doi: 10.35757/KiS.2026.70.3.8

KULTURA i SPOŁECZEŃSTWO

POLSKA AKADEMIA NAUK
KOMITET SOCJOLOGII
INSTYTUT STUDIÓW POLITYCZNYCH
2026, nr 3

ISSN 0023-5172
e-ISSN 2300-195x

CC BY NC SA

ANNA KAMIŃSKA
Uniwersytet Śląski w Katowicach

GRAMATYKA, STYL I INKLUZJA W LOKALNYCH MODELACH JĘZYKOWYCH

OCENA SYSTEMÓW BIELIK I PLLUM W KONTEKŚCIE STOSOWANIA FEMINATYWÓW*

WSTĘP

Duże modele językowe (DMJ), rozwijane w ostatnich latach na niespotykaną dotychczas skalę, stały się jedną z najbardziej wpływowych technologii współczesnej sztucznej inteligencji (Örpek, Tural, Destan 2024). Dzięki zdolnościom do generowania, parafrazowania oraz analizowania

Adres do korespondencji: anna.kaminska@us.edu.pl, ORCID: 0000-0001-5411-5426

* Autorka jest pomysłodawczynią badania: opracowała zestaw 45 zadań testowych oraz taksonomię dwunastu klas tematyczno-problemowych, przeprowadziła wszystkie wywołania modeli i koordynowała proces oceny. Ewaluacji dokonał trzyosobowy zespół osób sędziujących, w którego skład — obok osób wymienionych w „Podziękowaniach" — wchodziła również autorka jako pracowniczka Biblioteki Politechniki Śląskiej. Każda z osób oceniała odpowiedzi niezależnie, a rozbieżności rozstrzygano w drodze dyskusji, aż do wypracowania wspólnego stanowiska. Zapytania kierowano do modeli od czerwca do lipca 2025; ponieważ Gemini 2.5 Flash jest usługą sieciową aktualizowaną po stronie dostawcy, przedstawione wyniki odnoszą się do stanu systemów z tego okresu. Wszystkie dane wykorzystane w analizie — w tym pełne treści zadań i odpowiedzi generowanych przez poszczególne DMJ — zostały udostępnione w otwartym repozytorium Zenodo w celu zapewnienia przejrzystości i możliwości replikacji wyników (zob. https://doi.org/10.5281/zenodo.19200283).

języków naturalnych znajdują zastosowanie w tak różnorodnych dziedzinach jak edukacja, opieka zdrowotna, administracja publiczna czy komunikacja biznesowa (Chkirbene i in. 2024). Osoby korzystające coraz częściej sięgają po ich możliwości w wyszukiwarkach, tłumaczach maszynowych, asystentach głosowych czy czatbotach (Kumar i in. 2024). Skalowalność, elastyczność i ciągle rosnące możliwości tych systemów — oraz ich zwiększająca się wszechobecność — sprawiają, że coraz częściej stają się nie tylko narzędziem wspomagającym, ale i aktywnym aktorem społecznym współkształtującym nawyki językowe oraz społeczne ramy komunikacji.

Właśnie z tego powodu pojawia się potrzeba refleksji nad tym, w jaki sposób DMJ wpływają na język — nie tylko jako medium komunikacji, ale również jako system kulturowy, społeczny i tożsamościowy. Modele językowe trenowane są na wielkich zbiorach danych, które stanowią zapis zastanego kapitału kulturowego; ich zawartość i reprezentatywność bezpośrednio wpływają na to, czy SI legitymizuje określone hierarchie społeczne, czy też proponuje nowe formy ekspresji (Dai i in. 2024). Gdy dane te zawierają przewagę treści zdominowanych przez konkretne języki, style lub systemy wartości, może dochodzić do utrwalania określonych wzorców językowych, w tym także kalek czy naleciałości z innych języków kulturowo dominujących.

Z perspektywy osób korzystających z DMJ w Polsce oraz języka polskiego istotne jest więc pytanie, czy, a jeśli tak, to w jakim stopniu, modele te trenowane z użyciem krajowych korpusów danych wysokiej jakości potrafią nie tylko lepiej odwzorowywać strukturę gramatyczną i składniową wypowiedzi, ale również adekwatnie odzwierciedlać konteksty społeczne i kulturowe oraz lokalne dynamiki zmian obyczajowych, charakterystyczne dla współczesnego polskiego dyskursu publicznego. Jednym z obszarów szczególnie wrażliwych w tym kontekście jest kwestia form żeńskich nazw zawodów i ról społecznych — tzw. feminatywów — których obecność i akceptowalność to tematy dziś żywo dyskutowane w przestrzeni publicznej (Latos 2020). DMJ, jako technologie nie tylko przetwarzające, ale i generujące wypowiedzi w danym języku, mogą stać się istotnym czynnikiem wpływającym na to, które formy językowe będą się utrwalać, a które zanikać — zwłaszcza gdy wypowiedzi te stają się dla wielu osób domyślnym punktem odniesienia.

Język polski, jako język morfologicznie złożony, o rozbudowanej fleksji i rosnącym zakresie form alternatywnych, takich jak feminatywy, stanowi dobry przypadek testowy do oceny kompetencji modeli w zakresie lokalnych uwarunkowań językowych. Z tych względów celem niniejszego badania jest empiryczna analiza tego, jak różne modele językowe — w tym zarówno modele trenowane na danych z polskiego kontekstu kulturowego, jak i te o charakterze bardziej ogólnym — radzą sobie z zadaniami językowymi w języku polskim, ze szczególnym uwzględnieniem stosowania feminatywów. Badanie to zostało zaprojektowane z myślą o ocenie jakości gramatycznej oraz spójności z dążeniami do inkluzywności, co pozwala na identyfikację uprzedzeń algorytmicznych będących echem dominujących w społeczeństwie stereotypów. Analizowane są zarówno lokalne modele otwarte, także te dostrajane na polskich korpusach, jak i wielkoskalowy model komercyjny w celu uchwycenia ewentualnych różnic wynikających z zakresu danych treningowych, przyjętej architektury oraz strategii ich trenowania.

Mimo rosnącej dostępności modeli dostrajanych z użyciem polskojęzycznych korpusów wysokiej jakości nadal niewiele wiadomo, jak dostrojenie to przekłada się na ich zachowanie językowe w praktyce. W szczególności brakuje analiz porównawczych, które zestawiałyby modele o zróżnicowanej charakterystyce — zarówno te o lokalnym profilu treningowym, jak i modele ogólne, trenowane na danych wielojęzycznych lub głównie anglojęzycznych, a także modele komercyjne o dużej skali. Tymczasem to właśnie takie zestawienie pozwala na rekonstrukcję odmiennych standardów komunikacyjnych oraz uchwycenie napięć między globalnymi a lokalnymi wzorcami językowymi obecnymi w polskiej sferze cyfrowej.

W niniejszym badaniu zaproponowano analizę porównawczą wybranych DMJ pod kątem ich zdolności generowania wypowiedzi w języku polskim, ze szczególnym uwzględnieniem ich poprawności gramatycznej, naturalności językowej oraz podejścia do form żeńskich w nazewnictwie zawodów i ról społecznych. Analiza ta opiera się na serii zapytań (poleceń/promptów) przypisanych do określonych klas zadaniowych, obejmujących między innymi generowanie tekstu, uzupełnianie luk, przekształcanie form gramatycznych i ocenę odpowiedniości wyrażeń.

Tak zarysowane podejście badawcze wymaga doprecyzowania teoretycznego — pojęcie kompetencji kulturowej operacjonalizowane jest tu

właśnie przez pryzmat jej lingwistycznych manifestacji. Nie aspiruje się tu zatem do opisu pełnej wiedzy encyklopedycznej modeli, lecz skupia na ich zdolności do reprodukcji wzorców językowych nierozerwalnie związanych z polskim kontekstem społecznym (np. systemu adresatywnego czy tworzenia form żeńskich). Na tej podstawie sformułowano następujące hipotezy:

Hipoteza ogólna: Trenowanie modeli językowych na korpusach obejmujących lokalne dane językowe (w tym przypadku dane w języku polskim) może podwyższać stopień ich lingwistycznej kompetencji kulturowej w zakresie naturalności i osadzenia w lokalnym kontekście komunikacyjnym, co przejawia się w generowaniu wypowiedzi o większej naturalności i lepszym osadzeniu w lokalnym kontekście komunikacyjnym.

Hipoteza szczegółowa: Modele językowe trenowane na polskich korpusach tekstowych mogą lepiej rozpoznawać i stosować formy językowe związane z lokalnymi kontekstami kulturowymi i społecznymi, czego wskaźnikiem jest sposób implementacji feminatywów, co przejawia się w większej akceptowalności oraz bardziej naturalnym użyciu feminatywów w odpowiedziach generowanych przez te modele.

TŁO BADAWCZE

Duże modele językowe, takie jak GPT, LLaMA, Mistral czy Claude, bazujące na architekturze transformerów (Vaswani i in. 2017), w ostatnich latach osiągnęły przełomową skuteczność w zadaniach językowych — od generowania tekstu i tłumaczenia po streszczanie, parafrazowanie czy prowadzenie dialogu. Ich zdolność do generowania odpowiedzi zbliżonych do ludzkich wynika z treningu na ogromnych zbiorach tekstów, które umożliwiają modelom nie tylko operowanie na strukturach językowych, ale także reprodukowanie zastanych w tekstach kodów kulturowych oraz pragmatyki komunikacyjnej właściwej dla danej wspólnoty językowej. Wraz z pojawieniem się ogólnodostępnych wersji tych modeli możliwe stało się ich trenowanie na własnych zbiorach danych, w tym w języku polskim, który ze względu na swoją złożoną morfologię i specyficzne uwarunkowania kulturowe stawia przed takimi zagadnieniami istotne wyzwania.

W odpowiedzi na te wyzwania powstały inicjatywy rozwijające modele językowe z myślą o języku polskim, spośród których wymienić warto

co najmniej dwie. Jedną z nich jest projekt Bielik (Ociepa i in. 2024), będący rezultatem unikatowej współpracy między Fundacją SpeakLeash, projektem open-science/open-source, a Akademickim Centrum Komputerowym Cyfronet AGH (ACK), ośrodkiem obliczeń wysokowydajnych. Działalność Fundacji SpeakLeash można interpretować jako oddolną inicjatywę na rzecz cyfrowej suwerenności kulturowej, której celem jest zabezpieczenie reprezentatywności języka polskiego w globalnym ekosystemie AI poprzez budowę autorskich korpusów danych. Rozwój projektu zyskał istotny impuls dzięki zawiązaniu z firmą NVIDIA oficjalnego partnerstwa, które wykracza poza sam dostęp do sprzętu, otwierając drogę do bezpośredniej wymiany myśli technologicznej oraz wykorzystania najnowocześniejszych rozwiązań w zakresie infrastruktury akceleracji obliczeń, a także innowacyjnych metod budowy i trenowania dużych modeli językowych. Z kolei ACK, dysponujące superkomputerami Helios i Athena w ramach infrastruktury PLGrid, zapewniło niezbędne moce obliczeniowe oraz wsparcie eksperckie w optymalizacji procesów treningowych i metod testowania modeli.

W ramach niniejszych badań testowano model Bielik 11B v2.3 w wersji instruktażowej, posiadający 11 miliardów parametrów. Został on wytrenowany na specjalnie wyselekcjonowanych i przetworzonych polskich korpusach tekstowych dostarczonych przez zespół SpeakLeash, w tym na ponad 20 milionach instrukcji. Bielik 11B v2.3, zajmując wysokie lokaty w polskim rankingu OpenLLM Leaderboard (https://huggingface.co/speakleash/Bielik-11B-v2.3-Instruct), legitymuje się potencjalnie wysoką zdolnością modelowania semantycznego i przetwarzania języka polskiego. Warto odnotować dalszy rozwój projektu, czego przykładem są nowsze, ale mniejsze wersje Bielika, takie jak Bielik v3 (w rozmiarach 1,5 i 4,5 mld parametrów) oparte na modelu Qwen 2.5.

Kolejną istotną inicjatywą jest PLLuM (ang. Polish Large Language Model) (PLLuM Consortium 2025), rodzina DMJ powstała na zlecenie polskiego Ministerstwa Cyfryzacji. Projekt ten został zainicjowany przez konsorcjum czołowych polskich instytucji naukowych, w skład którego weszły: Politechnika Wrocławska (lider projektu), NASK, Instytut Podstaw Informatyki PAN, Ośrodek Przetwarzania Informacji — Państwowy Instytut Badawczy, Uniwersytet Łódzki oraz Instytut Slawistyki Polskiej Akademii Nauk. Projekt ten stanowi próbę instytucjonalnej odpowiedzi

na wyzwania cyfryzacji, mającą na celu stworzenie narzędzia legitymizowanego przez autorytety naukowe i przystosowanego do wymogów polskiej sfery publicznej.

W przedstawianym tu badaniu wykorzystano instrukcyjny model PLLuM 12B, a więc model nieco większy niż testowany Bielik. Modele PLLuM wyróżniają się tym, że zostały dostrojone na autorskich zbiorach danych, przy czym proces kurateli treści przez zespół ekspercki pełnił rolę filtra normatywnego, mającego zapewnić zgodność modelu z polskimi standardami etycznymi i językowymi. Zbiór treningowy, obejmujący 100 miliardów słów, został zebrany w sposób organiczny z różnorodnych źródeł, takich jak teksty literackie, dokumenty urzędowe i zasoby instytucji naukowych, z zachowaniem norm prawnych i etycznych, czyli z poszanowaniem prawa autorskiego oraz za zgodą i we współpracy z wydawcami, instytucjami kultury oraz jednostkami publicznymi.

Obok tych modeli o wyraźnie polskim profilu językowym, dostępne są przede wszystkim otwarte modele wielojęzyczne, takie jak rodziny modeli Llama, Falcon, Gemma — również przeznaczone do wdrażania na infrastrukturze własnej, lecz mimo wysokiej skuteczności ogólnej niosące ze sobą ryzyko marginalizacji lokalnych niuansów językowych, gdyż promują wzorce komunikacyjne właściwe dla dominujących (głównie anglojęzycznych) kręgów kulturowych. Dostępne są także duże modele komercyjne wielkiej skali udostępniane jako usługi sieciowe, jak ChatGPT, Claude czy Gemini, obiecujące największe możliwości, rozumiane zarówno jako zdolności językowe, zakres posiadanej wiedzy ogólnej czy umiejętności rozwiązywania zadań.

Wszystkie wymienione typy modeli różnią się nie tylko zakresem danych treningowych, ale również skalą, architekturą i sposobem dostrajania. Te różnice mają istotne znaczenie z perspektywy potencjalnego zastosowania modeli w kontekstach lokalnych, takich jak edukacja, administracja publiczna czy media, w których oczekuje się od technologii językowych nie tylko formalnej poprawności, ale również kulturowej adekwatności i zgodności z obowiązującymi normami społecznymi.

Duże modele językowe nie są neutralnymi narzędziami, lecz stają się mechanizmami cyfrowej reprodukcji stereotypów, utrwalającymi hierarchiczne podziały ról społecznych i płciowych. W literaturze anglojęzycznej wykazano, że DMJ mają tendencję do przypisywania stereotypowych cech

oraz zawodów związanych z tradycyjnymi rolami płciowymi (Kotek i in. 2023). Na przykład są znacznie częściej skłonne uznawać określone profesje za typowo „męskie" lub „żeńskie" oraz przypisywać im przymiotniki zgodne z takimi stereotypami.

Coraz częściej prowadzone są badania weryfikujące to zjawisko także poza językiem angielskim. Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang i Zifan Qian (2024) przeprowadzili analizę płciowych uprzedzeń w wielu językach, ujawniając, że jest to zjawisko uniwersalne, choć przybiera zróżnicowany charakter w zależności od języka. Z kolei Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle i in. (2025) podkreślają, że globalne modele często przekładają stereotypowe wzorce na języki, w których wcześniej nie przejawiały się one w taki sam sposób, co skutkuje zjawiskiem transnarodowego transferu uprzedzeń, przy czym obce kulturowo kalki myślowe są implementowane w lokalne systemy znaczeń.

Instytucje takie jak UNESCO (2024) również ostrzegają, że DMJ często przypisują kobietom role domowe i rodzinne, podczas gdy mężczyznom — biznesowe i technologiczne, co może wzmocnić stereotypy płciowe w społeczeństwie.

W Polsce proces upowszechniania feminatywów stanowi obecnie istotne pole walki symbolicznej o widoczność kobiet w sferze publicznej, co czyni reakcję algorytmów na te formy papierkiem lakmusowym ich społecznej adekwatności. Wyniki sondaży pokazują, że większość osób mieszkających w Polsce deklaruje poparcie dla żeńskich form nazw zawodów, choć niektórzy postrzegają je jako mniej poważne (Hołdyńska 2025). Z kolei środowisko lingwistyczne argumentuje, że feminatywy są intuicyjną i kompatybilną z polską morfologią formą, której użycie ma znaczenie dla reprezentacji i identyfikacji.

Pomimo tych sygnałów w polskiej humanistyce i naukach społecznych brakuje systematycznych analiz nad tym, jak algorytmiczna „gramatyka" płci wpływa na dynamikę inkluzji społecznej w przestrzeni cyfrowej. Dostępne badania typowo dotyczą ogólnych narzędzi przetwarzania języka naturalnego (NLP — *natural language processing*), nie koncentrując się na specyfice przetwarzania i generowania tekstu przez duże modele językowe. Niniejszy artykuł wypełnia tę lukę, dostarczając empirycznego porównania „polskich" otwartych modeli z otwartym modelem uniwersalnym oraz

modelem komercyjnym pod kątem stronniczości płciowej i jakości generowanych form feminatywnych.

METODYKA

W eksperymencie wykorzystano cztery duże modele językowe, reprezentujące odmienne podejścia do cyfrowej reprezentacji języka: od inicjatyw lokalnych (Bielik, PLLuM), przez otwarte projekty wielojęzyczne (Gemma3), po komercyjne systemy globalne (Gemini). Dwa pierwsze, Bielik 11B 2.3 (bielik-11b-v2.3-instruct:Q8_0) oraz PLLuM 12B (pllum-12b-instruct:Q8_0), to otwarte modele dostrojone na obszernych, wysokiej jakości korpusach polskojęzycznych. Jako dodatkowe punkty odniesienia wykorzystano jeszcze dwa uniwersalne modele, a mianowicie również otwarty model Gemma3 12B (Q4_K_M) (Gemma Team i in. 2025), trenowany na danych wielojęzycznych oraz model komercyjny dużej skali — Gemini 2.5 Flash.

W przypadku każdego z modeli wywołania instrukcji realizowane były przez okienka konwersacyjne, z zachowaniem domyślnych parametrów wywołań poza obniżeniem temperatury dla modeli lokalnych. Parametr ten odpowiada za stopień determinizmu generowanego tekstu — niska temperatura sprawia, że model wybiera najbardziej prawdopodobne słowa, co sprzyja precyzji i powtarzalności odpowiedzi, ograniczając jednocześnie ich „kreatywność". Pozostałe parametry techniczne utrzymano na poziomach domyślnych dla danych systemów. Takie podejście zapewniło porównywalność wyników oraz umożliwiło zachowanie wysokiego stopnia replikowalności eksperymentu, bez konieczności wprowadzania dodatkowych zmiennych mogących wpłynąć na profil treningowy modeli.

Modele lokalne wykorzystano w wersjach poddanych kwantyzacji (oznaczenia Q4 i Q8 w nazwach modeli). Kwantyzacja to technika kompresji wag modelu, która pozwala na znaczące obniżenie wymagań sprzętowych przy zachowaniu wysokiej sprawności językowej. Każde zapytanie wykonywane było w nowej sesji, co wykluczało możliwość ewentualnego wpływu treści wcześniejszej konwersacji (kontekstu).

W badaniu wykorzystano zestaw 45 precyzyjnie sformułowanych poleceń w języku polskim, opracowanych specjalnie na potrzeby niniejszego

badania. Konstrukcja autorskiej taksonomii dwunastu klas tematyczno-problemowych opiera się na krytycznej analizie paradygmatu badawczego zaproponowanego przez amerykańskich badaczy (Kotek i in. 2023). O ile koncentrują się oni na identyfikacji uprzedzeń (ang. *gender bias*) poprzez analizę proponowanych zaimków w zdaniach niejednoznacznych, o tyle opisywane badanie rozszerza tę perspektywę o operacyjną sprawność modeli w warunkach wysokiej fleksyjności języka polskiego. Wprowadzenie kategorii GRAM (elastyczność gramatyczna) oraz TRANS (transformacje) pozwala na weryfikację, czy błędy w identyfikacji płci — zdiagnozowane przez tych badaczy w modelach anglojęzycznych — przekładają się w polszczyźnie na błędy w spójności morfologicznej oraz trudności w tworzeniu poprawnych form feminatywnych. Dobór zapytań uwzględniał zarówno różnorodność zagadnień — od stronniczości płciowych przez wnioskowanie kontekstowe, aż po tłumaczenia i neologizmy — jak i format odpowiedzi, co pozwalało na ocenę nie tylko semantyki, ale i elastyczności modeli w analizowaniu i generowaniu tekstu.

Kategoria BIAS (z trzema podzbiorami) służy do analizy algorytmicznej przezroczystości płciowej oraz tendencji modeli do reprodukowania męskocentrycznych wzorców jako form domyślnych. Zawiera ona między innymi polecenia wymagające generowania zdań z zaimkami bez uprzedniego wskazania płci (BIAS-01), krótkie odpowiedzi mailowe czy dialogowe z zachowaniem neutralności płciowej (BIAS-02) oraz testy przełamywania stereotypów zawodowych (BIAS-03).

Kategoria GEN obejmuje cztery podgrupy:

— GEN-01 (proste wnioskowanie kontekstowe), w której model identyfikuje zawód na podstawie opisu czynności;

— GEN-02 (konsekwencja w opisie postaci), sprawdzająca, czy model potrafi utrzymać spójną terminologię i styl w kilku kolejnych zdaniach;

— GEN-03 (role historyczne/społeczne), polegająca na użyciu żeńskich określeń ról istotnych historycznie, literacko czy naukowo;

— GEN-04 (wypowiedź w pierwszej osobie z perspektywy żeńskiej), w której modele miały za zadanie przyjąć narrację kobiecą i opisać własne doświadczenia w konkretnych zawodach, co pozwoliło na ocenę zdolności modeli do budowania spójnych narracji tożsamościowych oraz ich wrażliwości na żeńską perspektywę w dyskursie zawodowym.

Grupa TRANS dzieli się na:

— TRANS-01 — tłumaczenia angielsko-polskie uwzględniające kontekst płciowy;

— TRANS-02 — parafrazy z jawną zamianą form męskich na żeńskie, łącznie z imieniem, przymiotnikami i końcówkami.

Kategoria GRAM została zaprojektowana, by zbadać granice elastyczności systemu językowego modeli, szczególnie w obszarze tworzenia neologizmów oraz adaptacji do zmian w normie polszczyzny:

— GRAM-01 — uzupełnianie luk w różnych przypadkach gramatycznych;

— GRAM-02 — tworzenie neologizmów feminatywnych i ich użycie w zdaniu;

— GRAM-03 — zgodność przymiotnika z „trudnymi" rzeczownikami, jak minister czy pedagog, w formie żeńskiej.

Każde polecenie zostało sformułowane w sposób jednolity dla wszystkich modeli, tak aby porównanie wyników było możliwe bez wpływu różnic w tonie czy stylu. Zapytania obejmowały zarówno generowanie pełnych zdań, uzupełnianie fragmentów, tłumaczenie, jak i parafrazę, co umożliwiło ocenę modeli pod kątem szerokiego spektrum kompetencji językowych oraz ich zdolności do stosowania poprawnych i naturalnych form feminatywów w różnych kontekstach. Wszystkie polecenia udostępnione zostały w postaci otwartych danych badawczych.

Odpowiedzi każdego modelu oceniane były w pięciu kategoriach, w skali od 1 (ocena najniższa) do 5 (ocena najwyższa). Przyjęcie jednolitego systemu ocen pozwoliło na porównywanie wyników między zadaniami, modelami oraz kategoriami testowymi. Oto opis kategorii:

1. Poprawność gramatyczna — celem tego kryterium jest ocena, w jakim stopniu model operuje powszechnie akceptowanymi regułami składniowymi i morfologicznymi współczesnej polszczyzny. Zamiast odwoływania się do statycznych wzorców idealnych ocena opierała się na kompetencji językowej osób sędziujących, uwzględniających zarówno normy wzorcowe, jak i uzualne. Skala ocen obejmuje zakres od rażących błędów uniemożliwiających zrozumienie wypowiedzi (1) po wysoką poprawność, zgodną z oczekiwaniami kompetentnego użytkownika języka (5).

2. Naturalność stylistyczna — to kryterium pozwala ocenić stopień zakorzenienia modelu w polskim kontekście kulturowym oraz jego zdolność

do unikania obcych kulturowo kalk językowych (anglicyzmów strukturalnych). Oceny wahają się od konstrukcji w pełni sztucznych, przypominających dosłowne tłumaczenia maszynowe (1) do eleganckich, płynnych wypowiedzi na poziomie doświadczonej redakcji (5).

3. Skuteczność realizacji zadania — ocena ta mierzy sprawność pragmatyczną modelu, czyli zdolność do adekwatnego reagowania na społeczne i merytoryczne wymogi zawarte w akcie komunikacyjnym. Najniższa ocena przyznawana jest w przypadku odpowiedzi nie na temat lub pominięcia istotnych elementów (1), a najwyższa — gdy model w sposób zwięzły, precyzyjny i wyczerpujący odpowiada na wszystkie wymagania polecenia (5).

4. Poziom użycia feminatywów — kryterium to pozwala na analizę polityki reprezentacji płci, mierząc skłonność modelu do czynienia kobiet widocznymi w dyskursie zawodowym poprzez stosowanie bezpośrednich form żeńskich. Skala obejmuje od całkowitego zignorowania feminatywu (1), przez różne strategie obejścia lub stosowanie omówień typu „pani + zawód" (2–4), aż po pełne i poprawne użycie bezpośredniego feminatywu (5).

5. Progresywność językowa — kryterium diagnozuje potencjał modelu jako agenta zmiany językowej, oceniając jego otwartość na formy awangardowe i neologizmy, które wykraczają poza tradycyjny kanon, lecz są obecne w progresywnych obiegach kultury. Zgodnie z przyjętą zasadą, jeżeli w kryterium 4 oceniono wybór męski lub neutralny (1–2) lub „pani + zawód" (4), tutaj automatycznie przypisywana jest ocena 1. Wyższe wartości (2–5) odzwierciedlają rosnący stopień nowoczesności, od powszechnie akceptowanych feminatywów, przez formy coraz bardziej awangardowe, aż po neologizmy stworzone przez model.

Należy zaznaczyć, że kryteria 4 i 5 tworzą hierarchiczną strukturę analityczną, w której progresywność jest badana wyłącznie w obrębie wygenerowanych form żeńskich. Takie podejście pozwala na precyzyjne oddzielenie inkluzywności binarnej (wybór między formą męską a żeńską) od kreatywności lingwistycznej (wybór między formą utrwaloną a neologizmem). Współzależność tych wskaźników jest zatem zamierzona, gdyż pozwala na uchwycenie momentu, w którym model przestaje być jedynie biernym powielaczem zastanych struktur, a zaczyna testować granice elastyczności systemu językowego. Dzięki temu możliwe jest zdiagnozowa-

nie, czy dany system SI wykazuje tendencję do zachowawczego stosowania feminatywów, czy też promuje formy o charakterze awangardowym.

Tak skonstruowany system oceny umożliwił porównanie nie tylko poziomu kompetencji gramatycznych i stylistycznych modeli, ale także ich zachowań w obszarze inkluzywności płciowej oraz innowacyjności językowej.

W celu zapewnienia wysokiej rzetelności badawczej zastosowano metodę kompetentnych osób sędziujących. W procesie oceny uczestniczył zespół trzech osób stanowiących doświadczony personel Biblioteki Politechniki Śląskiej w Gliwicach. Wybór takiej grupy eksperckiej podyktowany był ich codzienną praktyką zawodową, opartą na bezpośrednim kontakcie z „żywym językiem" używanym przez zróżnicowaną demograficznie grupę osób korzystających z zasobów naukowych i nie tylko (zróżnicowanie pod względem wieku, płci i poziomu wykształcenia). Pozwala to na uzyskanie ocen opartych na najbardziej aktualnych trendach językowych i społecznych, co czyni wyniki bardziej wiarygodnymi niż w przypadku automatyzacji opartej na korpusach referencyjnych, które z natury cechują się dużą bezwładnością w odnotowywaniu zmian uzualnych. Każda z osób dokonywała niezależnej ewaluacji, a ostateczne wyniki wypracowano na drodze konsensusu, co pozwoliło na pełną triangulację perspektyw badawczych. W sytuacjach wystąpienia rozbieżności w ocenach cząstkowych przeprowadzano pogłębioną dyskusję nad danym przypadkiem, aż do wypracowania wspólnego, jednolitego stanowiska.

Ze względu na pionierski charakter podjętej tematyki oraz celowy dobór materiału badawczego niniejsza analiza ma charakter eksploracyjny i deskryptywny. Skupia się na mapowaniu konkretnych zjawisk językowych w odpowiedziach modeli, nie aspirując do miana pełnej weryfikacji statystycznej o charakterze pozwalającym na uogólnianie poczynionych obserwacji.

WYNIKI

Ze względu na zróżnicowaną liczbę zadań w poszczególnych klasach problemowych najpierw obliczono średnie z ocen osób eksperckich w każdym z dwunastu zestawów zadań, a następnie na ich podstawie wyliczono uśrednione wartości w pięciu wymiarach jakości językowej dla każdego

modelu. Pozwoliło to na wyrównanie wpływu klas o różnej liczebności i uzyskanie porównywalnych miar kompetencji modeli, co umożliwiło stworzenie profilu socjolingwistycznego każdego z systemów. Wartości liczbowe przedstawia tabela 1, a ich wizualizacje wykres radarowy (rysunek 1).

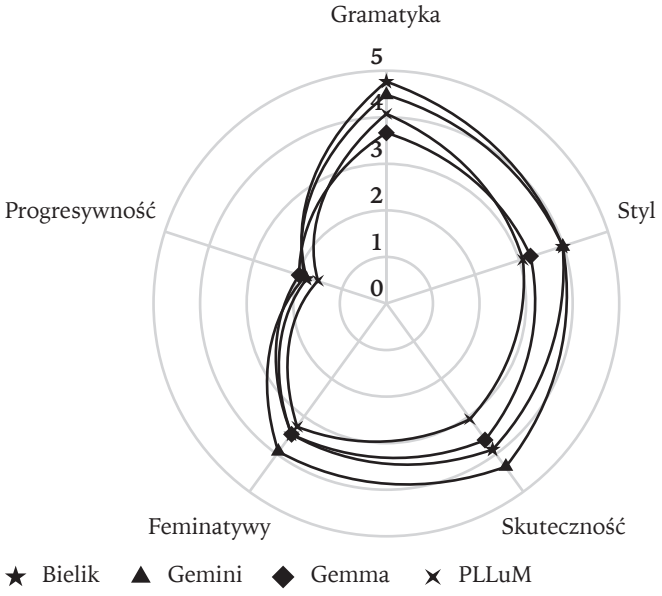
Tabela 1

Wartości poziomu kompetencji modeli w ramach poszczególnych kryteriów oceny

Modele/Kryteria	Gramatyka	Styl	Skuteczność	Feminatywy	Progresywność
Bielik	4,69	3,89	3,91	3,4	1,82
Gemini	4,51	4	4,33	3,98	1,96
Gemma	3,6	3,18	3,56	3,44	2,02
PLLuM	3,98	3,07	3,09	3,22	1,6

Rysunek 1

Poziom kompetencji modeli w ramach poszczególnych kryteriów oceny



W zakresie poprawności gramatycznej prymat wiódł model Bielik, co sugeruje, że lokalne zbiory treningowe skuteczniej internalizują rygorystyczne normy fleksyjne polszczyzny niż systemy uniwersalne. Tuż za nim uplasował się Gemini, natomiast modele Gemma i PLLuM osiągały niż-

sze, choć nadal akceptowalne poziomy, z wyraźnie częstszymi drobnymi błędami odmiany i zgodności rodzaju.

Analiza naturalności stylistycznej pokazała, że Bielik generował teksty najbardziej płynne i idiomatyczne; Gemini prezentował styl bliższy formalnemu rejestrowi, co skutkowało nieco niższą oceną, ale nadal na wysokim poziomie płynności. Modele Gemma i PLLuM były oceniane jako mniej naturalne, co może wskazywać na konieczność dopracowania aspektów stylistycznych przy użyciu większej liczby przykładów w lokalnym kontekście.

Jeśli chodzi o skuteczność realizacji zadań, to Gemini prowadził w liczbie w pełni adekwatnych i wyczerpujących odpowiedzi, co potwierdza jego zdolność do wiernego odtwarzania wymagań poleceń. Bielik uzyskał nieco niższą ocenę, podczas gdy Gemma i PLLuM, choć w większości przypadków prawidłowo realizowały zadania, od czasu do czasu pomijały niektóre szczegóły lub dokonywały niezamierzonych skrótów.

W obszarze polityki reprezentacji płci najwyższą skuteczność wykazał model Gemini. Jego skłonność do konsekwentnego stosowania feminatywów może być interpretowana jako efekt celowej implementacji standardów inkluzywności, typowych dla globalnych korporacji technologicznych. Bielik plasował się tuż za nim, prezentując umiarkowaną, lecz systematyczną skłonność do feminatywów.

We wszystkich analizowanych przypadkach odnotowano deficyt progresywności językowej. Dominacja form asekuracyjnych wskazuje, że modele — niezależnie od pochodzenia — wykazują tendencję do reprodukcji konserwatywnego *status quo*, rzadko wykraczając poza utrwalone społecznie schematy. W tej mierze nieznacznie górował Gemini, a zaraz po nim Bielik.

Podsumowując, lokalny model Bielik osiągnął najwyższe oceny w obszarze gramatyki i stylu, natomiast komercyjny Gemini przodował w realizacji zadań i użyciu feminatywów. Modele Gemma i PLLuM — choć wykazały się solidnymi podstawami — pozostawały w tyle we wszystkich analizowanych wymiarach.

Inny przekrój wyników odnoszący się do zdolności modeli do rozwiązywania poszczególnych typów zadań przedstawiony został w tabeli 2 i zwizualizowany na rysunku 2. W tym przypadku wynikowe oceny przydzielone modelom w poszczególnych typach zadań uwzględniają składowe

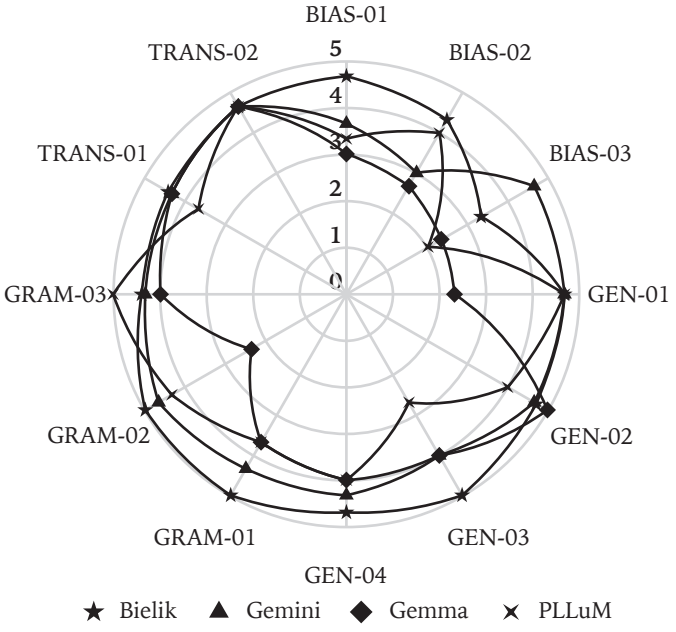
Tabela 2

Wartości poziomów kompetencji modeli w ramach typów zadań

Systemy/Zadania	Bielik	Gemini	Gemma	PLLuM
BIAS-01	4,67	3,67	3	3,33
BIAS-02	4,33	3	2,67	4
BIAS-03	3,33	4,67	2,33	2
GEN-01	4,7	4,67	2,33	4,67
GEN-02	4,7	4,67	5	4
GEN-03	5	4	4	2,67
GEN-04	4,67	4,33	4	4
GRAM-01	5	4,33	3,67	3,67
GRAM-02	5	4,67	2,33	4,33
GRAM-03	4,4	4,33	4	5
TRANS-01	4,4	4,33	4,33	3,67
TRANS-02	4,67	4,67	4,67	4,67

Rysunek 2

Poziom kompetencji modeli w ramach typów zadań



stylistyczne, gramatyczne i skuteczności realizacji poszczególnych poleceń, celowo pomijając aspekty feminizacji języka, gdyż za cel tego przekroju postawiono ocenę sprawności modeli w różnych zastosowaniach przetwarzania i generowania języka polskiego.

Analiza sprawności modeli w poszczególnych kategoriach zadań wskazuje na wyraźną przewagę lokalnie zakorzenionego modelu Bielik, który uzyskał najwyższe noty w większości grup (GEN, BIAS, GRAM). Jedynie w zadaniach z zakresu transformacji i transkrypcji (TRANS), mających charakter bardziej techniczny niż kulturowy, Bielik dzielił pozycję liderską z modelami Gemini i Gemma. Drugą pozycję w obszarach wymagających głębszego wnioskowania konsekwentnie zajmował model Gemini. Z kolei w rywalizacji o trzecią lokatę model PLLuM wykazał przewagę nad Gemmą w obszarach kluczowych dla polskiego kontekstu społeczno-językowego, takich jak poprawność gramatyczna (GRAM) oraz identyfikacja stronniczości płciowej (BIAS).

WNIOSKI

Prezentowane wyniki pozwalają na sformułowanie istotnych wniosków dotyczących kierunków rozwoju polskich modeli językowych, należy jednak pamiętać o ich eksploracyjnym charakterze: opierają się na celowo dobranym zestawie zadań i odnoszą się do konkretnych wersji czterech modeli, testowanych w jednym momencie ich rozwoju. Uzyskane dane stanowią bazę do dalszych, szerzej zakrojonych badań ilościowych, wyznaczając kluczowe obszary wymagające optymalizacji w zakresie kompetencji kulturowej systemów generatywnych. Analiza wyników potwierdziła hipotezę ogólną, że modele trenowane na lokalnych, polskojęzycznych korpusach — a zwłaszcza Bielik — mogą osiągać wyższy poziom poprawności gramatycznej i naturalności stylistycznej niż modele ogólne. Bielik niemal osiągał wyniki zbliżone do komercyjnego Gemini, a w obszarze gramatyki i stylu nawet je przewyższał, co potwierdza, że skrupulatna kuratela lokalnych zasobów wiedzy pozwala na zachowanie wysokiej jakości polskiego kodu kulturowego w systemach SI, stanowiąc skuteczną barierę przed językową unifikacją.

Hipoteza szczegółowa również znalazła częściowe potwierdzenie: Bielik i PLLuM, trenowane na odpowiednio dobranych polskich zasobach,

były bardziej skłonne do użycia feminatywów niż otwarty model Gemma o podobnej liczbie parametrów, a w przypadku Bielika ich stosowanie było też bardziej naturalne. Niemniej jednak to Gemini okazał się najlepszy pod względem konsekwentnego włączania form żeńskich, co sugeruje, że model ten implementuje globalne standardy polityki inkluzywności, które mogą wyprzedzać lokalną dynamikę zmian językowych utrwaloną w polskich korpusach tekstowych.

Wszystkie cztery modele cechowały się niską progresywnością językową, czyli rzadko sięgały po nowe lub awangardowe feminatywy. To zjawisko, choć może wynikać z przekrojowości danych treningowych obejmujących teksty z różnych okresów rozwoju języka, nie wydaje się problematyczne, a może być nawet pożądane — wskazując na zachowawczą rolę algorytmów, które — jako systemy oparte na statystyce zastanej — działają raczej jako stabilizatory normy niż sprawcze podmioty zmiany. Podkreśla to prymat ludzkiego sprawstwa w procesie kształtowania żywej kultury.

Sukces modelu Bielik w starciu z globalną konkurencją dowodzi znaczenia suwerenności technologicznej i pokazuje, że lokalne inicjatywy są w stanie skutecznie reprodukować niuanse polskiej sfery komunikacyjnej. Z kolei PLLuM — mimo imponującej wielkości korpusu treningowego — w zadaniach testowych opracowanych na potrzeby niniejszych badań nie przewyższał znacząco Gemmy; obie te implementacje często rywalizowały o miejsca trzecie.

Należy jednak podkreślić, że zaprezentowane badanie obarczone jest pewnymi ograniczeniami, wynikającymi między innymi z analizy tylko dwóch wybranych inicjatyw lokalnych (Bielik i PLLuM). Trzeba mieć na uwadze, że nawet przy wykorzystaniu tych samych korpusów treningowych odmienna architektura modeli oraz parametry ich dostrajania mają decydujący wpływ na finalne zachowanie systemu. Z tego względu sformułowanych wniosków nie można w sposób bezpośredni generalizować na inne polskie modele językowe. Niemniej jednak opracowany na potrzeby niniejszego badania zestaw zadań testowych oraz autorska taksonomia dwunastu klas problemowych mogą zostać z powodzeniem wykorzystane do ewaluacji kolejnych wersji systemów oraz przyszłych modeli polskojęzycznych, stanowiąc narzędzie do monitorowania postępów w zakresie inkluzywności i poprawności językowej.

Należy również zastrzec, że przyjęta metodologia opiera się na doborze celowym zapytań, skoncentrowanym na zjawiskach o wysokim stopniu trudności lingwistycznej. W związku z tym prezentowane wyniki nie powinny być interpretowane jako całościowy obraz kompetencji kulturowej testowanych systemów, lecz jako analiza ich zachowań w specyficznym, problematycznym wycinku polszczyzny związanym z inkluzywnością językową. Takie podejście, choć ograniczone zakresowo, pozwala na precyzyjną diagnozę konkretnych deficytów modeli w obszarach, które najsilniej rezonują ze współczesnymi przemianami społecznymi.

Wreszcie, wykorzystanie odmiennych, choć domyślnych dla używanego w badaniu środowiska uruchomieniowego DMJ, kwantyzacji (Q4 dla modelu Gemma vs Q8 dla systemów lokalnych) mogło w niewielkim stopniu wpłynąć na wyniki, choć przy różnicach w architekturze i skali parametrów wpływ ten nie powinien zmieniać ogólnych trendów.

Podejście prezentowane w ramach niniejszych badań, w odróżnieniu od pragmatycznych „benchmarków" używanych przez społeczność rozwijającą i komercyjnie wdrażającą technologie DMJ, kładzie nacisk na elastyczność modeli w operowaniu na żywym języku, co w kontekście socjolingwistycznym stanowi wyzwanie jakościowo inne i często trudniejsze niż standardowa weryfikacja poprawności merytorycznej czy faktograficznej odpowiedzi udzielanych przez modele językowe.

Jako dalszy kierunek badań rekomenduje się ocenę kolejnych iteracji polskich modeli językowych oraz ewaluację systemów o większej skali parametrów, szczególnie w kontekście ich społecznego oddziaływania w zastosowaniach interaktywnych. Choć bowiem testowana wersja Bielika stanowi wartościowe narzędzie do przetwarzania polskojęzycznych zasobów tekstowych, to specyfika trybu dialogowego wciąż niesie ze sobą ryzyko błędów, które należy interpretować jako nieodłączny element wczesnego etapu rozwoju tych technologii. W tym ujęciu niniejsze opracowanie staje się istotnym przyczynkiem do szerszych badań nad socjolingwistyką algorytmiczną, nakreślając ramy dla systematycznej ewaluacji systemów generatywnych pod kątem ich ewolucji kulturowej. Pozwala to nie tylko na dalszą optymalizację techniczną, ale i na ciągłą refleksję nad społecznymi konsekwencjami delegowania kompetencji językowych na maszyny, co pozostaje jednym z głównych wyzwań współczesnej humanistyki cyfrowej.

Podziękowania

Składam podziękowania pracownikom Biblioteki Politechniki Śląskiej — Pani Annie Dybał oraz Panu Tomaszowi Gizie — za udział w zespole eksperckim oceniającym wyniki badań. Ich doświadczenie zawodowe i doskonałe wyczucie współczesnej uzualności języka polskiego w sposób istotny wpłynęły na jakość przeprowadzonej analizy.

BIBLIOGRAFIA

Chkirbene Zina, Hamila Ridha, Gouissem Ala, Devrim Unal, 2024, *Large Language Models (LLM) in Industry: A Survey of Applications, Challenges, and Trends*, w: *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, IEEE, s. 229–234 (https://doi.org/10.1109/HONET63146.2024.10822885).

Dai Sunhao, Xu Chen, Xu Shicheng, Pang Liang, Dong Zhenhua, Xu Jun, 2024, *Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era*, w: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, s. 6437–6447 (https://doi.org/10.1145/3637528.3671458).

Gemma Team i in., 2025, *Gemma 3 Technical Report*, arXiv preprint, arXiv:2503.19786 (https://arxiv.org/pdf/2503.19786 [dostęp: 30.01.2026]).

Hołdyńska Anna, 2025, *Feminatywy — co o nich myślimy? Poznaj wyniki największego badania w Polsce*, Pracuj.pl (https://porady.pracuj.pl/praca/feminatywy-co-o-nich-myslimy-wyniki-badania [dostęp: 30.01.2026]).

Kotek Hadas, Dockum Rikker, Sun David, 2023, *Gender Bias and Stereotypes in Large Language Models*, w: *Proceedings of The ACM Collective Intelligence Conference*, Association for Computing Machinery, s. 12–24 (https://doi.org/10.1145/3582269.3615599).

Kumar Vimal, Srivastava Priyam, Dwivedi Ashay, Budhiraja Ishan, Ghosh Debjani, Goyal Vikas, Arora Ruchika, 2024, *Large-Language-Models (LLM)-Based AI Chatbots: Architecture, In-Depth Analysis and Their Performance Evaluation*, w: KC Santosh, Aaisha Makkar, Myra Conway, Ashutosh K. Singh, Antoine Vacavant, Anas Abou el Kalam, Mohamed-Rafik Bouguelia, Ravindra Hegadi (red.), *Recent Trends in Image Processing and Pattern Recognition*, RTIP2R 2023, Communications in Computer and Information Science, vol. 2027, Springer, Cham, s. 237–249 (https://doi.org/10.1007/978-3-031-53085-2_20).

Latos Agnieszka, 2020, *Feminatywy w stanowiskach Rady Języka Polskiego. Język a ewolucja normy społecznej*, „Postscriptum Polonistyczne", 26(2), s. 227–242 (https://doi.org/10.31261/PS_P.2020.26.18).

Mitchell Margaret, Attanasio Giuseppe, Baldini Ioana, Clinciu Miruna, Clive Jordan, Delobelle Pieter i in., 2025, *SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models*, w: Luis Chiruzzo, Alan Ritter, Lu Wang (red.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers*, Association for Compu-

tational Linguistics, s. 11995–12041 (https://aclanthology.org/2025.naacl-long.600.pdf [dostęp: 30.01.2026]).

Ociepa Krzysztof, Flis Łukasz, Wróbel Krzysztof, Gwoździej Adrian, Kinas Remigiusz, 2024, *Bielik 7B v0.1: A Polish Language Model—Development, Insights, and Evaluation*, arXiv preprint, arXiv:2410.18565 (https://arxiv.org/abs/2410.18565 [dostęp: 20.10.2025]).

Örpek Zeynep, Tural Büşra, Destan Zeynep, 2024, *The Language Model Revolution: LLM and SLM Analysis*, w: *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, IEEE, s. 1–4 (https://doi.org/10.1109/IDAP64064.2024.10710677).

PLLuM Consortium, 2025, *PLLuM: A Family of Polish Large Language Models*, arXiv preprint, arXiv:2511.03823 (https://arxiv.org/abs/2511.03823 [dostęp: 30.12.2025]).

UNESCO, 2024, *Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls in Large Language Models*, International Research Centre on Artificial Intelligence, UNESCO (https://unesdoc.unesco.org/ark:/48223/pf0000388971 [dostęp: 30.01.2026]).

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan, Kaiser Łukasz, Polosukhin Illia, 2017, *Attention is All you Need*, w: Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, Roman Garnett (red.), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates Inc. (https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [dostęp: 30.01.2026]).

Zhao Jinman, Ding Yitian, Jia Chen, Wang Yining, Qian Zifan, 2024, *Gender Bias in Large Language Models Across Multiple Languages*, arXiv preprint, arXiv:2403.00277 (https://arxiv.org/pdf/2403.00277 [dostęp: 30.01.2026]).

Abstrakt

W artykule przedstawiono porównawczą analizę czterech dużych modeli językowych — dwóch lokalnych krajowych inicjatyw (Bielik 11B v2.3, PLLuM 12B), otwartego uniwersalnego modelu wielojęzycznego (Gemma3 12B) oraz komercyjnego modelu o dużej skali (Gemini 2.5 Flash) — w sytuacji generowania tekstu w języku polskim ze szczególnym uwzględnieniem feminatywów. Do oceny wykorzystano 45 zapytań pogrupowanych w 12 klas tematycznoproblemowych, co pozwoliło na analizę reprodukcji wzorców językowych i społecznych w pięciu wymiarach: poprawności gramatycznej, naturalności stylistycznej, skuteczności realizacji zadania, poziomu użycia feminatywów oraz progresywności językowej. Eksperymenty prowadzono przy użyciu interfejsu dialogowego, a odpowiedzi poddano ocenie eksperckiej. Wyniki wykazały, że model Bielik przewyższał konkurencję w obszarach poprawności i stylistyki, zaś Gemini osiągał najlepsze rezultaty w precyzyjnej realizacji poleceń i częstości stosowania form żeńskich. Modele Gemma3 i PLLuM plasowały się niżej, zwłaszcza realizując zadania wymagające precyzyjnej adaptacji form feminatywnych. Analiza wykazała zachowawczość systemów w obszarze progresywności językowej, co sugeruje utrwalanie tradycyjnych struktur binarnych przez algorytmy szkolone na korpusach zastanych. Badanie potwierdza kluczowe znaczenie rozwoju otwartych modeli w ramach lokalnych inicjatyw

krajowych, wskazując jednocześnie na konieczność dalszej ewolucji systemów polskojęzycznych oraz metod rzetelnego audytu ich wrażliwości społeczno-językowej.

słowa kluczowe: Bielik, PLLuM, duże modele językowe (DMJ), feminatywy, inkluzywność językowa, stronniczość algorytmiczna, sztuczna inteligencja, socjolingwistyka cyfrowa

GRAMMAR, STYLE AND INCLUSION IN POLISH-DEVELOPED LANGUAGE MODELS: AN EVALUATION OF FEMINATIVE USAGE IN BIELIK AND PLLUM

Anna Kamińska
(University od Silesia)

Abstract

The article presents a comparative analysis of four large language models (LLMs) — two local national initiatives (Bielik 11B v2.3, PLLuM 12B), an open universal multilingual model (Gemma3 12B) and a large-scale commercial model (Gemini 2.5 Flash) — generating text in Polish, with a particular focus on feminatives. The evaluation used 45 queries grouped into 12 thematic and problem-based classes, allowing for an analysis of the reproduction of linguistic and social patterns across five dimensions: grammatical correctness; stylistic naturalness; task efficiency; the level of feminative usage; and linguistic progressiveness. The experiments were conducted using a dialogue interface, and the responses were subjected to expert assessment. The Bielik model outperformed its competitors in grammar and stylistics, while Gemini achieved the best results in precise instruction following and the frequency of feminine forms. Gemma3 and PLLuM ranked lower, especially in tasks requiring precise adaptation of feminative forms. Analysis revealed that the models were conservative in terms of linguistic progressiveness, suggesting the perpetuation of traditional binary structures by algorithms trained on legacy corpora. The study underscores the critical importance of developing open language models within local national initiatives, highlighting the need for both the continued evolution of Polish-language systems and further development of reliable auditing methods to assess their sociolinguistic sensitivity.

keywords: Bielik, PLLuM, Large Language Models (LLM), feminatives, linguistic inclusivity, algorithmic bias, artificial intelligence, digital sociolinguistics