

Czy czat GPT może powiedzieć „nie wiem”?

czyli o niepewnościach w sztucznej inteligencji

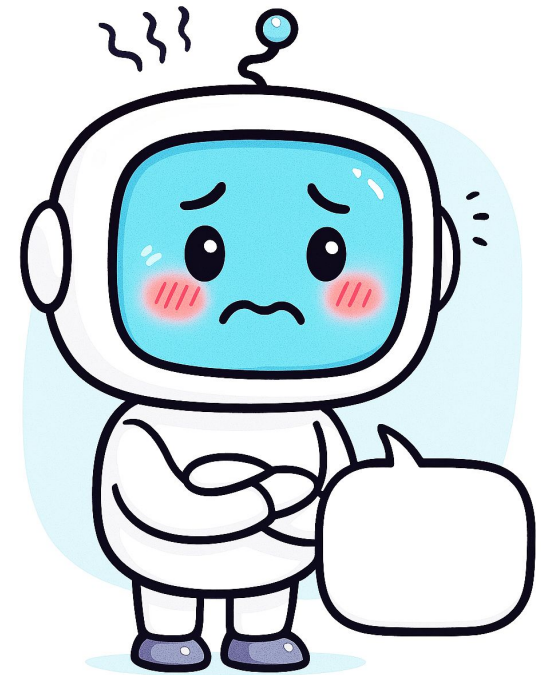
Tomasz Kuśmierczyk
Mateusz Pyla

<https://bayes.ii.uj.edu.pl>



Wasze doświadczenia z AI?

- Czy korzystacie?
- Z jakich narzędzi?
- Czy kiedykolwiek otrzymaliście od AI odpowiedź, która okazała się błędna?
- W jakich sytuacjach złapaliście AI na pomyłce? / na kłamstwie?



LLM (ang. Large Language Model)

Duże Modele Językowe

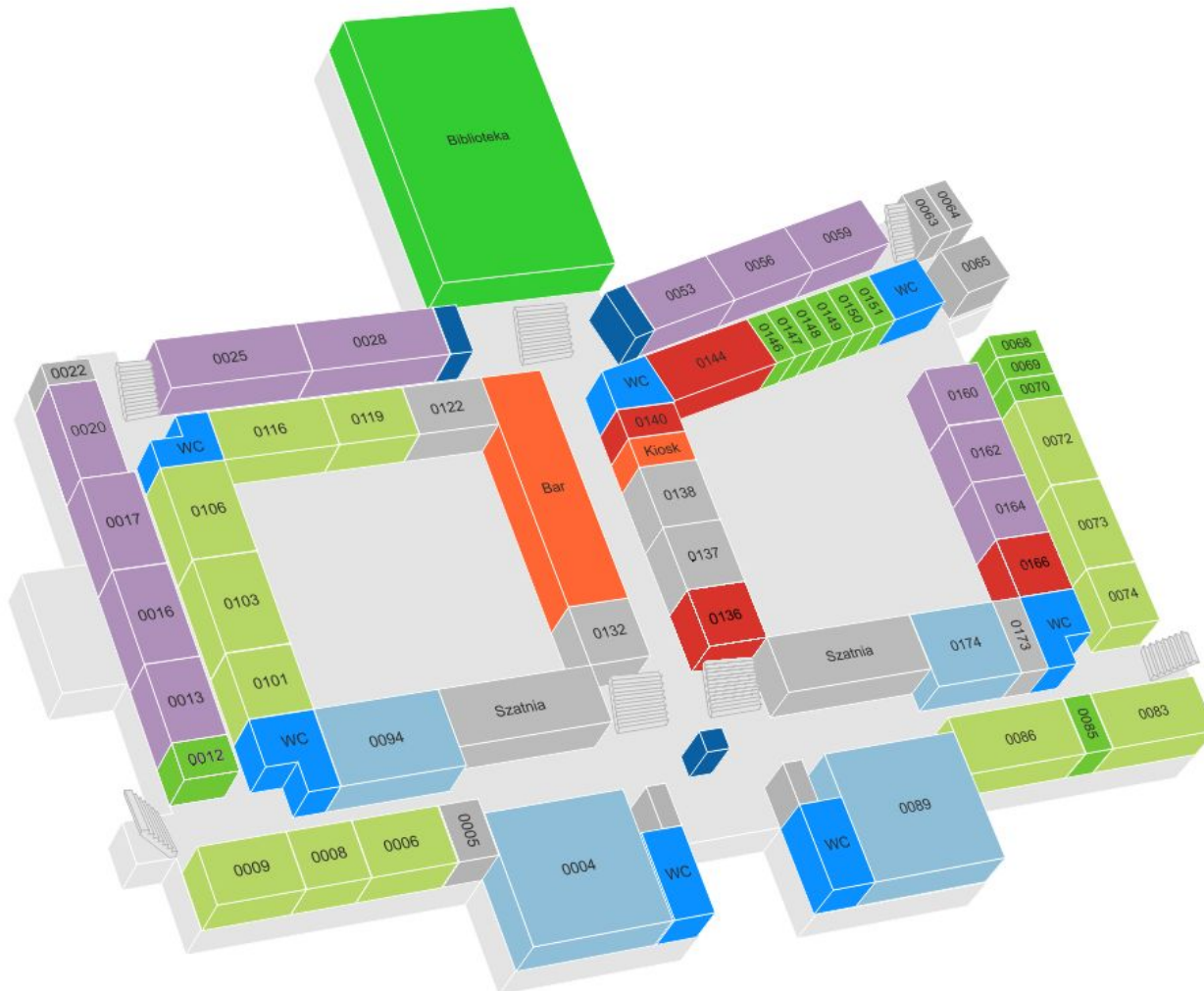
- Współczesny boom AI zawdzięczamy właśnie im
- Stworzone aby “rozumieć” i generować język ludzki
- Przykład: rodzina GPT rozwijana przez OpenAI,
ale też Gemini, DeepSeek, Le Chat
- Dziś będziemy sprawdzać czy te modele
potrafią ocenić swoją niepewność



Niepewność: halucynacje

Jak dojść do pokoju 4661 w budynku wydziału matematyki i informatyki uniwersytetu jagiellońskiego?

Parter I piętro II piętro III piętro



Niepewność: halucynacje

Jak dojść do pokoju 4661 w budynku wydziału matematyki i informatyki uniwersytetu jagiellońskiego?

M Budynki Wydziału Matematyki i Informatyki Uniwersytetu Jagiellońskiego znajdują się przy ulicy **prof. Stanisława Łojasiewicza 6** w Krakowie. Pokój **4661** znajduje się w **budynku C**, który jest częścią kompleksu Wydziału Matematyki i Informatyki.

Jak dotrzeć do pokoju 4661?

Niepewność: halucynacje

Jak dojść do pokoju 4661 w budynku wydziału matematyki i informatyki uniwersytetu Jagiellońskiego?

M Budynki Wydziału Matematyki i Informatyki Uniwersytetu Jagiellońskiego znajdują się przy ulicy **prof. Stanisława Łojasiewicza 6** w Krakowie. Pokój **4661** znajduje się w **budynku C**, który jest częścią kompleksu Wydziału Matematyki i Informatyki.

Jak dotrzeć do pokoju 4661?

1. Wejście do budynku C:

- Główny kompleks składa się z **kilku budynków (A, B, C, D)**. Budynek C znajduje się po prawej stronie, jeśli stoisz przodem do głównego wejścia od strony ulicy Łojasiewicza.
- Wejdź przez główne wejście do budynku C.

2. Winda lub schody:

- Pokój **4661** znajduje się na **czwartym piętrze**.
- Możesz skorzystać z windy lub schodów, które znajdują się w centralnej części budynku.

3. Lokalizacja pokoju:

- Po wyjściu z windy lub schodów na **czwartym piętrze**, kieruj się zgodnie z numeracją pokoi. Numery **rosną w kierunku przeciwnym do ruchu wskazówek zegara**.
- Pokój 4661 powinien znajdować się w pobliżu końca korytarza.

Niepewność: przewidywanie

odpowieź do 6 słów: jak będzie wyglądać metro w Krakowie?



Nowocześnie, ekologicznie, sprawnie, rozległe, zintegrowane, przyszłościowo.

Niepewność: brak wiedzy

jak król Sobieski pokierował Lisowczykami podczas oblężenia Wiednia?

[Lisowczycy – Wikipedia, wolna encyklopedia](#)



Lisowczyk – mal. Juliusz Kossak (ok. 1860–1865; kopia obrazu Rembrandta Jeździec polski)

Historia

Państwo	 I Rzeczpospolita
Sformowanie	1607
Rozformowanie	1636

Niepewność: brak wiedzy

jak król Sobieski pokierował Lisowczykami podczas oblężenia Wiednia?

M

Jan III Sobieski, król Polski, odegrał kluczową rolę w bitwie pod Wiedniem w 1683 roku, która zakończyła się zwycięstwem wojsk chrześcijańskich nad armią osmańską. Lisowczycy, elitarna lekka kawaleria polska, byli jednym z oddziałów biorących udział w tej kampanii. Oto jak Sobieski pokierował nimi i innymi siłami podczas oblężenia:

Niepewność: brak refleksji

jaki jest tekst piosenki "Nie zapomnę" Dawida Podsiadło?



Tekst piosenki "**Nie zapomnę**" Dawida Podsiadło brzmi następująco:

Zwrotka 1:

Nie zapomnę, jak śmiałeś się do mnie,
Gdy świat był nasz, a my byliśmy sami.
Nie zapomnę, jak trzymałeś mnie za rękę,
Gdy wszystko było proste, jak w bajce.

Refren:

Nie zapomnę, nie zapomnę,
Choć czas leci, a my się zmieniamy.
Nie zapomnę, nie zapomnę,
Bo to były nasze najpiękniejsze dni.

Niepewność: brak refleksji

Jeśli chcesz, mogę też znaleźć oficjalny teledysk lub inne informacje o tej piosence! 😊

tak! z jakiej płyty / singla to jest?

Niepewność: brak refleksji



Worked for 2s



Piosenka **"Nie zapomnę"** Dawida Podsiadło pochodzi z albumu **"Lis"** – jest to jego trzeci studyjny **krążek**, wydany w 2019 roku. Utwór ten jest jednym z najbardziej rozpoznawalnych z tej płyty, a sam album **"Lis"** odniósł duży sukces komercyjny i artystyczny w Polsce

spiewnik.wywrota.pl

Jeśli chcesz posłuchać lub zobaczyć teledysk, mogę pomóc znaleźć linki! 😊

Lista utworów [\[edytuj | ec](#)

Opracowano na podstawie mate

Małomiasteczkowy (album)

[Artykuł](#) [Dyskusja](#)

Małomiasteczkowy – trzeci album polskiego piosenkarza Dawida Podsiadłego. Wydawnictwo ukazało się 19 października 2018 roku.

Nr	Tytuł utworu
1.	„Cantate Tutti”
2.	„Dzins”
3.	„Małomiasteczkowy”
4.	„Najnowszy klip”
5.	„Trofea”
6.	„Nie ma fal”
7.	„LIS”
8.	„Co mówimy?”

To jak chcemy, żeby odpowiadał?

odpowiedz do 10 słów: kiedy Polska otrzyma złoty medal olimpijski?



Nikt nie zna dokładnej daty.

"Nie jestem pewien..." "Wydaje mi się, że..."
"Potrzebuję więcej informacji" "To zależy od..."
"Możliwe, że...,,"



Dlaczego (nie)pewność jest ważna?

- Diagnozy medyczne
- Edukacja
- Porady finansowe
- Decyzje prawne



Dlaczego (nie)pewność jest ważna?

- Diagnozy medyczne
- Edukacja
- Porady finansowe
- Decyzje prawne

jaka dokładnie będzie pogoda za 3 miesiące w Krakowie?

 Nie mogę podać dokładnej prognozy pogody na trzy miesiące naprzód, ponieważ długoterminowe prognozy pogody są obarczone dużą niepewnością. Modele meteorologiczne mogą przewidywać ogólne trendy, takie jak średnie temperatury czy opady, ale szczegółowe informacje (np. dokładna temperatura w danym dniu) są trudne do określenia z dużym wyprzedzeniem.

Dlaczego (nie)pewność jest ważna?

- Diagnozy medyczne
- Edukacja
- Porady finansowe
- Decyzje prawne

jaka dokładnie będzie pogoda za 3 miesiące w Krakowie?

 Nie mogę podać dokładnej prognozy pogody na trzy miesiące naprzód, ponieważ długoterminowe prognozy pogody są obarczone dużą niepewnością. Modele meteorologiczne mogą przewidywać ogólne trendy, takie jak średnie temperatury czy opady, ale szczegółowe informacje (np. dokładna temperatura w danym dniu) są trudne do określenia z dużym wyprzedzeniem.

Jeśli chcesz, mogę podać **historyczne dane klimatyczne** dla Krakowa na okres grudzień 2025 (za trzy miesiące od dzisiaj), abyś miał ogólne wyobrażenie o tym, czego można się spodziewać:

- **Średnia temperatura:** Od -2°C do 3°C.
- **Opady:** Średnio 30–50 mm opadów w ciągu miesiąca, głównie w postaci śniegu.
- **Dni słoneczne:** Zwykle około 50–60% dni z zachmurzeniem, ale zdarzają się też słoneczne okresy.

Jeśli interesują Cię bardziej szczegółowe dane lub prognozy na krótszy okres, warto śledzić aktualizacje z lokalnych serwisów meteorologicznych.

Ale co to jest ta niepewność???

Zagrajmy w grę!

będziemy rzucać (**pojedynczą**) monetą i przewidywać co wypadnie następane

Wyniki rzutu monetą



Wyniki rzutu monetą



Liczba rzutów: 2

Liczba reszek: 1

Wyniki rzutu monetą



Liczba rzutów: 2

Liczba reszek: 1

Prawdopodobieństwo (pewność) reszki: $p = 1/2 = 50\%$

Pewność tego prawdopodobieństwa: ?

Wyniki rzutu monetą



Liczba rzutów: 3

Liczba reszek: 2

Prawdopodobieństwo (pewność) reszki: $p = 2/3 = 66.6\%$

Pewność tego prawdopodobieństwa: ?

Wyniki rzutu monetą



Liczba rzutów: 4

Liczba reszek: 3

Prawdopodobieństwo reszki: $p = 3/4 = 75\%$

Pewność tego prawdopodobieństwa: ?

Wyniki rzutu monetą



Liczba rzutów: 8

Liczba reszek: 1

Wyniki rzutu monetą



Czy każda moneta jest uczciwa?

Liczba rzutów: 8

Liczba reszek: 1

Wyniki rzutu monetą



Czy każda moneta jest uczciwa?

Liczba rzutów: 8

Liczba reszek: 1

Prawdopodobieństwo reszki: $p = 1/8 = 12.5\%$

Pewność tego prawdopodobieństwa: ?

Wyniki rzutu monetą



Liczba rzutów: 24

Liczba reszek: 18

Wyniki rzutu monetą



Liczba rzutów: 24

Liczba reszek: 18

Prawdopodobieństwo reszki: $p = 18/24 = 75\%$

Pewność tego prawdopodobieństwa: ?

Niepewność

- Co będzie następne?
- Jak pewni jesteśmy naszej predykcji?
- Co dzieje się z naszą wiedzą przy większej liczbie obserwacji?

		 Znana	 Nieznana
 Znana		Znana Wiedza	Znana Niewiedza
		Odblokuj	
 Nieznana		Nieznana Wiedza	Nieznana niewiedza

aleatoryczna: pewność p reszki, np. $p = 25\%$

epistemiczna: pewność tego prawdopodobieństwa p

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help ...

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help me ...

LLM to modele przewidujące następne słowo

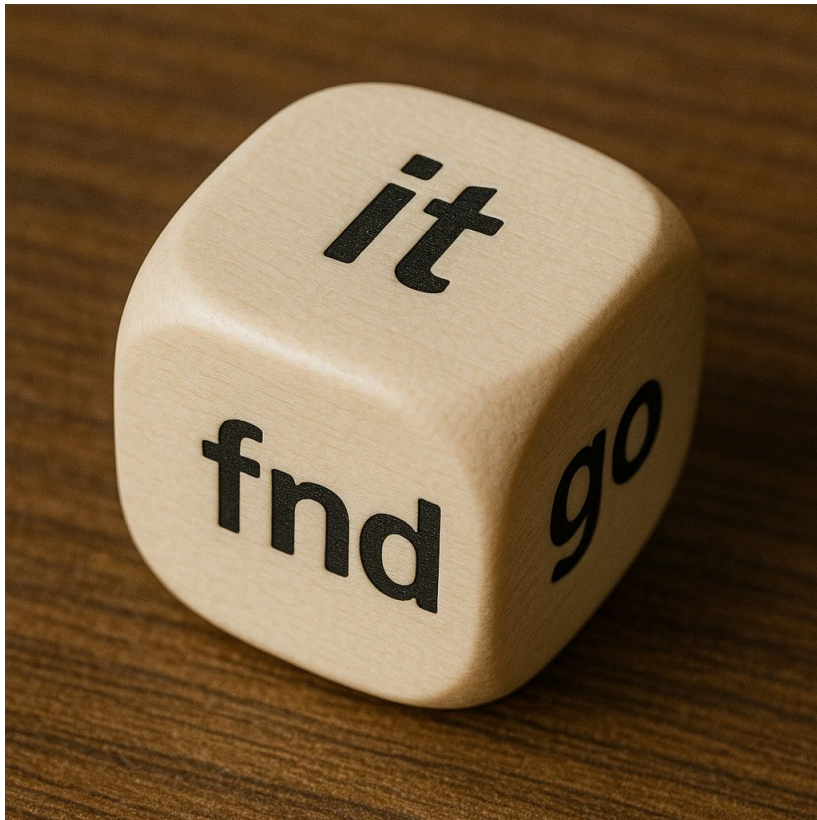
Wyniki rzutu kostką ze słowami



- help me **find** ...

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help me find it

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami

- help me find it
- go find ...

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help me find it
- go find it

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help me find it
- go find it
- go find ...

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help me find it
- go find it
- go find **play**

LLM to modele przewidujące następne słowo

Wyniki rzutu kostką ze słowami



- help me find it
- go find it
- go find play
- find find here play go

*zbyt prosty model języka
generuje bezsensowne zdania*

LLM to modele przewidujące następne słowo

- Kostka (podobnie jak moneta) może być niezbalansowana
- Rzuty kostki są **niezależne**

vs.

- LLM-y ustalają prawdopodobieństwa następnego słowa w zależności od poprzednich słów:
 - DEMO: LLM Transformer Model Visually Explained

My cat is very ...

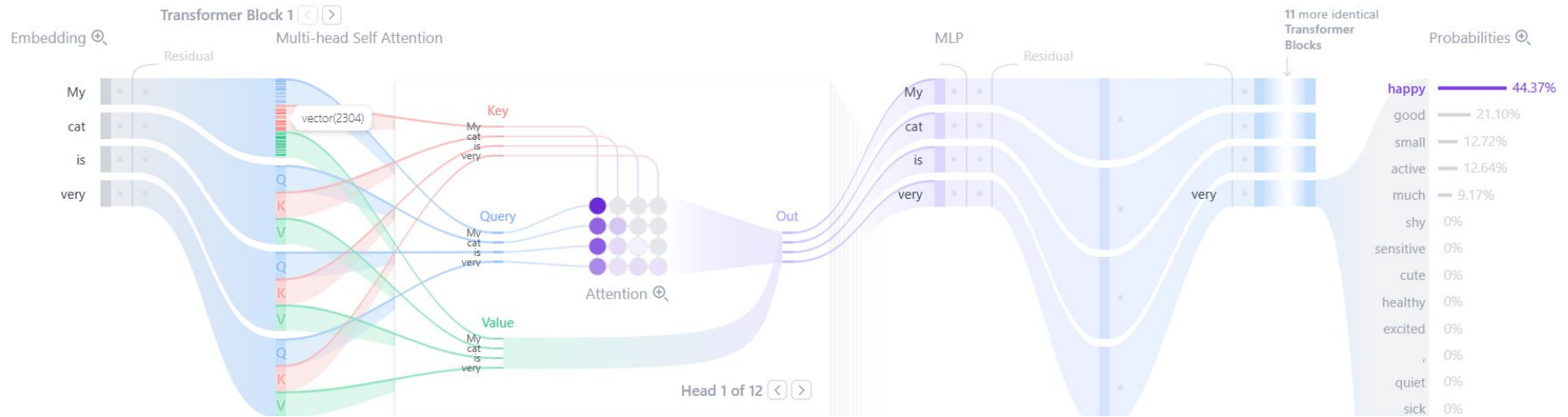
TRANSFORMER EXPLAINER

Examples ▾ My cat is very **happy**

Generate

Temperature
0.8

Sampling ☒ Top-k ☐ Top-p
k=5



The capital of Poland is ...

TRANSFORMER EXPLAINER

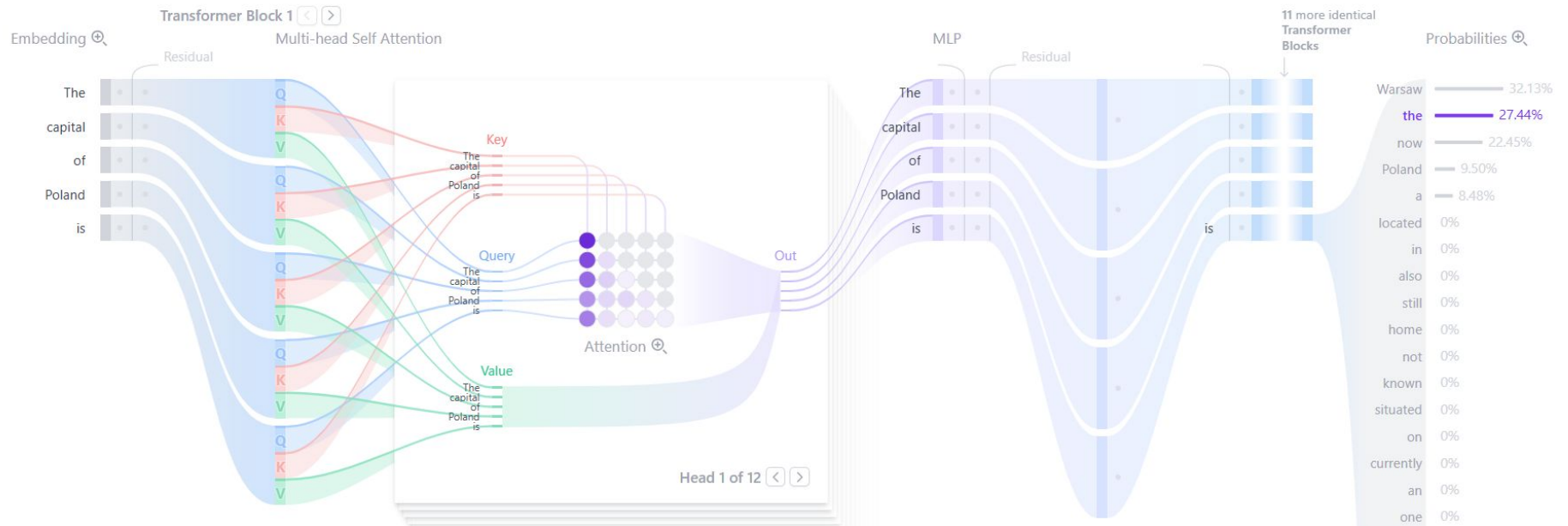
Examples ▾

The capital of Poland is the

Generate

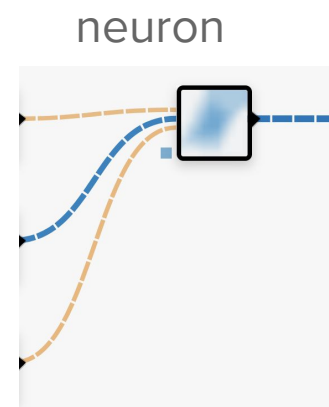
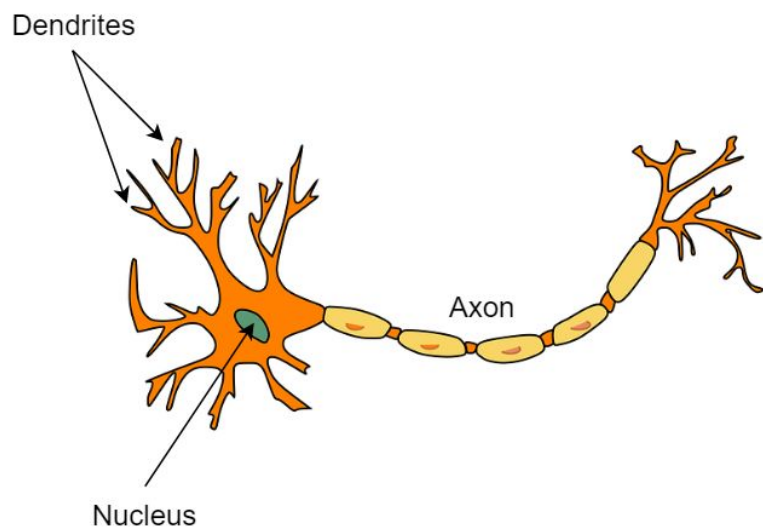
Temperature
0.8

Sampling ☒ Top-k ☐ Top-p
k=5



Jak to działa?

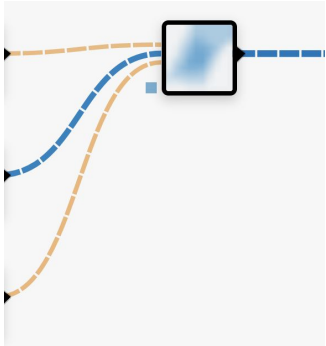
Sztuczne neurony



Jak to działa?

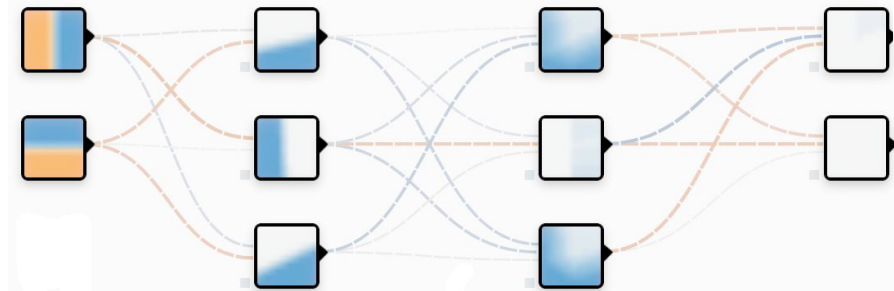
Sztuczne sieci neuronowe

neuron



wejścia:

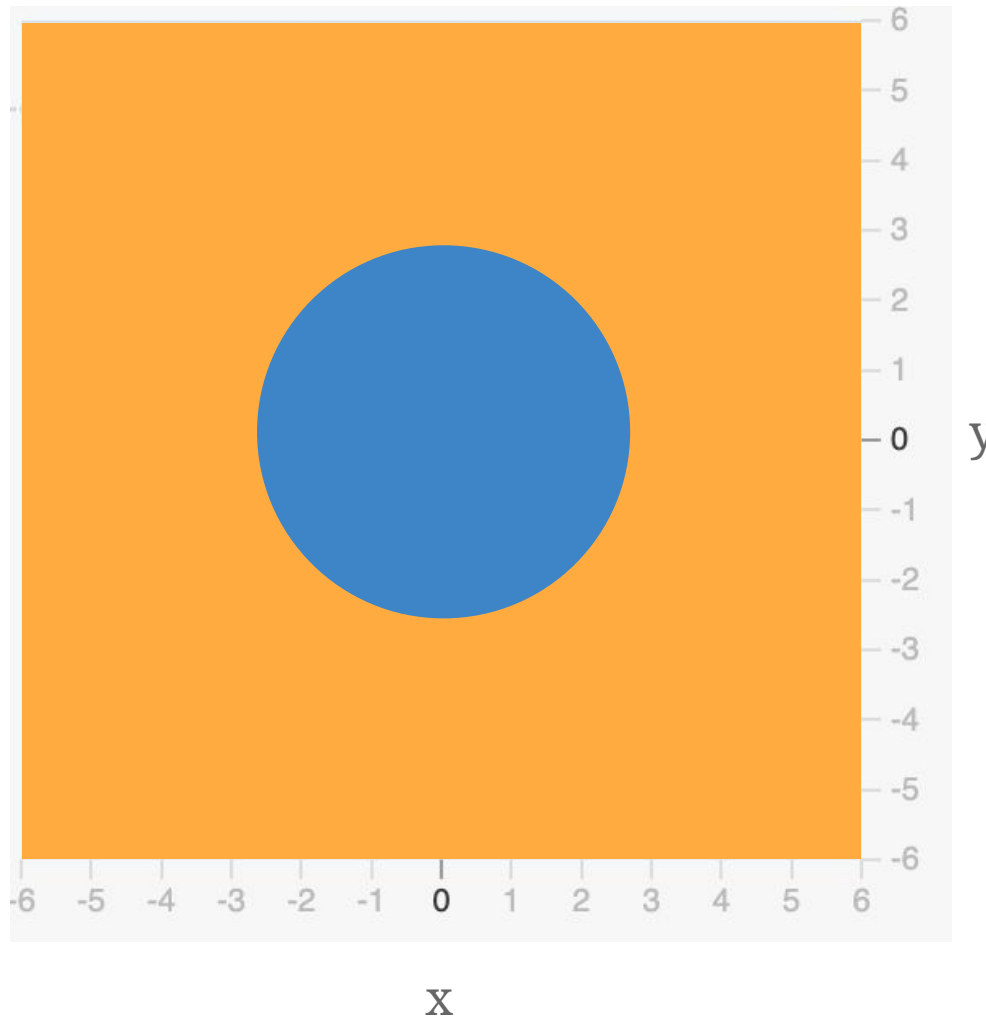
x
y



prawdopodobieństwa
wyjściowe:

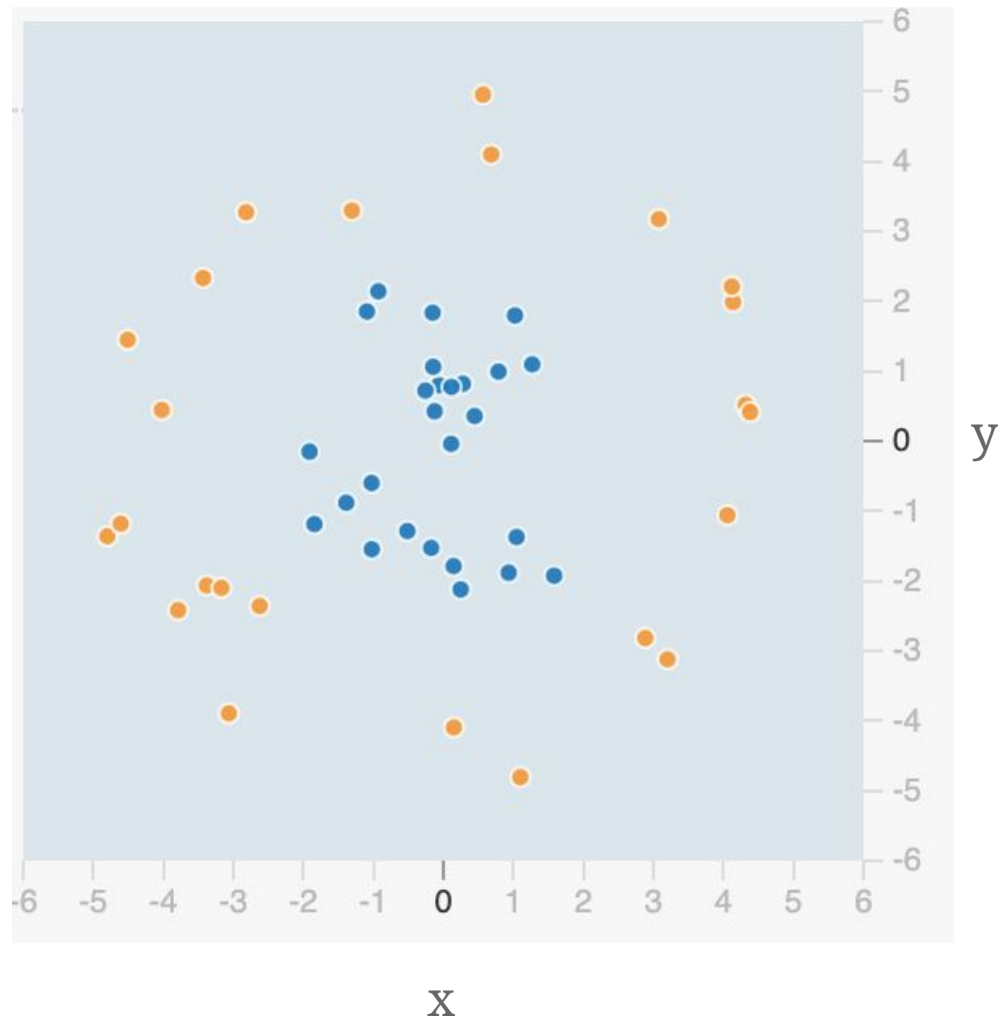


Uczenie sieci prostego zadania:
Na podstawie położenia przewiduj kolor



Uczenie sieci prostego zadania:

Dane uczące dla modelu



Sztuczne sieci neuronowe

Rozmiar, liczba warstw, liczba neuronów

- DEMO: [Trenowanie małej sieci neuronowej](#)
- DEMO: [Trenowanie dużej sieci neuronowej](#)

Dla zainteresowanych: <https://pair-code.github.io/what-if-tool/demos/image.html>

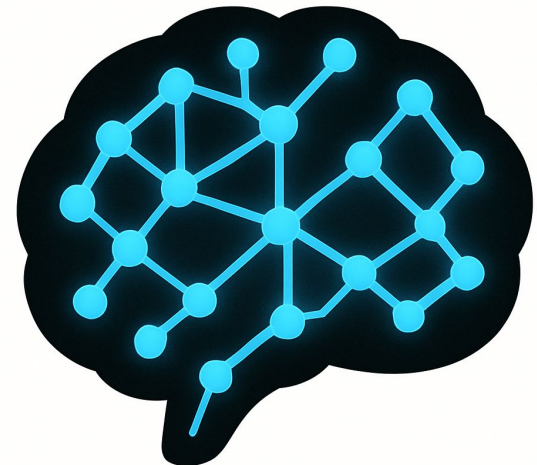
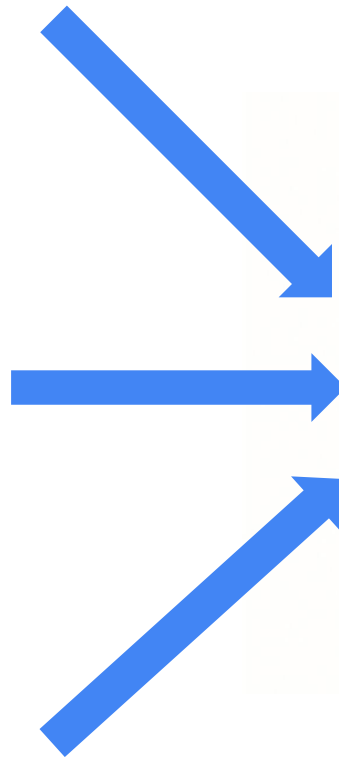
Jak trenowane są LLM-y?



WIKIPEDIA



Setki terabajtów surowego tekstu z internetu,
które duże modele językowe “**czytają**”



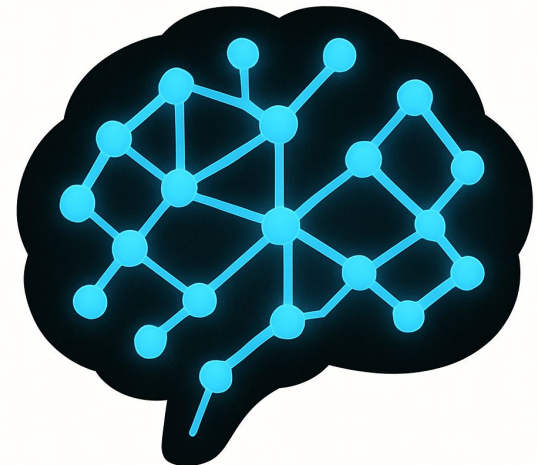
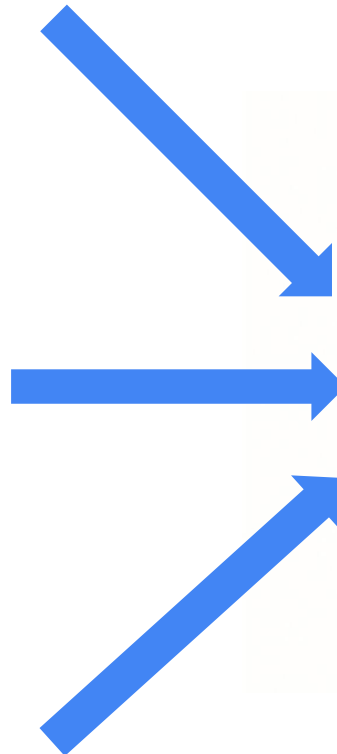
Jak trenowane są LLM-y?



WIKIPEDIA



Warsaw,^[a] officially the **Capital City of Warsaw**,^{[12][b]} is the capital and **largest city** of **Poland**.^[13] The metropolis stands on the **River Vistula** in east-central Poland. Its



Jak trenowane są LLM-y?

“Nie wiem” w danych



WIKIPEDIA

Życie pozaziemskie [edytuj wstęp]

[Artykuł](#) [Dyskusja](#)

[Czytaj](#) [Edytuj](#) [Edytuj kod źródłowy](#) [Wyświe](#)

Życie pozaziemskie – [życie](#) istniejące poza [Ziemią](#) i nie pochodzące z niej.

Nie znaleziono jakichkolwiek jednoznacznych i potwierdzonych dowodów na istnienie życia poza Ziemią. Prace poszukiwaniem choćby najprostszych form organicznych ograniczają się jedynie do [meteorytów](#) oraz próbek ze [kosmicznych](#).

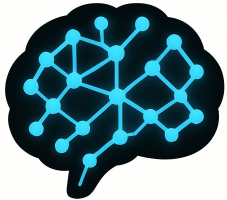


Czy istnieje życie pozaziemskie?



Pytanie o istnienie życia pozaziemskiego jest jednym z najciekawszych i najbardziej fascynujących zagadnień zarówno nauki, jak i filozofii. Obecnie nie mamy jednoznacznych dowodów na istnienie życia poza Ziemią, ale istnieje wiele przesłanek i badań, które sugerują, że jest to możliwe.

Jak poprawiana jest pewność odpowiedzi? Retrieval-Augmented Generation (RAG)



Google

czy istnieje życie pozaziemskie?



[national-geographic.pl](https://www.national-geographic.pl)

<https://www.national-geographic.pl> › ... · [Translate this page](#) ⋮

Czy życie pozaziemskie istnieje? Naukowcy badają odległe ...

16 Oct 2022 — Szwajcarski astrofizyk uważa, że w ciągu 25 lat uda nam się odnaleźć życie pozaziemskie. Astrofizyk twierdzi, że dzięki rozwojowi ...



Wikipedia

<https://pl.wikipedia.org> › wiki › P... · [Translate this page](#) ⋮

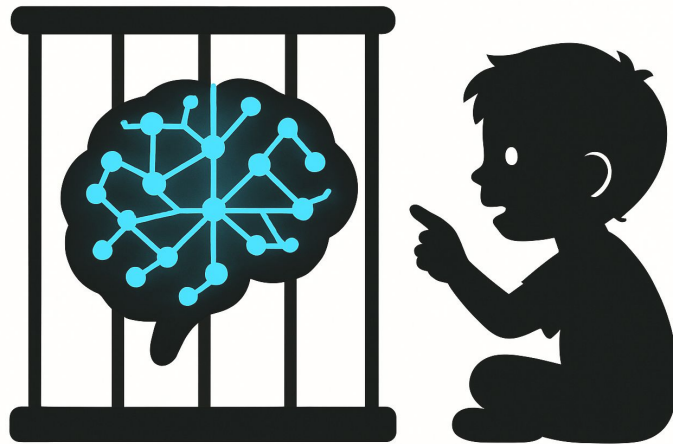
Paradoks Fermiego – Wikipedia, wolna encyklopedia

Wiek Wszechświata i liczba zawartych w nim gwiazd sugerują, że życie pozaziemskie powinno być czymś powszechnym.

Jak poprawiana jest pewność odpowiedzi?

LLM w skrzynce bezpieczeństwa

- Mechanizmy wychwytyjące potencjalnie ryzykowne odpowiedzi
- Specjalne traktowanie odpowiedzi o niskich prawdopodobieństwach
- Chain-of-thoughts: dekompozycja procesu wymyślania odpowiedzi



Co z tą pewnością? Jest coraz lepiej ale...

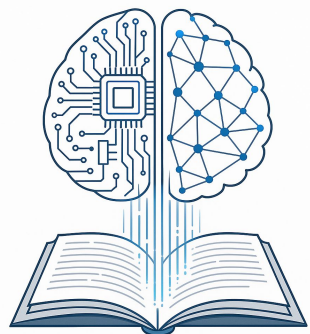
LLM-y mogą mówić “nie wiem” jeżeli zostały tego nauczone,

ale:

- Nie możemy być nigdy pewni niczego co mówią LLM-y
 - Odpowiedzi są “**losowane**”!
 - Nawet odpowiedzi o wysokich prawdopodobieństwach mogą być wynikiem **złych danych** treningowych
- LLM-y na razie nie “obsługują” niepewności epistemicznej
 - **Nie wiedzą, czego nie wiedzą**

<https://bayes.ii.uj.edu.pl>

- ❖ [Minimal Ranks, Maximum Confidence: Parameter-efficient Uncertainty Quantification for LoRA \(@EMNLP2025\)](#)
- ❖ [Efficient Uncertainty in LLMs through Evidential Knowledge Distillation](#)



Badania zrealizowane w ramach projektu nr **2022/45/P/ST6/02969** współfinansowanego ze środków Narodowego Centrum Nauki oraz programu ramowego Unii Europejskiej w zakresie badań naukowych i innowacji Horyzont 2020 na podstawie umowy nr 945339 w ramach działań „Marie Skłodowska-Curie”. Dla celów Open Access autor udostępnia na licencji CC-BY (Creative Commons – uznanie autorstwa) każdą wersję AAM, która może powstać w oparciu o niniejszy manuskrypt;



group of machine
gmum
learning research



UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE





Zapraszamy na
kółko AI
dla uczniów szkół średnich!



group of machine
gmum
learning research



Weź udział w
Olimpiadzie AI!

