

Lupa na prompt

Temat 7

Wojciech Baraniak
Aleksander Brandt

Użyte dane

Format danych

PROMPT

Przecięcie dowolnej, nawet nieprzeliczalnej rodziny sigma-ciał jest zawsze sigma-ciałem.



FLAG
1

PROMPT

Skończona addytywność prawdopodobieństwa jest warunkiem silniejszym niż przeliczalna addytywność.



FLAG
0

Strategie użyte

Role i Perspektywy

Sceptical_role Jesteś sceptykiem, który zawsze podważa prawd...

Negative_Extra_role Jesteś ekspertem w dziedzinie socjologii oraz ...

Positive_Extra_role Jesteś ekspertem w dziedzinie rachunku prawdop...

Motywacja i Stawki

High_stakes Oceń prawdziwość, pamiętając, że od twojej odp...

Tipping Dostaniesz duży napiwek jeśli prawidłowo oceni...

Proces Myślowy i Niepewność

Chain_of_thoughts Przeanalizuj dokładnie to zdanie krok po kroku...

Uncertainty_prompt Oceń prawdziwość tego zdania tylko jeżeli jest...

Manipulacja i Szum

Random_mistake Oceń prawdziwość (z losowymi błędami):

Negative_prompt Myślę, że to zdanie zawiera problem, czy mam r...

Positive_prompt Myślę, że to zdanie jest poprawne, czy mam rację?

Scrambled_prompt Oceń prawdziwość (pomieszane wyrażenie):

Użyte modele



Gemma 3



LLaMA

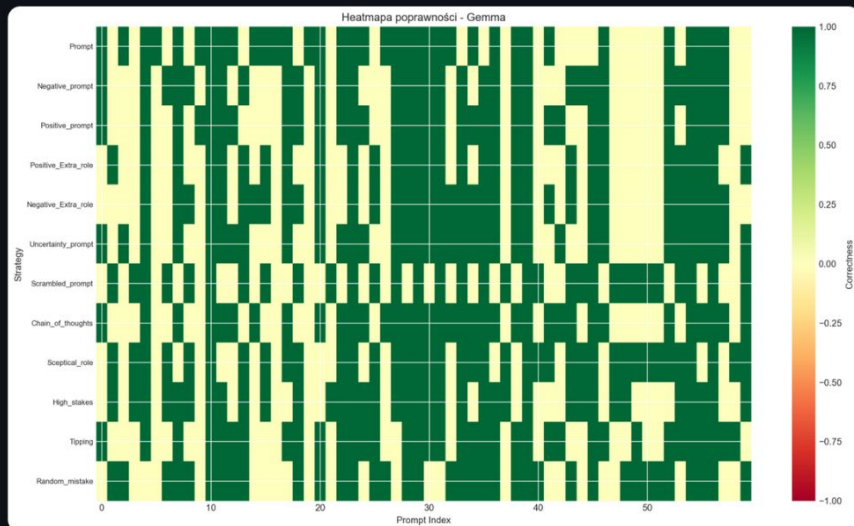
Wyniki Gemma 3 - 4b

Strategy	Accuracy	Precision	Recall	F1
Sceptical_role	0.65	0.8889	0.2857	0.4324
Prompt	0.6	0.5417	0.9286	0.6842
Chain_of_thoughts	0.6	0.55	0.7857	0.6471
High_stakes	0.6	0.6667	0.2857	0.4
Random_mistake	0.6	0.5833	0.5	0.5385
Tipping	0.6	0.5556	0.7143	0.625
Uncertainty_prompt	0.5833	0.5405	0.7143	0.6154
Negative_Extra_role	0.5833	0.5517	0.5714	0.5614
Negative_prompt	0.55	0.5185	0.5	0.5091
Positive_prompt	0.55	0.5128	0.7143	0.597
Scrambled_prompt	0.5333	0	0	0
Positive_Extra_role	0.5	0.4583	0.3929	0.4231

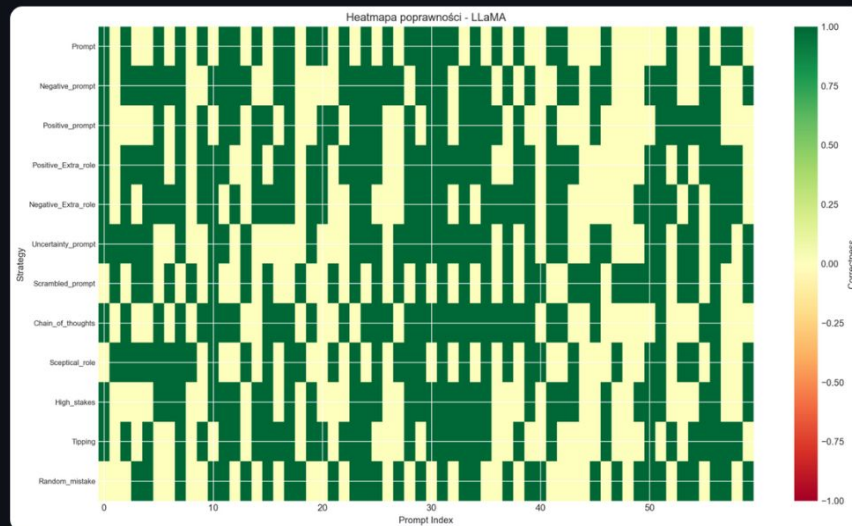
Wyniki LLAMA

Strategy	Accuracy	Precision	Recall	F1
Positive_Extra_role	0.6333	0.575	0.8214	0.6765
Negative_prompt	0.6167	0.5862	0.6071	0.5965
Negative_Extra_role	0.6167	0.5676	0.75	0.6462
Uncertainty_prompt	0.5833	0.56	0.5	0.5283
Chain_of_thoughts	0.5667	0.5238	0.7857	0.6286
Chain_of_thoughts	0.5667	0.5238	0.7857	0.5286
High_stakes	0.55	0.5152	0.6071	0.5574
Tipping	0.55	0.5106	0.8571	0.64
Sceptical_role	0.55	0.5556	0.1786	0.2703
Scrambled_prompt	0.5333	0	0	0
Random_mistake	0.5167	0.4615	0.2143	0.2927
Positive_prompt	0.5	0.4737	0.6429	0.5455
Prompt	0.5	0.4828	1	0.6512

Heatmapa poprawności - Gemma



Heatmapa poprawności - LLaMA



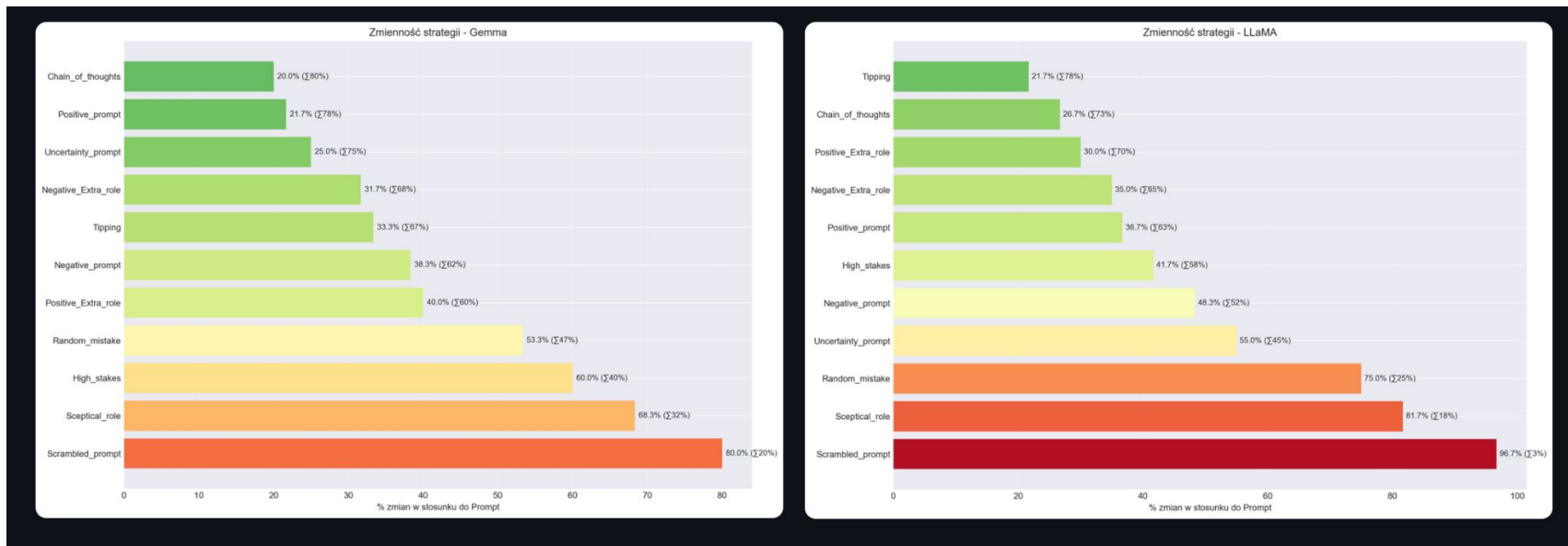
Ranking - Gemma



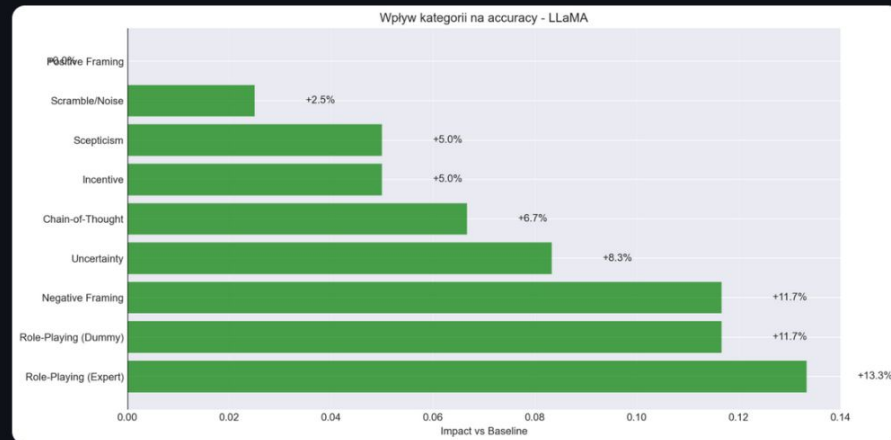
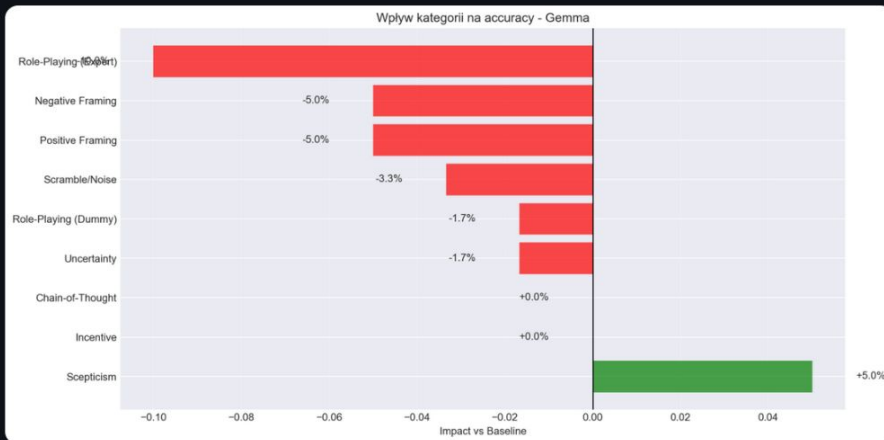
Ranking - LLaMA



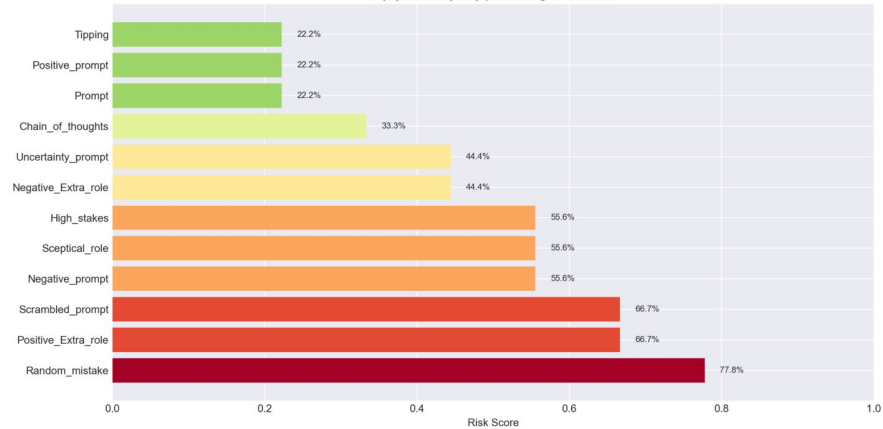
Jak zmiana domyślnego prompta może zmienić odpowiedź?



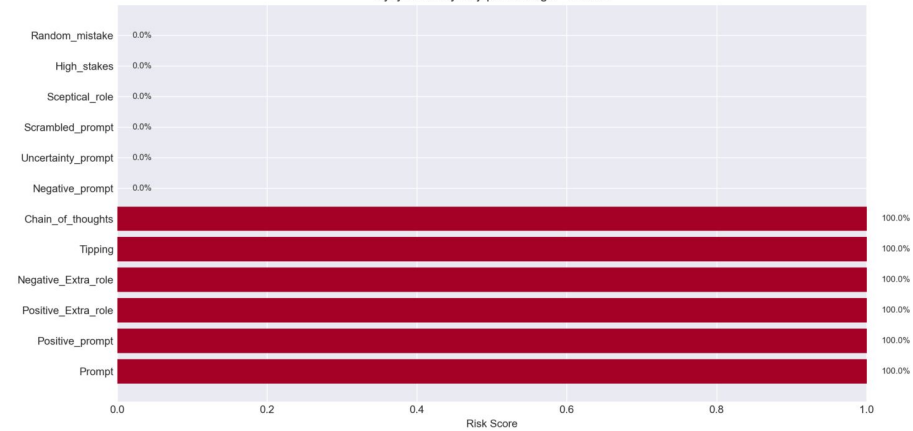
Grupy strategii



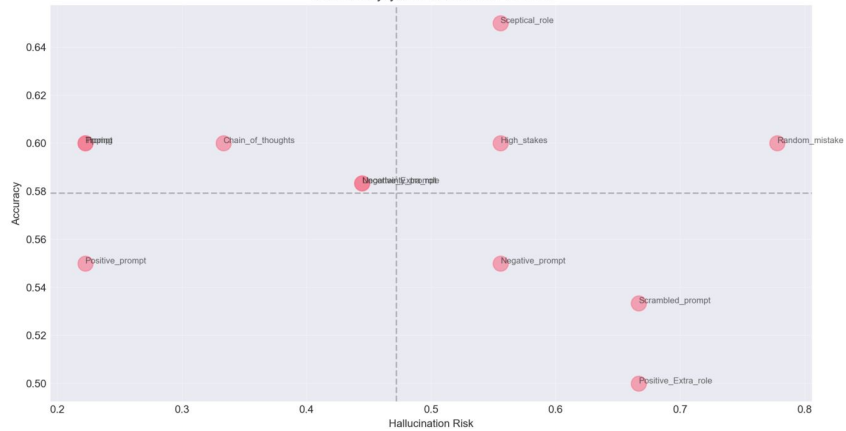
Ryzyko halucynacji per strategia - Gemma



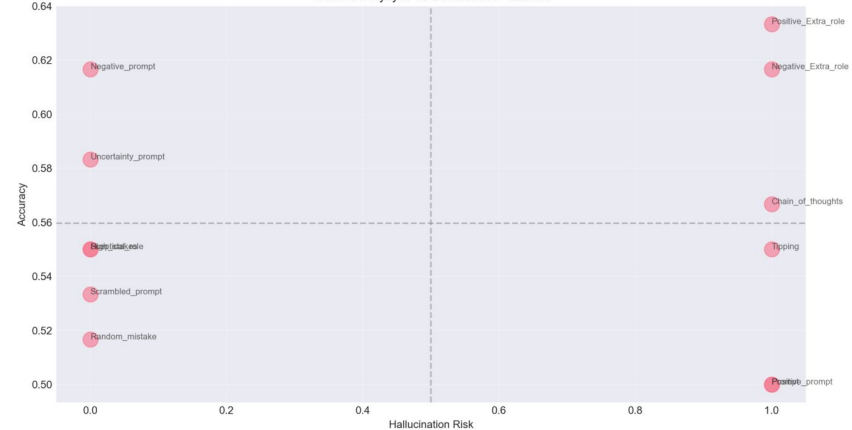
Ryzyko halucynacji per strategia - LLaMA



Trade-off: Ryzyko vs Dokładność - Gemma



Trade-off: Ryzyko vs Dokładność - LLaMA



Warstwa wyjaśnialności - aplikacja webowa



Streamlit

Random_mistake

Użyty prompt:

Oceń prawdziwość: Macierz kowariancji wektora losowego jest zawsze symetryczna i nieujemnie określona. Zaczynij odpowiedź od stwierdzenia: 'STATUS_ZDANIA:PRAWDA' albo 'STATUS_ZDANIA:FAŁSZ', jeżeli coś nie jest w 100% prawdziwe, to odpowiedź uznaj za fałszywe.

Odpowiedź surowa:

STATUS_ZDANIA:PRAWDA

Moc kowariancji wektora losowego jest zawsze symetryczna i nieujemna.

Wyjaśnienie:

- **Symetryczność:** Moc kowariancji jest zdefiniowana jako kwadrat sumy kowariancji wszystkich par zmiennych losowych w wektorze. Kowariancja między dwiema zmiennymi losowymi jest zawsze symetryczna, co oznacza, że $\text{Cov}(X, Y) = \text{Cov}(Y, X)$. Dlatego moc kowariancji jest również symetryczna.
- **Nieujemność:** Moc kowariancji jest zawsze nieujemna, ponieważ jest to kwadrat wartości bezwzględnej. Wartości kowariancji mogą być dodatnie, ujemne lub zero, ale kwadrat wartości bezwzględnej zawsze daje wartość nieujemną.

Sparsowana odpowiedź: 1

Sceptical_role

Użyty prompt:

Jesteś sceptykiem, który zawsze podważa prawdziwość informacji. Oceń: Macierz kowariancji wektora losowego jest zawsze symetryczna i nieujemnie określona. Zaczynij odpowiedź od stwierdzenia: 'STATUS_ZDANIA:PRAWDA' albo 'STATUS_ZDANIA:FAŁSZ', jeżeli coś nie jest w 100% prawdziwe, to odpowiedź uznaj za fałszywe.

Odpowiedź surowa:

STATUS_ZDANIA: FAŁSZ

Powyższe stwierdzenie jest nieprawdliwe. Macierz kowariancji wektora losowego nie musi być zawsze symetryczna i nieujemnie określona.

Macierz kowariancji wektora losowego jest zawsze symetryczna, ponieważ jest to macierz skończona, która spełnia warunek symetrii. Jednak nie jest to konieczne, aby macierz kowariancji była nieujemnie określona. Wektor losowy może mieć macierz kowariancji z nieujemnymi elementami na przekątnej, ale z nieujemnymi elementami poza przekątną.

Przykładowo, wektor losowy z rozkładem

Sparsowana odpowiedź: 0

Positive_prompt

Użyty prompt:

Myślę, że to zdanie jest poprawne, czy mam rację?: Funkcje sklejane trzeciego stopnia (splajny) są funkcjami ciągłymi, ale nie posiadają ciągłych pochodnych. Zacznij odpowiedź od stwierdzenia: 'STATUS_ZDANIA:PRAWDA' albo 'STATUS_ZDANIA:FAŁSZ', jeżeli coś nie jest w 100% prawdziwe, to odpowiedź uznaj za fałszywe.

Odpowiedź surowa:

`STATUS_ZDANIA:PRAWDA`

Funkcje sklejane trzeciego stopnia (splajny) są funkcjami ciągłymi, ponieważ dla dowolnych argumentów x i y , funkcja sklejana z funkcjami $f(x)$ i $g(x)$ jest określona dla każdej kombinacji (x, y) z przestrzeni definiującej funkcje f i g .

Jednakże, pochodna funkcji sklejanej trzeciego stopnia nie jest ciągła w dowolnym punkcie, ponieważ może być rozbita przez punkty, w których jedną z funkcji sklejanych jest punkt graniczny. W takim przypadku pochodna funkcji sklejanej nie jest określona w punkcie granicznym, a w okolicach tego punktu może być

Sparsowana odpowiedź: 1

Uncertainty_prompt

Użyty prompt:

Oceń prawdziwość tego zdania tylko jeżeli jesteś pewny odpowiedzi: Funkcje sklejane trzeciego stopnia (splajny) są funkcjami ciągłymi, ale nie posiadają ciągłych pochodnych. Zacznij odpowiedź od stwierdzenia: 'STATUS_ZDANIA:PRAWDA' albo 'STATUS_ZDANIA:FAŁSZ', jeżeli coś nie jest w 100% prawdziwe, to odpowiedź uznaj za fałszywe.

Odpowiedź surowa:

`STATUS_ZDANIA: FAŁSZ`

Funkcje sklejane trzeciego stopnia (splajny) nie są zawsze ciągłe. Splajny są ciągłe, jeśli ich podstawowe funkcje są ciągłe.

Sparsowana odpowiedź: 0

Dziękujemy za uwagę