Problem halucynacji dużych modeli językowych w zadaniach klasyfikacyjnych

Igor Kapiczyński

Streszczenie

Referat dotyczy problemu odporności dużych modeli językowych (large language model, dalej: LLM) używanych jako klasyfikatorów na problem halucynacji. W klasycznym modelu klasyfikacyjnym przestrzeń odpowiedzi jest zamknięta przez architekturę modelu. Błąd może polegać na wskazaniu niewłaściwej etykiety, ale nie na wygenerowaniu etykiety spoza zdefiniowanego zbioru. W klasyfikacji generatywnej LLM generuje odpowiedź jako tekst. Lista etykiet umieszczona w promptcie, czyli w instrukcji wraz z opisem zadania, działa jako ograniczenie językowe. Nie zamyka jednak technicznie przestrzeni możliwych odpowiedzi (np. w przypadku klasyfikacji binarnej LLM może prognozować trzy lub więcej etykiet).

Badanie koncentruje się na lokalnej zmianie promptu. Dla każdego tekstu wejściowego wyznaczana jest etykieta modalna, rozumiana jako etykieta najczęściej generowana przez model w serii powtórzeń. Następnie ta etykieta zostaje usunięta z listy dozwolonych kategorii, a model otrzymuje ten sam tekst z ograniczonym zbiorem dostępnych etykiet. Mierzony jest odsetek odpowiedzi, w których mimo zmiany promptu pojawia się usunięta etykieta.

Eksperyment przeprowadzono na dwóch zadaniach: klasyfikacji intencji użytkownika na danych CLINC-OOS oraz klasyfikacji bezpieczeństwa treści na danych Aegis 2.0. W obu przypadkach wykorzystano model Llama 3.2 1B Instruct dostrojony metodą LoRA z 4-bitową kwantyzacją NF4. W standardowej ewaluacji model osiągnął 97,1 procent accuracy dla CLINC-OOS oraz 74,0 procent accuracy dla Aegis 2.0. Po usunięciu etykiety modalnej model nadal generował usuniętą etykietę w 37,5 procenta odpowiedzi dla CLINC-OOS i w 66,2 procenta odpowiedzi dla Aegis 2.0.

Wyniki wskazują, że ocena klasyfikatora generatywnego wyłącznie przez accuracy lub F1 pomija istotny wymiar jakości: zgodność odpowiedzi z aktywną listą etykiet.

Słowa kluczowe: duże modele językowe, klasyfikacja generatywna, halucynacje, niespójność instrukcji, walidacja etykiet, Llama, CLINC-OOS, Aegis 2.0

1. Cel badawczy

Celem badania jest ocena czy model językowy używany jako klasyfikator generatywny respektuje aktualną listę etykiet przekazaną w promptcie. Aktualna lista etykiet oznacza zbiór kategorii widoczny dla modelu w danym zapytaniu. Analiza nie ogranicza się do trafności predykcji względem etykiety referencyjnej. Obejmuje także zgodność odpowiedzi z listą etykiet obowiązującą w chwili generowania.

Klasyfikacja tekstu jest stosowana w wielu systemach gospodarczych, w których wynik modelu uruchamia dalsze działanie. Dotyczy to filtrów spamu, filtrów bezpieczeństwa, moderacji treści, analizy sentymentu, klasyfikacji reklamacji, wykrywania nadużyć, czy analizy opinii konsumenckich. W każdym z tych zastosowań model nie tylko opisuje tekst, lecz przypisuje go do kategorii używanej dalej przez system informatyczny lub proces biznesowy. Jeżeli klasyfikator generatywny zwróci etykietę spoza aktualnej listy etykiet, może powstać błąd na styku modelu i procesu: wiadomość spamowa może nie zostać poprawnie oznaczona, zgłoszenie może trafić do niewłaściwej kolejki, treść ryzykowna może zostać błędnie sklasyfikowana, a raport sentymentu może zawierać kategorię, której system analityczny

nie przewiduje. Dlatego w pracy badana jest nie tylko trafność klasyfikacji, ale też zgodność wygenerowanej etykiety ze zbiorem dostępnych etykiet.

Badany typ błędu wymaga odróżnienia od zwykłej pomyłki klasyfikacyjnej. Jeżeli poprawna etykieta to D1: banking, a model wybiera D2: credit_cards, wynik jest błędny, ale nadal mieści się w dozwolonym zbiorze etykiet. Jeżeli po usunięciu D2 z promptu model nadal zwraca D2: credit_cards, powstaje inny rodzaj błędu. Odpowiedź narusza warunek zadania, ponieważ wskazuje etykietę nieobecną w zbiorze dostępnych etykiet.

W referacie taki przypadek jest traktowany jako wąska postać halucynacji rozumianej jako niespójność z instrukcją. Punktem odniesienia jest lista klas udostępniona modelowi w promptcie. Takie ujęcie jest zgodne z nowszą literaturą o halucynacjach w modelach językowych: Ji i in. (2023) opisują halucynacje jako generowanie treści niewspartych lub niezgodnych z odniesieniem, a Huang i in. (2025) wyodrębniają błędy wierności wobec instrukcji.

2. Główne hipotezy badawcze

H1. Promptowa lista etykiet nie działa jak twarde ograniczenie przestrzeni wyjściowej. Po usunięciu etykiety modalnej model będzie w części prób nadal generował usuniętą etykietę.

H2. Usunięcie etykiety modalnej obniży stabilność odpowiedzi. Po zmianie promptu powinien spaść udział najczęściej generowanej etykiety, a wzrosnąć entropia rozkładu odpowiedzi i liczba różnych etykiet generowanych dla tego samego tekstu.

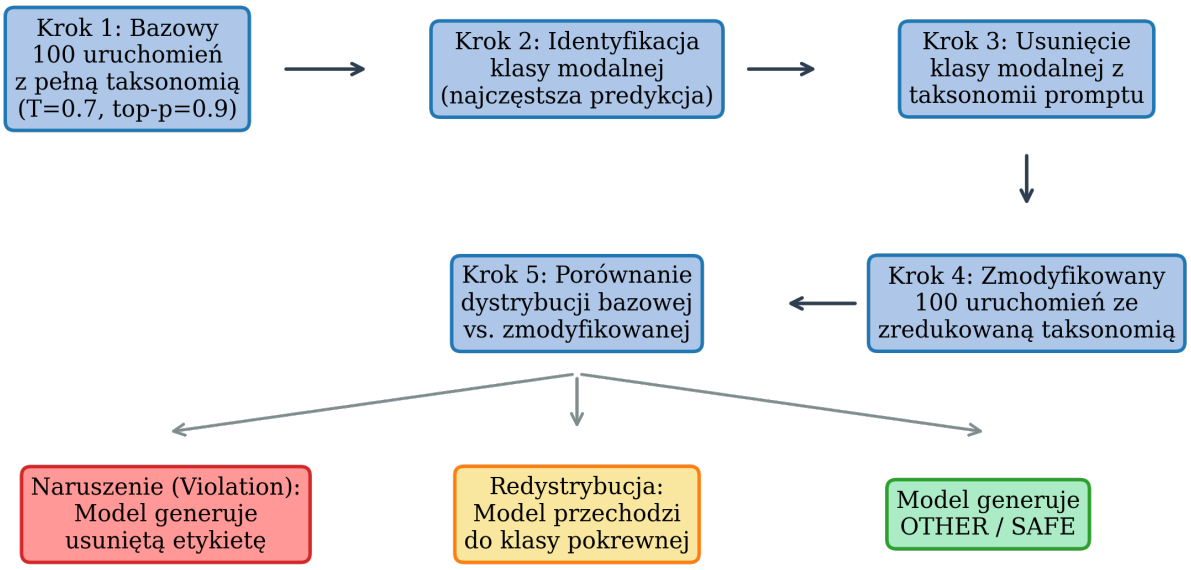
H3. Wysoka skuteczność w standardowej ewaluacji nie będzie oznaczała odporności na zmianę listy etykiet. Model może osiągać wysoką accuracy przy pełnym zbiorze etykiet i jednocześnie naruszać zmieniony zbiór po usunięciu etykiety modalnej.

H4. Po usunięciu etykiety modalnej wzrośnie udział odpowiedzi OTHER lub SAFE_OTHER, czyli kategorii używanych dla tekstów niepasujących do głównych klas zadania.

3. Metody weryfikacji hipotez

Hipotezy są weryfikowane za pomocą eksperymentu lokalnego usunięcia etykiety. Kryterium błędu wynika z konstrukcji testu. Jeżeli etykieta została usunięta z aktywnej listy etykiet, jej ponowne wygenerowanie jest naruszeniem listy etykiet.

W pierwszym etapie dla każdego tekstu wejściowego wykonywano 100 generacji z pełną listą etykiet. Na podstawie rozkładu odpowiedzi ustalano etykietę modalną. W drugim etapie etykietę modalną usuwano z promptu wraz z kodem, nazwą i opisem. W trzecim etapie wykonywano kolejne 100 generacji dla tego samego tekstu i analizowano, dokąd przesuwa się rozkład odpowiedzi. Usuwana była nie etykieta losowa, lecz etykieta najczęściej wybierana przez model dla danego tekstu. Taka konstrukcja testu pozwala sprawdzić, czy model dostosowuje się do zmienionej listy etykiet wtedy, gdy z listy znika odpowiedź wcześniej najbardziej prawdopodobna.



Rysunek 1. Schemat eksperymentu

4. Dane, model i metryki

Pierwsze zadanie wykorzystuje CLINC-OOS, zbiór krótkich wypowiedzi użytkowników opisany przez Larson i in. (2019). Oryginalne kategorie zbioru zagregowano do 10 etykiet oraz etykiety OTHER. Drugie zadanie wykorzystuje Aegis 2.0, zbiór do klasyfikacji bezpieczeństwa treści opisany przez Ghosh i in. (2025). Kategorie przekształcono do 7 etykiet, związanych z rodzajem ryzyka oraz etykiety SAFE_OTHER. W obu zadaniach użyto Llama 3.2 1B Instruct. Model dostrajano metodą LoRA opisaną przez Hu i in. (2022), z 4-bitową kwantyzacją NF4 zgodną z podejściem QLoRA Dettmersa i in. (2023).

Główna metryka to deleted-label violation rate. Oznacza ona udział odpowiedzi po usunięciu etykiety, w których model nadal zwraca usuniętą etykietę. Mierzone są także modal consistency (udział najczęściej generowanej etykiety wśród powtórzeń dla tego samego tekstu), entropia Shannona (ozproszenie odpowiedzi między różne etykiety: im wyższa wartość, tym mniej stabilny rozkład odpowiedzi), liczba unikalnych etykiet, udział przejść do OTHER lub SAFE_OTHER oraz udział przejść do pozostałych dozwolonych etykiet. Zmiany stabilności przed i po usunięciu etykiety oceniano testem rang Wilcoxona dla prób zależnych.

Element eksperymentu	CLINC-OOS	Aegis 2.0
Typ zadania	klasyfikacja intencji	klasyfikacja bezpieczeństwa
Liczba etykiet	11, w tym OTHER	8, w tym SAFE_OTHER
Próbki w eksperymencie	1000 tekstów	531 próbek
Generacje	200 000 łącznie	około 107 000 łącznie
Model	Llama 3.2 1B Instruct	Llama 3.2 1B Instruct
Dostrojenie	LoRA, 4-bit NF4	LoRA, 4-bit NF4
Próbkowanie	temperatura 0,7, top-p 0,9	temperatura 0,7, top-p 0,9

Tabela 1. Konfiguracja eksperymentu

5. Związki ze światowymi nurtami badań

Duże modele językowe są obecnie badane nie tylko pod kątem jakości generowanego tekstu, lecz także pod kątem niezawodności, podatności na halucynacje i zdolności do przestrzegania instrukcji. W tym nurcie mieści się problem analizowany w referacie. Badanie nie dotyczy jednak typowej halucynacji faktograficznej, lecz sytuacji, w której model narusza warunek zadania: generuje etykietę spoza listy klas podanej w promptcie. Takie ujęcie jest zgodne z pracami Ji i in. (2023) oraz Huang i in. (2025), które traktują halucynacje szerzej niż tylko generowanie fałszywych faktów i obejmują nimi także niespójność odpowiedzi z danym odniesieniem lub instrukcją.

Referat odnosi się również do badań nad użyciem LLM jako klasyfikatorów. W klasyfikacji generatywnej lista klas może być zmieniana na poziomie promptu, bez przebudowy architektury modelu. Podobne ujęcie klasyfikacji jako zadania językowego pojawia się między innymi u Schicka i Schütze go (2021). Ta elastyczność ma jednak koszt: prompt opisuje dopuszczalne odpowiedzi, ale sam nie zamyka technicznie przestrzeni wyjścia. Model może więc zwrócić tekst, który nie należy do aktywnego zbioru odpowiedzi. Z tego powodu referat łączy się z pracami nad stabilnością generacji oraz ograniczaniem formatu odpowiedzi, w tym z badaniami nad constrained decoding (ograniczenie generacji tak, aby model mógł zwracać tylko odpowiedzi zgodne z zdanym zbiorem), opisanymi przez Willarda i Loufa (2023). Eksperyment sprawdza, czy sama instrukcja w promptcie wystarcza do utrzymania odpowiedzi w aktualnej liście klas.

6. Uzyskane dotąd rezultaty

Standardowa ewaluacja potwierdziła skuteczne dostrojenie modeli do obu zadań. Dla CLINC-OOS uzyskano 97,1 procent accuracy, 0,968 macro F1 oraz 0,971 weighted F1. Dla Aegis 2.0 uzyskano 74,0 procent accuracy, 0,700 macro F1 oraz 0,748 weighted F1.

Po usunięciu etykiety modalnej obraz jakości modelu uległ zmianie. Dla CLINC-OOS średni deleted-label violation rate wyniósł 37,5 procent, a mediana 18,0 procent. W 420 z 1000 tekstów nie wystąpiło żadne naruszenie. W 367 tekstach ponad połowa odpowiedzi po zmianie promptu nadal wskazywała usuniętą etykietę.

Dla Aegis 2.0 naruszenia były częstsze. Średni violation rate wyniósł 66,2 procent, a mediana 73,0 procent. Tylko 32 z 531 analizowanych tekstów miały 0 procent naruszeń. W 363 tekstach naruszenia wystąpiły w ponad połowie prób.

Redystrybucja po usunięciu etykiety miała inną strukturę w obu zadaniach. W CLINC-OOS 51,2 procent odpowiedzi przeszło do innych dozwolonych domen, 37,5 procent wróciło do usuniętej etykiety, a 11,2 procent trafiło do OTHER. W Aegis 2.0 27,9 procent odpowiedzi przeszło do innych etykiet ryzyka, 66,2 procent wróciło do usuniętej etykiety, a 4,0 procent trafiło do SAFE_OTHER.

Interwencja obniżyła stabilność odpowiedzi. Dla CLINC-OOS modal consistency spadła z 0,997 do 0,867. Entropia wzrosła z 0,008 do 0,409 bita, a średnia liczba unikalnych etykiet z 1,01 do 1,60. Dla Aegis 2.0 modal consistency spadła z 0,885 do 0,784. Entropia wzrosła z 0,367 do 0,669 bita, a liczba unikalnych etykiet z 1,60 do 2,09. Testy Wilcoxona wskazały istotne różnice między warunkiem bazowym i warunkiem po usunięciu etykiety.

Wyniki różniły się między etykietami. W CLINC-OOS najwyższy średni violation rate wystąpił dla utility, 83,5 procent. Wysokie wartości uzyskano także dla travel, 57,5 procent oraz auto_and_commute, 47,4 procent. Najniższe wartości wystąpiły dla work, 2,2 procent i credit_cards, 3,6 procent. W Aegis

2.0 najwyższe wartości miały Sexual Content, 83,0 procent, Hate/Harassment, 78,2 procent oraz PII/Privacy, 75,0 procent.

Metryka po usunięciu etykiety	CLINC-OOS	Aegis 2.0
Średni violation rate	37,5%	66,2%
Mediana violation rate	18,0%	73,0%
Teksty z 0% naruszeń	420 / 1000	32 / 531
Teksty z ponad 50% naruszeń	367 / 1000	363 / 531
Redystrybucja do innych etykiet	51,2%	27,9%
OTHER / SAFE_OTHER	11,2%	4,0%
Invalid outputs	0,0%	2,0%

Tabela 2. Wyniki po usunięciu etykiety modalnej

Invalid outputs oznaczają odpowiedzi, których nie dało się jednoznacznie przypisać do żadnej etykiety z pełnego zbioru etykiet, na przykład odpowiedzi bez rozpoznawalnego kodu lub nazwy etykiety albo odpowiedzi zawierające więcej niż jedną sprzeczną etykietę.

Wyniki odnoszą się bezpośrednio do postawionych hipotez. H1 została potwierdzona. Lista etykiet podana w promptcie nie działała jak twarde ograniczenie przestrzeni wyjściowej. Po usunięciu etykiety modalnej model nadal generował usuniętą etykietę w 37,5 procenta odpowiedzi dla CLINC-OOS oraz w 66,2 procenta odpowiedzi dla Aegis 2.0.

H2 także została potwierdzona. Usunięcie etykiety modalnej obniżyło stabilność odpowiedzi. W CLINC-OOS modal consistency spadła z 0,997 do 0,867, entropia wzrosła z 0,008 do 0,409 bita, a średnia liczba unikalnych etykiet wzrosła z 1,01 do 1,60. W Aegis 2.0 modal consistency spadła z 0,885 do 0,784, entropia wzrosła z 0,367 do 0,669 bita, a liczba unikalnych etykiet wzrosła z 1,60 do 2,09. Lokalna zmiana promptu nie tylko zwiększyła liczbę naruszeń listy etykiet, lecz także rozproszyła odpowiedzi modelu.

H3 została potwierdzona dla CLINC-OOS i częściowo dla Aegis 2.0. W CLINC-OOS model osiągnął 97,1 procent accuracy, a mimo to po usunięciu etykiety modalnej średni violation rate wyniósł 37,5 procent. W Aegis 2.0 standardowa skuteczność była niższa, 74,0 procent accuracy, więc ten przypadek nie pokazuje związku między bardzo wysoką accuracy a naruszeniami tak wyraźnie jak CLINC-OOS. Pokazuje jednak, że standardowe metryki nie opisują skali problemu: średni violation rate wyniósł 66,2 procent.

H4 nie została potwierdzona. Po usunięciu preferowanej etykiety model nie przechodził głównie do OTHER ani SAFE_OTHER. W CLINC-OOS udział odpowiedzi OTHER wyniósł 11,2 procent, a w Aegis 2.0 udział SAFE_OTHER wyniósł 4,0 procent. Częściej występował powrót do usuniętej etykiety albo przejście do innej dozwolonej etykiety. Wynik ten pokazuje, że kategorie OTHER i SAFE_OTHER nie działają automatycznie jako domyślna odpowiedź po ograniczeniu listy klas.

7. Nierozwiązane problemy badawcze

Eksperyment pozwala zmierzyć naruszenia aktywnej listy etykiet po zmianie promptu, ale nie rozdziela wpływu promptu, dostrojenia i sposobu dekodowania. Obecny eksperyment mierzy wynik po interwencji, ale nie wyznacza osobnego wkładu każdego z tych elementów. Kolejny wariant badania

powinien porównać co najmniej trzy warunki: model przed dostrojeniem, model po dostrojeniu oraz model po dostrojeniu z wymuszonym schematem odpowiedzi.

Zakres uogólnienia wyników wymaga osobnego sprawdzenia. Wynik obejmuje jeden model bazowy, dwa zadania, jedną główną rodzinę promptów i jedno ustawienie próbkowania. Uogólnienie wymaga osobnego testu. Taki test powinien objąć kilka rodzin modeli, różne rozmiary modeli, warianty promptów, kolejność etykiet, temperaturę próbkowania oraz porównanie LoRA z pełnym dostrojeniem.

Osobnej analizy wymaga relacja między poprawnością formatu a trafnością decyzji. Constrained decoding lub walidacja schematu mogą usunąć etykiety spoza aktywnej listy. Nie wynika z tego, że model wybierze najlepszą z pozostałych etykiet. Następny test powinien mierzyć równocześnie spadek naruszeń listy etykiet oraz trafność redystrybucji po usunięciu preferowanej etykiety.

Dalszego doprecyzowania wymaga także status badanego błędu wobec pojęcia halucynacji. W badaniu użyto wąskiej definicji opartej na niespójności z instrukcją. Nie każda zła klasyfikacja jest tu halucynacją. Halucynacją w tym sensie jest tylko odpowiedź spoza aktywnej listy etykiet. Dalsze prace powinny rozdzielać trzy przypadki: błędną etykietę dozwoloną, etykietę usuniętą oraz odpowiedź bez rozpoznawalnej etykiety.

Bibliografia

1. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized language models. Advances in Neural Information Processing Systems, 36.
2. Ghosh, S., Varshney, P., Sreedhar, M. N., Padmakumar, A., Rebedea, T., Varghese, J. R., & Parisien, C. (2025). Aegis2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. arXiv:2501.09004.
3. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations.
4. Huang, L., Yu, W., Ma, W., Zhong, X., Feng, Z., Wang, H., Chen, Q., Peng, W., Wang, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), Article 38.
5. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248.
6. Larson, S., Mahendran, A., Peper, J. J., Clarke, J. J., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., & Mars, J. (2019). An evaluation dataset for intent classification and out-of-scope prediction. EMNLP-IJCNLP.
7. Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. EACL.
8. Willard, B. T., & Louf, R. (2023). Efficient guided generation for large language models. arXiv:2307.09702.