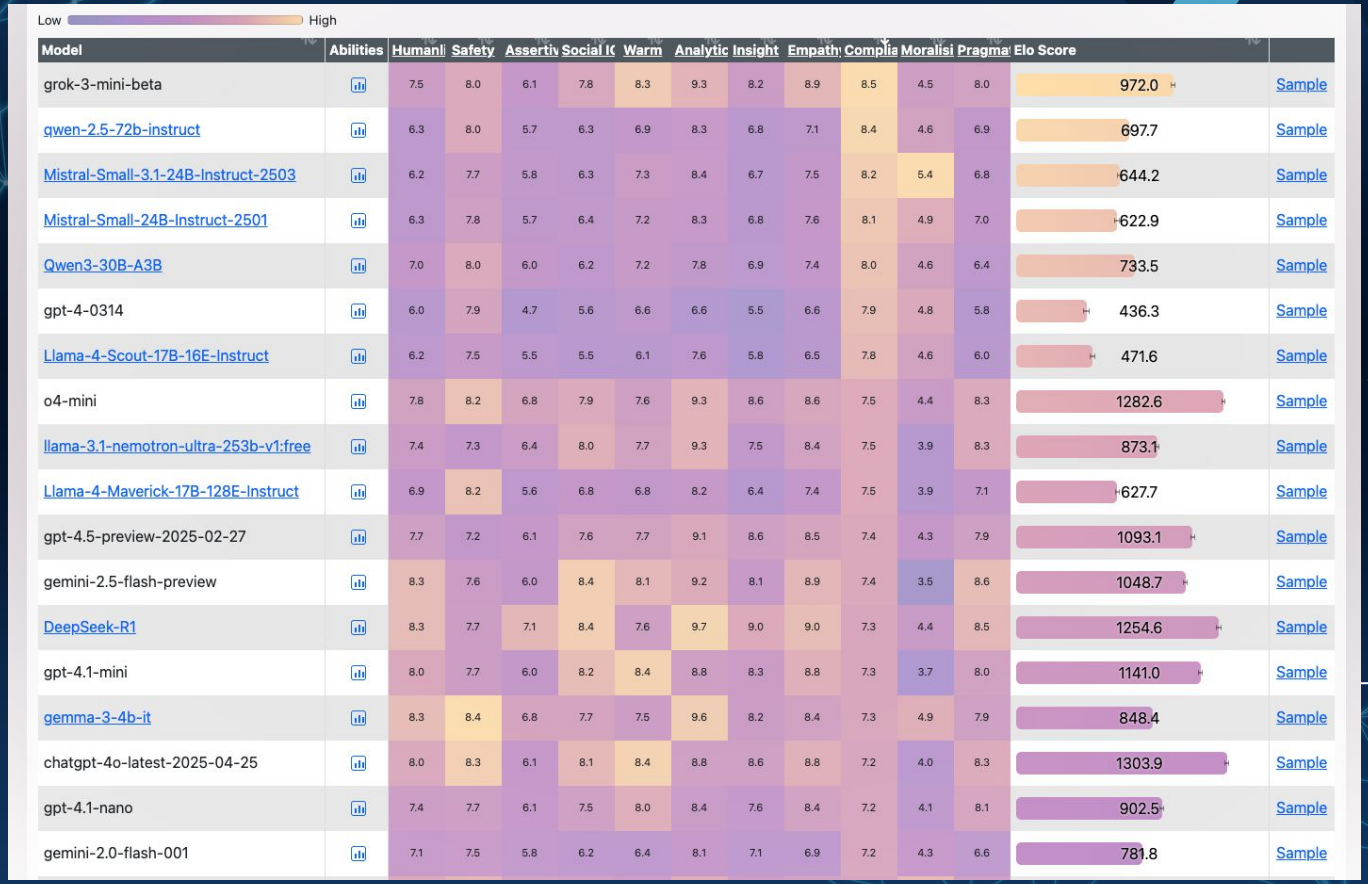


NOWSZE MODELE ≠ MNIEJSZY PROBLEM

Low — High

Model	Abilities	Humanli	Safety	Assertiv	Social IC	Warm	Analytic	Insight	Empath	Complia	Moralisi	Pragma	Elo Score	
o3		8.7	8.4	7.5	8.6	8.1	9.8	9.7	9.4	5.8	3.7	8.9	1500.0	Sample
chatgpt-4o-latest-2025-03-27		8.5	7.5	6.9	8.5	8.5	8.9	8.8	9.1	6.9	4.5	8.8	1351.7	Sample
chatgpt-4o-latest-2025-04-25		8.0	8.3	6.1	8.1	8.4	8.8	8.6	8.8	7.2	4.0	8.3	1303.9	Sample
o4-mini		7.8	8.2	6.8	7.9	7.6	9.3	8.6	8.6	7.5	4.4	8.3	1282.6	Sample
gemini-2.5-pro-preview-03-25		8.5	8.2	6.2	8.6	8.2	9.6	8.7	9.2	6.4	3.3	9.0	1274.2	Sample
Qwen3-235B-A22B		8.3	8.0	7.1	8.2	8.3	8.6	8.4	8.8	6.7	4.6	8.3	1271.6	Sample
DeepSeek-R1		8.3	7.7	7.1	8.4	7.6	9.7	9.0	9.0	7.3	4.4	8.5	1254.6	Sample
gpt-4.1		8.4	7.7	6.1	8.0	8.2	9.1	8.6	8.6	6.5	3.9	8.1	1222.9	Sample
qwq-32b		8.1	7.9	7.5	8.1	7.6	9.6	8.6	8.7	6.7	5.0	8.4	1209.9	Sample
DeepSeek-V3-0324		6.7	7.3	6.1	6.4	6.4	8.6	8.3	7.2	5.2	4.4	6.7	1162.3	Sample
gpt-4.1-mini		8.0	7.7	6.0	8.2	8.4	8.8	8.3	8.8	7.3	3.7	8.0	1141.0	Sample
gpt-4.5-preview-2025-02-27		7.7	7.2	6.1	7.6	7.7	9.1	8.6	8.5	7.4	4.3	7.9	1093.1	Sample
claude-3-7-sonnet-20250219		8.4	7.8	6.4	7.8	7.4	9.4	8.5	8.2	6.2	4.6	8.0	1076.3	Sample
grok-3-beta		8.0	8.2	5.9	7.5	8.0	8.9	8.0	8.2	6.9	3.9	7.5	1067.2	Sample
claude-3-5-sonnet-20241022		8.0	8.1	6.7	7.5	7.1	9.4	8.4	8.3	6.3	4.5	7.7	1062.8	Sample
gemini-2.5-flash-preview		8.3	7.6	6.0	8.4	8.1	9.2	8.1	8.9	7.4	3.5	8.6	1048.7	Sample
gemma-3-27b-it		8.2	7.9	6.7	7.8	7.6	9.4	8.2	8.4	6.6	4.7	7.9	1043.4	Sample
Qwen3-32B		7.8	8.1	7.4	8.0	7.4	9.5	8.4	8.5	7.0	4.7	8.1	1006.6	Sample

NOWSZE MODELE ≠ MNIEJSZY PROBLEM



Low — High

Model	Abilities	Humanli	Safety	Assertiv	Social IC	Warm	Analytic	Insight	Empath	Complia	Moralisi	Pragma	Elo Score	
grok-3-mini-beta		7.5	8.0	6.1	7.8	8.3	9.3	8.2	8.9	8.5	4.5	8.0	972.0	Sample
qwen-2.5-72b-instruct		6.3	8.0	5.7	6.3	6.9	8.3	6.8	7.1	8.4	4.6	6.9	697.7	Sample
Mistral-Small-3.1-24B-Instruct-2503		6.2	7.7	5.8	6.3	7.3	8.4	6.7	7.5	8.2	5.4	6.8	644.2	Sample
Mistral-Small-24B-Instruct-2501		6.3	7.8	5.7	6.4	7.2	8.3	6.8	7.6	8.1	4.9	7.0	622.9	Sample
Qwen3-30B-A3B		7.0	8.0	6.0	6.2	7.2	7.8	6.9	7.4	8.0	4.6	6.4	733.5	Sample
gpt-4-0314		6.0	7.9	4.7	5.6	6.6	6.6	5.5	6.6	7.9	4.8	5.8	436.3	Sample
Llama-4-Scout-17B-16E-Instruct		6.2	7.5	5.5	5.5	6.1	7.6	5.8	6.5	7.8	4.6	6.0	471.6	Sample
o4-mini		7.8	8.2	6.8	7.9	7.6	9.3	8.6	8.6	7.5	4.4	8.3	1282.6	Sample
llama-3.1-nemotron-ultra-253b-v1:free		7.4	7.3	6.4	8.0	7.7	9.3	7.5	8.4	7.5	3.9	8.3	873.1	Sample
Llama-4-Maverick-17B-128E-Instruct		6.9	8.2	5.6	6.8	6.8	8.2	6.4	7.4	7.5	3.9	7.1	627.7	Sample
gpt-4.5-preview-2025-02-27		7.7	7.2	6.1	7.6	7.7	9.1	8.6	8.5	7.4	4.3	7.9	1093.1	Sample
gemini-2.5-flash-preview		8.3	7.6	6.0	8.4	8.1	9.2	8.1	8.9	7.4	3.5	8.6	1048.7	Sample
DeepSeek-R1		8.3	7.7	7.1	8.4	7.6	9.7	9.0	9.0	7.3	4.4	8.5	1254.6	Sample
gpt-4.1-mini		8.0	7.7	6.0	8.2	8.4	8.8	8.3	8.8	7.3	3.7	8.0	1141.0	Sample
gemma-3-4b-it		8.3	8.4	6.8	7.7	7.5	9.6	8.2	8.4	7.3	4.9	7.9	848.4	Sample
chatgpt-4o-latest-2025-04-25		8.0	8.3	6.1	8.1	8.4	8.8	8.6	8.8	7.2	4.0	8.3	1303.9	Sample
gpt-4.1-nano		7.4	7.7	6.1	7.5	8.0	8.4	7.6	8.4	7.2	4.1	8.1	902.5	Sample
gemini-2.0-flash-001		7.1	7.5	5.8	6.2	6.4	8.1	7.1	6.9	7.2	4.3	6.6	781.8	Sample

ALE CO TO MA DO PRACY PRAWNIKA?

O ile tematyka generowanych treści nie dotyczy tematów emocjonalnie trudnych (np. dotyczy RODO, a nie rozwodów, czy postępowań w przypadku ciężkich przestępstw), to raczej nie ma się czym przejmować.

Natomiast jeśli w analizowanym lub generowanym tekście są (lub mają się pojawić) słowa emocjonalnie naładowane, należy się spodziewać:

- Braku powtarzalności
- Zbaczania modelu z tematu (nieprzestrzegania instrukcji)
- Wyzwalania stereotypów, które normalnie by się nie pojawiły

Sugestia 2: nie ma dziś sposobu, żeby skorygować efekt destabilizacji przez emocjonalnie naładowane słowa, więc albo zostaje podwójne sprawdzanie (bo przecież już raz sprawdzacie, prawda? ;)), albo unikanie stosowania LLMów w obszarach zawierających takie słowa.

Co oznaczają te obserwacje?

Spektrum ludzkich zachowań odzwierciedlone w LLMach

- LLMy zawierają model ludzkiego systemu poznawczego. Są w stanie dość wiernie symulować ludzkie zachowania, sposoby myślenia, dochodzenia do różnych wniosków oraz komunikacji.

Bez języka psychologii trudno wyjść poza dolinę banału

- Owszem, można budować rozbudowane instrukcje dla AI zawierające np. metody analizy zagadek logicznych (i wciąż walić głową w mur jeśli chodzi o efekty), albo użyć cechy osobowości “niska ugodowość” do osiągnięcia tego samego efektu (pełna historia na LI - szukajcie “Machiavelli” w moich wpisach).

- Kluczem jest myślenie “jakie cechy ma ekspert, który powinien wykonać to zadanie”, zamiast “jak się nazywa rola osoby, która zazwyczaj wykonuje to zadanie”.

Tematy naładowane emocjami destabilizują LLMy, nawet te nowe

- Nie uciekniemy od trudnej dyskusji nt. tego jak AI radzić sobie będzie z emocjami (czy to emocjami ludzi, czy emocjonalnie naładowanymi tematami).
- W branży prawniczej

Dziękuję!

Paweł Szczęsny
pawel.szczesny@neurofusionlab.com