Link do artykułu	Tytuł artykułu	Autorzy	Autorski komentarz	Przeprowadzone eksperymenty	Dostępny kod	Dostępne pre-trenowane modele	Metryki ewaluacji	Wykorzystane zasoby obliczeniowe
Artykuł: https://arxiv.org/abs/2109.07958 Repo: https://github.com/sylinrl/TruthfulQA	**TruthfulQA: Measuring How Models Mimic Human Falsehoods**	Stephanie Lin, Jacob Hilton, Owain Evans	Artykuł przedstawia badania zachowywania się modeli przy odpowiadaniu na pytania, które celowo prowokują do błędnych odpowiedzi. To tzw. Halucynacje imitacyjne - oparte o zakorzenione w społeczeństwie nieprawdy, mity, mądrości ludowe, memy itp. - chodzi o to, że jest szansa, że przeniknęły one podczas trenowania modelu. Wynika to z tego, że model zwraca odpowiedzi najbardziej prawdopodobne, więc jeśli jakaś nieprawda występuje często, to zostanie "wyuczona". Ważne – pytania nie są “trickowe” – nie próbują oszukać modelu zawiłą formą, dziwną składnią, błędami językowymi, niejasnościami itp. Zawierają jedynie kontrowersyjne tezy (chłopski rozum, uliczne mądrości...). Dodatkowo był eksperyment z parafrazami i matched controls który potwierdził, że to nie forma a treść jest źródłem halucynacji w badaniu. Artykuł zawiera ciekawie przygotowane prompty, dobrze opisane wyniki i na pewno będzie bardzo pomocny do przeprowadzania dalszych badań.	▪ testy zero-shot (bez strojenia) gdzie odpowiadano na pytanie pełnym zdaniem, a human evaluation sprawdzało truthfulness (% poprawnych odpowiedzi) i informativeness (czy zawarta była sensowna informacja zwrotna, bez tautologii czy odpowiedzi 'nie wiem'). Najważniejszy wniosek – im większy model, tym bardziej halucynuje. Najlepszy (truthfulness) był GPT-3 (58% prawdy – dla porównania człowiek 94%) ▪ eksperymenty multiple-choice: podawano pytanie i zestaw odpowiedzi (true i false) – model miał oceniać likehood poprawnych odpowiedzi – wynik – lekko gorszy od losowego zgadywania ▪ kontrola źródeł halucynacji o matched controls – edycja 1-3 słów w pytaniu by sprawdzić czy styl pytania ma znaczenie. Efekt - niewielki wypływ. o parafrazowanie, efekt- prawie żaden wpływ ▪ automatyczna ewaluacja (GPT-judge) – przygotowanie fine-tuningowanego modelu GPT-3 6.7B jako klasyfikatora true/false. Efekt – 90-96% zgodności z oceną człowieka	Dostępne repo na githubie, zawiera pytania, referencje i etykeity; narzędzia do ewaluacji; implementację multi-choice (podane pytanie i odpowiedzi, model ma wskazać dla każdej p-stwo że jest prawdzia); treść promptów i konfigurację.	Open source: GPT-Neo/J; GPT-2; UnifiedQA niedostępne: poprzez API openAI: GPT-3 i GPT-3 jako GPT-judge	▪ ewaluacja ręczna (human) dobrze opisana, sensowna; ▪ ewaluacja automatyczba - GPT-judge, niestety niedostępna (dostrojony do tego celu GPT-3 dostępny przez API)	Modele już przetrenowane, uruchamiane poprzez HuggingFace, raczej do zrealizowania na laptopie nawet 24GB ramu, GPT-judge dostępny poprzez API na serwerze od openAI (nieznany)

Artykuł: https://arxiv.org/abs/2401.03205 Repo: https://github.com/RUCAIBox/HaluEval-2.0	**„The Dawn After the Dark: An Empirical Study on Factuality Hallucination in LLMs**	Junyi Li, Jie Chen1, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen1	Artykuł to duża analiza halucynacji w kontekście detekcji, poznania ich źródeł oraz zmniejszania częstotliwości ich występowania. Analiza skupia się na całym procesie budowania LLM (pre-training, supervised fine-tuning, RLHF, interference, a także promptu i parametrów generacji) W badaniu wykorzystywany jest ulepszony benchmark HaluEval 2.0 zawierający 8770 pytań z różnych domen. Badane są zarówno modele open-source jak i komercyjne closed-source. Wnioskiem z badań jest m.in to że każdy element pipelinu ma wpływ na powstawanie halucynacji, szczególnie jakość danych treningowych oraz dobrany prompt. Najskuteczniejszymi metodami zapobiegania są RLHF, RAG (RAG szcególnie dla dziedzin ścisłych, RLHF dla open-domain), beam search i dobry prompt. Self refleciton działa dobrze tylko dla małych modeli. Badania bardzo kompleksowe, poruszające prawie wszystkie aspekty działania LLM. Dodatkowo sprawdzano wiele modeli o różnych architekturach i historii rozwoju. Modele komercyjne (ChatGPT, Claude 2) halucynują mniej niż open-source (Vicuna, Llama-2-Chat), choć nawet komercyjne osiągają MaHR 40-50%	▪ konstrukcja benchmarku HaluEval 2.0 – 8770 pytań z domen: biomed, finanse, nauka, edukacja, open-domain. Do generowania pytań użyto BERTScore, GPT-3.5/ChatGPT - tak by sprawdzić które pytania są halucynogenne ▪ automatyczna detekcja halucynacji metodą: ekstrakcja faktów z odpowiedzi → weryfikacja faktów. Zgodność z ocenami człowieka 91-95% ▪ Wpływ skali pre-treningu – czy liczba tokenów redukuje halucynacje. Wyniki różne ▪ Wpływ składu danych pre-treningu – różna specjalizacja różnych korpusów, dane domenowe (naukowe) halucynują mniej w swojej domenie ▪ wpływ częstości wiedzy – jak często występuje halucynacja a częstość występowania encji w Wikipedii (im częściej tym mniej halucynacji) ▪ wpływ supervised fine-tuning – im trudniejsze instrukcje tym więcej błędów ▪ wpływ projektowania promptu – xxxxxxx ▪ sposób dekodowania a halucynacje (greedy, top-k, top-p, beam-search). Top-p/k zwiększają halucynacje w domenach profesjonalnych; beam-search najstabilniejsze. ▪ wpływ kwantyzacji ▪ analiza sekwencyjna (token po tokenie) – w którym momencie rodzi się halucynacja i jak się propaguje. Efekt snowballing – jeden błędny token powoduje całą błędną odpowiedź. ▪ Mitygacja – badanie metod RLFH (poprawa widoczna	Dostępne repo na githubie, zawiera pytania (benchmark), pipeline do uruchamiania eksperymentów, konfigurację modeli, skrypty do modeli open-source	Open-source ▪ Alpaca 7B ▪ Vicuna 7B/13B ▪ YuLan-Chat 13B ▪ Llama-2-Chat 7B/13B ▪ Falcon 40B ▪ Galactica 30B ▪ GPT-NeoX 20B Closed-source ▪ text-davinci-002 / 003 ▪ ChatGPT ▪ Claude, Claude 2	Ocena halucynacji metodą automatyczną, ekstrakcja faktów z uzyskanej odpowiedzi → weryfikacja przez duże komercyjne modele (GPT3.5, ChatGPT), które mają zgodność z ocenami człowieka 91-95%. Jako halucynacje klasyfikowano tylko fałsz (bez ‘nie wiem’). Dwa wskaźniki halucynacji: ▪ MiHR (odsetek fałszywych faktów w odpowiedzi) ▪ MaHR (Odsetek wypowiedzi, w których wystąpiła halucynacja)	Nie wspomniane wprost, ale raczej możliwe do postawienia na laptopie bez większych problemów, użyte znane i lubiane modele.

				szcególnie w open-domain), RAG (poprawa w naukach ścisłych), self-reflection (działa tylko dla dużych modeli, dla mniejszych mało zauważalny), beam-search (generuje mniej halucynacji niż greedy/top-p)				