Studium Licencjackie

Kierunek Metody Ilościowe w Ekonomii i Systemy Informacyjne

Antoni Czołgowski
Nr albumu 121306

Czy sztuczna inteligencja dyskryminuje? Analiza stronniczości modelu językowego Gemma 2 2B

Praca licencjacka
pod kierunkiem naukowym
dr Magdalena Smyk-Szymańskiej
Kolegium Analiz Ekonomicznych
Instytut Statystyki i Demografii

Warszawa 2025

Spis treści

1. Wprowadzenie

Trwająca obecnie rewolucja technologiczna, napędzana gwałtownym rozwojem sztucznej inteligencji (AI), w istotny sposób wpłynie na niemal wszystkie aspekty naszego życia. Zmieni nie tylko sposób i efektywność pracy, ale także nasze codzienne zwyczaje, sposoby spędzania wolnego czasu oraz budowania relacji międzyludzkich. Co jednak najważniejsze, jak każda znacząca transformacja, będzie miała istotne konsekwencje społeczne. Z jednej strony, dzięki nowym technologiom obniży się próg wejścia do branż wcześniej niedostępnych dla części społeczeństwa, co zwiększy szanse wielu osób na rynku pracy. Z drugiej jednak, pojawi się nowa grupa, która dzięki efektywniejszemu wykorzystaniu AI zdobędzie przewagę, generując ponadprzeciętne zyski i prowadząc do wzrostu nierówności majątkowych.

Istotnym zadaniem współczesnych badań jest dokładne opisanie wykorzystywanych narzędzi oraz skierowanie uwagi opinii publicznej na szanse i zagrożenia, jakie niesie ze sobą rozwój AI. Obecny intensywny wyścig technologiczny często uniemożliwia czołowym firmom poświęcenie odpowiednich zasobów na analizę negatywnych konsekwencji ich rozwiązań, mimo rosnących oczekiwań społecznych w tym zakresie. Dlatego szczególną rolę odgrywają organy regulacyjne, które muszą dostosować poziom kontroli tak, aby nie hamować innowacyjności, a jednocześnie zapewnić użytkownikom poczucie bezpieczeństwa i ochronę przed potencjalnymi zagrożeniami. Właśnie z tych względów zagadnienia związane z etycznym wykorzystaniem narzędzi sztucznej inteligencji zyskują coraz większe znaczenie.

Przeprowadzone w pracy badanie empiryczne miało na celu zweryfikowanie czy model językowy równie skutecznie przewiduje odpowiedzi dotyczące kwestii społecznych, politycznych oraz ekonomicznych wśród reprezentantów różnych grup społeczno-demograficznych z jedenastu krajów (wybranych w taki sposób, żeby odzwierciedlały globalne zróżnicowanie). Identyfikacja ewentualnych niedoreprezentowanych (w rozumieniu: niedoszacowanych przez AI) grup pozwoli decydentom na podjęcie ukierunkowanych działań regulacyjnych oraz przeprowadzenie pogłębionych analiz przyczyn ewentualnych uprzedzeń w działaniu modelu.

W pracy przyjęto metodologię „czarnej skrzynki” (ang. black box), polegającą na analizie modelu językowego bez ingerencji w jego architekturę czy szczegółowe parametry techniczne. Szczegóły dotyczące tej metody zostały szerzej opisane w rozdziale drugim. W ramach badania modelowi językowemu zadano zestaw pytań, prosząc go o wcielenie się w role reprezentujące różne kombinacje cech realnych respondentów, a następnie porównano jego odpowiedzi z

odpowiedziami udzielonymi przez reprezentantów grup na podstawie ich odpowiedzi w badaniu World Values Survey (Wave 7). Można by się spodziewać, że różnice między odpowiedziami udzielonymi przez model językowy a przeciętnie przez przedstawicieli grup społeczno-demograficznych powinny być niewielkie a odchylenia podobne dla każdej z analizowanych grup. Ma to duże znaczenie, ponieważ brak obiektywizmu modelu językowego w tym przypadku oznacza, że model dopasowuje się lepiej do jednych grup niż do innych, co w odpowiednim kontekście może przełożyć się na zachowania dyskryminujące. Zdaniem autora niniejszej pracy współczesne badania powinny kłaść szczególny nacisk na zapewnienie równości dla grup niedoreprezentowanych.

Główną motywacją do podjęcia tego tematu jest coraz powszechniejsze wykorzystanie algorytmów sztucznej inteligencji przy jednocześnie ograniczonej świadomości społecznej dotyczącej ryzyka związanego z tymi technologiami. Firmy i instytucje często błędnie utożsamiają postęp technologiczny jedynie z automatyzacją procesów i przekazywaniem decyzji algorytmom AI, pomijając fakt, że systemy te nie są wolne od uprzedzeń czy błędów, a ich stosowanie może prowadzić do poważnych, niezamierzonych konsekwencji społecznych. Większość literatury z tej dziedziny koncentruje się na opisie dużych modeli korporacyjnych takich jak ChatGPT lub Gemini, podczas gdy modele z serii Gemma są najczęściej wykorzystywane jedynie jako punkt odniesienia. Z tego względu szczególnie istotne jest przeprowadzenie badań, które pozwolą lepiej zrozumieć mechanizmy działania również tych modeli, umożliwiając tym samym rzetelną ocenę przeprowadzanych testów oraz przeciwdziałanie potencjalnym negatywnym skutkom ich wdrożenia.

Praca składa się z trzech rozdziałów. W pierwszym z nich przedstawiono przegląd aktualnej literatury przedmiotu. Rozdział drugi poświęcono szczegółowemu opisowi metodologii badania, natomiast w rozdziale trzecim omówiono przebieg przeprowadzonego eksperymentu oraz przedstawiono uzyskane wyniki. W zakończeniu zawarto kluczowe wnioski płynące z analizy, a także praktyczne rekomendacje sformułowane na ich podstawie.

1.1 Repozytorium z kodem i bazą danych

https://github.com/AntoniCzolgowski/Examining_Gemma_2_2B_discrimination

2. Przegląd literatury

2.1 Stronniczość modeli językowych

Już w 2021 roku literatura naukowa wskazywała na szerokie spektrum obszarów, w których systemy oparte na sztucznej inteligencji wykazywały tendencje do stronniczości. Badanie The Greenlining Institute ujawniło problemy z tym związane w sektorach tak różnorodnych jak opieka zdrowotna, zatrudnienie, edukacja, finanse, mieszkalnictwo czy optymalizacja cen (The Greenlining Institute, 2021). Raport ten ilustruje, jak algorytmy – definiowane jako zestawy reguł podejmujących decyzje na podstawie danych – mogą nieświadomie utrwalać istniejące nierówności. Przykładami są faworyzowanie mężczyzn w procesach rekrutacyjnych (przypadek Amazona) czy zawyżanie kosztów kredytów dla osób należących do mniejszości rasowych. Autorzy badania podkreślają, że źródła stronniczości tkwią zarówno w subiektywnych wyborach projektantów systemów, jak i w tendencyjnych danych treningowych.

Istotnym wyzwaniem identyfikowanym w literaturze jest również brak jednoznacznego rozróżnienia między stronniczością (ang. bias) a dyskryminacją. Badacze z King' s College London już w 2021 argumentowali, że ocena, czy dany model językowy jest dyskryminacyjny, zależy w dużej mierze od kontekstu i sposobu jego praktycznego zastosowania (Ferrer, et al., 2021).

Należy jednak odróżnić potencjalną stronniczość algorytmu od systemów, które w realnych wdrożeniach ewidentnie działały na szkodę określonych grup społecznych (Granberg & Geiger, 2024). W 2024 roku ujawniono algorytm szwedzkiej agencji ubezpieczeń społecznych, który niesprawiedliwie kierował grupy marginalizowane do kontroli pod kątem oszustw socjalnych. Do tej samej grupy algorytmów autor zalicza również wszelkie systemy używane przez organy ścigania, których błędne zalecenia w stronę grup marginalizowanych, co jakiś czas pojawiają się w nagłówkach gazet (Heaven, 2020), (Johnson & Johnson, 2023), (Skelton, 2025).

Praca nad sztuczną inteligencją, w obliczu narastającego problemu braku transparentności twórców modeli (Heinrichs, 2021), wiąże się z licznymi wyzwaniami etycznymi, prawnymi i społecznymi. Warto przy tym postrzegać analizę stronniczości modeli językowych nie tylko jako identyfikację problemu, ale również jako narzędzie do głębszego zrozumienia i potencjalnego adresowania istniejących uprzedzeń społecznych. Badając tendencyjność modeli, możemy lepiej zrozumieć, dlaczego i w jaki sposób perspektywy pewnych grup są niedostatecznie reprezentowane lub pomijane. Stronniczość ta często odzwierciedla bowiem

poglądy twórców lub wzorce obecne w danych treningowych, które same w sobie mogą być odbiciem realnych nierówności.

Zadaniem środowiska naukowego i współpracujących z nim organów jest identyfikacja obszarów ryzyka i proponowanie środków zaradczych. Warto zauważyć postęp w sferze regulacyjnej – podczas gdy w 2020 roku, według autorów publikacji „Bias and Discrimination in AI: a cross-disciplinary perspective", europejskie ramy prawne były niewystarczające do efektywnej walki z opisanymi wyzwaniami (Ferrer, et al., 2021), obecnie Unia Europejska stała się liderem w regulacji systemów AI, co stanowi ważny krok naprzód.

Aby jednak dogłębnie zrozumieć mechanizmy leżące u podstaw tych zjawisk, niezbędne jest przyjrzenie się samym narzędziom. Zrozumienie, czym są współczesne modele językowe, jakie metodologie stosuje się do ich badania – zwłaszcza w kontekście systemów o nieprzejrzystej architekturze, tzw. „czarnych skrzynek" – oraz jakie są ich fundamentalne ograniczenia sprzyjające powstawaniu uprzedzeń, stanowi punkt wyjścia do dalszej, szczegółowej analizy problematyki stronniczości i dyskryminacji przedstawionej w kolejnych częściach tego przeglądu.

2.2. Czym jest czarna skrzynka?

Do listopada 2022 roku, kiedy swoją premierę miał pierwszy publiczny model językowy firmy OpenAI - ChatGPT, dominującym podejściem w badaniach nad systemami AI, włączając w to także duże modele językowe, było faworyzowanie modeli z otwartym kodem oraz dostępnymi publicznie danymi treningowymi. W kręgach akademickich, wyrażano szczególną niechęć wobec systemów stworzonych algorytmicznie, gdzie rola twórców, była sprowadzana do zapewnienia danych treningowych, co później, uniemożliwiało im nawet zrozumienie sposobu, w jaki model wydawał predykcje i komunikował się z rozmówcami. W 2019 roku, Harvardzcy badacze zaproponowali także rozwiązanie, żeby przy podejmowaniu wszystkich odpowiedzialnych decyzji, mogących mieć konsekwencje w rzeczywistym świecie, nie używać zamkniętych modeli uczenia maszynowego, chyba że upewniono się, że nie można skonstruować żadnego interpretowalnego modelu, który osiągnąłby ten sam poziom dokładności (Rudin, 2019). Porównano takie modele do czarnych skrzynek (eng. black box) i jasno postulowano, że przekonanie, że aby uzyskać wysoką dokładność modeli AI, trzeba poświęcić ich interpretowalność jest błędne. Innymi słowy, wiele osób zakładało, że modele muszą być skomplikowane i nieprzejrzyste, żeby osiągać najlepsze wyniki, co według autorów było nieprawdziwym założeniem.

Jednak pojawienie się komercjalnej wersji ChatGPT, która w miesiąc osiągnęła pułap 100 milionów użytkowników, stając się najszybciej rozwijającą się aplikacją konsumencką w historii, zburzyło tą pozorną równowagę w obszarze sztucznej inteligencji, a szczególnie wśród dużych modeli językowych (Hu, 2023). Po 3 latach nieustannego wyścigu między spółkami technologicznymi o prymat w szerzeniu innowacji, przewagę osiągnęły modele korporacyjne. Modele w założeniu z zamkniętym kodem oraz niedostępnymi danymi treningowymi, tak zwane modele zamknięte (Minaee, et al., 2024), aktualnie posiadają prawie 100% udziałów w rynku modeli językowych w USA (Bailyn, 2025). Opisana sytuacja wywarła także znaczący wpływ na trendy w środowisku akademickim. Pojęciem czarnej skrzynki oznacza się teraz, zyskującą coraz większą popularność metodologię badania modeli językowych bez dostępu do wewnętrznej architektury i zaawansowanych parametrów LLM (Lapid, et al., 2024).

Metodologia „czarnej skrzynki" (ang. black box) znajduje szerokie zastosowanie w wielu dziedzinach, m.in. w analizach prawniczych (Schroeder & Wood-Doughty, 2025), eksperymentach sprawdzających odporność modeli językowych na próby manipulacji (Lapid, et al., 2024) czy też badaniach nad wpływem zniekształconych danych treningowych (Gloaguen, et al., 2025). W niniejszej pracy również wykorzystano tę metodologię ze względu na możliwość przeprowadzenia analizy modelu językowego bez konieczności ingerowania w jego wewnętrzną architekturę lub parametry techniczne. Dzięki temu możliwe było skupienie się wyłącznie na jakości i bezstronności generowanych odpowiedzi, co bezpośrednio odpowiada na postawione pytania badawcze.

2.3. Poglądy modeli językowych

2.3.1. Wady modeli językowych

Literatura (Kirk, et al., 2021) pokazuje, że ogólnodostępne modele, mimo iż w założeniu powinny być bezstronne, ogólnodostępne i niepogłębiające dyskryminacji mają swoje poglądy, pogłębiając istniejące już uprzedzenia lub jak wykazano w dalszej części rozdziału, umiejscawiają się w lewej, dolnej ćwiartce kompasu politycznego. Zanim jednak omówione zostaną wyniki tych badań, należy zaznaczyć, że współczesne modele językowe, takie jak ChatGPT, obciążone są istotnymi wadami, prowadzącymi do błędów w generowanych przez nie treściach (Spennemann, 2023). Wśród tych niedoskonałości wymienić można generowanie fikcyjnych źródeł (tzw. halucynacje danych), powierzchowność i brak spójności w argumentacji, a także uproszczenia wynikające z nieuwzględnienia różnorodności kulturowej i kontekstowej. Szczególnie istotnym problemem tych systemów, jak wskazuje literatura, jest brak faktycznej bezstronności i obecność niejawnych uprzedzeń, które wpływają na ich

odpowiedzi, pogłębiając nierówności lub utrwalając określone narracje. Uwzględnienie tych ograniczeń jest kluczowe dla zrozumienia dalszych wyników i wniosków przedstawionych w niniejszej pracy.

2.3.2 Sposoby na analizę poglądów modelu językowego

Publikacje przeanalizowane przez autora opisują szerokie spektrum metodologii, jakie zostały wykorzystane podczas badań nad zrozumieniem i zbadaniem przyczyn obecności określonych poglądów w dużych modelach językowych. Autorzy analizy opublikowanej przez spółkę Anthropic stworzyli zbiór danych GlobalOpinionQA, wykorzystując pytania z badania World Values Survey, aby ocenić, jak model językowy reprezentuje globalne opinie (Durmus, et al., 2023). Dla każdego pytania model generuje rozkłady prawdopodobieństwa, które porównywane są z uśrednionymi odpowiedziami krajowymi za pomocą metryki opartej na 1 minus odległość Jensena-Shannona. Badanie obejmuje trzy warunki eksperymentalne: domyślny prompt, prompt z kontekstem narodowym oraz zapytania (prompty) w różnych językach. Wyniki pokazują, że model domyślnie odzwierciedla opinie krajów zachodnich, a modyfikacje kontekstowe przesuwają odpowiedzi, często generując uproszczone stereotypy, przy czym zmiana języka nie gwarantuje lepszej reprezentacji lokalnych perspektyw. Badanie to stanowi ważny wkład w analizę reprezentacji kulturowej w dużych modelach językowych.

Podobna publikacja ze Stanford University przedstawia test do oceny zgodności dystrybucyjnej generowanych przez modele odpowiedzi z rzeczywistymi rozkładami opinii (Meister, et al., 2024). Autorzy porównali różne metody przedstawiania rozkładu odpowiedzi generowanych przez modele. Pierwszą metodą są logarytmiczne prawdopodobieństwa (ang. log-probabilities), gdzie model określa prawdopodobieństwo kolejnych tokenów – czyli pojedynczych elementów tekstu, takich jak słowa lub fragmenty słów. Pozostałe metody to sekwencyjne próbkowanie, które polega na generowaniu wielu odpowiedzi w celu oszacowania ich zmienności, oraz bezpośrednia werbalizacja, gdzie model przedstawia rozkład w formie czytelnego tekstu (np. „A: 40%, B: 35%, C: 25%”). Dodatkowo, testowano metody sterowania, takie jak persona steering, polegające na dodaniu do prompta specyficznego opisu grupy docelowej (np. „odpowiedz jak typowy obywatel USA”), oraz few-shot steering, gdzie model otrzymuje kilka przykładów reprezentujących odpowiedzi danej grupy. Wyniki wskazują, że metoda bezpośredniej werbalizacji lepiej oddaje złożoność rozkład odpowiedzi respondentów, ponieważ eliminuje problem nadmiernej koncentracji prawdopodobieństwa, typowej dla logarytmicznych prawdopodobieństw, a zastosowanie kilku przykładów kontekstowych (few-shot steering) znacząco poprawia precyzję symulacji, mimo że nadal występuje tzw.

knowledge-to-simulation gap – czyli różnica między wiedzą modelu o dystrybucji a jego zdolnością do jej realistycznego próbkowania. Podobnie jak w publikacji Anthropic, badania ujawniają, że domyślne ustawienia modeli faworyzują perspektywy krajów zachodnich, a interwencje kontekstowe, choć przesuwają wyniki w stronę specyficznych, lokalnych opinii, często prowadzą do uproszczeń i stereotypizacji. Połączenie obu publikacji podkreśla, że aby modele mogły lepiej symulować zróżnicowane opinie społeczne, konieczne są dalsze badania nad precyzyjnym tworzeniem promptów do modeli, na co zwracają uwagę sami twórcy ChatGPT (Brockman, 2025).

Wnioski z przywołanych wyżej prac naukowych są potwierdzone poprzez inne badania obejmujące znacznie mniej abstrakcyjny obszar tematyczny oraz prowadzące do łatwiejszych do implementacji w praktyce zastosowań. Wyniki badania przeprowadzonego na University of California, wskazują, że precyzyjne formułowanie zapytań (promptów) odgrywa kluczową rolę w dokładnym odwzorowaniu lokalnych opinii (Zhao, et al., 2023). W tej metodzie do bazowego modelu językowego dodawany jest specjalny moduł – transformator – który, ucząc się na kilku przykładach (czyli parach prompt–odpowiedź), potrafi skutecznie dostosować model do preferencji danej grupy. Nawet przy ograniczonej liczbie przykładów, eksperymenty wykazały, że metoda GPO uzyskała o 7,1% wyższą zgodność (eng. alignment score) w porównaniu z metodą In-context Finetune, gdzie model jedynie korzysta z dostarczonych przykładów bez modyfikacji parametrów.

Praktyczne zastosowanie powyższych metod wykorzystano także w symulacji sondaży społecznych, których wyniki jednoznacznie wskazują, że modele takie jak GPT-3.5 lepiej odzwierciedlają opinie danej kultury, gdy prompt jest formułowany w języku dominującym dla tej kultury (AlKhamissi, et al., 2024). Według badaczy, kluczową rolę odgrywa tutaj skład danych treningowych – modele o bardziej zrównoważonym, wielojęzycznym pretrainingu skuteczniej oddają pluralizm kulturowy. W tej pracy zaproponowano także metodę Anthropological Prompting, opartą na ramionach myślenia antropologicznego, która pozwala modelom głębiej uwzględniać kontekst społeczno-kulturowy, zwłaszcza w odniesieniu do mniejszości.

Opisane podejścia łączy obserwacja, że domyślne ustawienia LLM-ów sprzyjają reprodukcji uproszczonych, zachodnich perspektyw, niezależnie od specyfiki kulturowej. Zarówno praca badaczy z University of California, jak i badania nad kulturową zgodnością podkreślają, iż precyzyjne, kontekstowe formułowanie promptów jest niezbędne dla prawidłowego

odwzorowania lokalnych opinii. Niewielkie interwencje w strukturę zapytań (promptów) mogą znacząco poprawić jakość generowanych odpowiedzi, choć jednocześnie niosą ryzyko uproszczenia złożonych przekazów. Wspólnym mianownikiem opisanych metod jest potrzeba interdyscyplinarnego podejścia – łączącego techniki uczenia maszynowego z umiejętnością dopasowania metodologii do kontekstu badania.

2.3.3. Skąd biorą się poglądy modeli językowych?

Staje się więc faktem, że modele językowe mają swoje poglądy i przekonania, w rozumieniu, że ich odpowiedzi nie są neutralnym odbiciem rzeczywistości, lecz rezultatem wpływu danych treningowych oraz metod uczenia, które preferują określone narracje i perspektywy kulturowe. Badania wskazują, że domyślne ustawienia modeli odzwierciedlają przede wszystkim zachodni punkt widzenia, a próby dostosowania ich wypowiedzi do konkretnych lokalnych społeczności często kończą się niezamierzonymi uproszczeniami lub stereotypizacją. Przykładowo, badanie Davida Rozado (Rozado, 2024) wykazało, że większość nowoczesnych konwersacyjnych modeli językowych, takich jak GPT-4, Google Gemini czy modele z serii Llama, obiera lewicowe preferencje polityczne, podczas gdy modele bazowe (niekomercyjne) pozostają neutralne, choć mniej spójne w odpowiedziach. We wnioskach jasno podkreślono, że preferencje te ujawniają się po procesie fine-tuningu (dostrajania), a ich źródło może tkwić w danych treningowych lub kulturowych normach, choć pozostaje to niejasne. Pokazano również, że celowe dostosowanie polityczne modeli jest możliwe przy użyciu niewielkich, ukierunkowanych zbiorów danych.

Rynek modeli językowych zdominowany jest przez modele zamknięte a ich twórcy, będąc w wygodnej sytuacji umożliwiającej powołanie się na tajemnice przedsiębiorstwa, mogą wpływać na wyniki i zachowanie modelu sposobami takimi jak Reinforcement Learning with Human Feedback (RLHF), co może nieświadomie wprowadzać kulturowe uprzedzenia lub uproszczenia (Glukhov, et al., 2024). Precyzyjne formułowanie promptów a następnie analiza odpowiedzi udzielonych przez model jest w stanie odpowiedzieć na wiele nasuwających się pytań, dlatego niniejsze badanie konsekwentnie przeprowadzono metodą czarnej skrzynki. Istniejąca literatura, jak udowodniono w tym rozdziale, zawiera wiele opracowań w tej konwencji.

Bardzo ciekawą hipotezę na temat opisywanego zjawiska dostarcza praca „Large Language Models Reflect the Ideology of their Creators” (Buyl, et al., 2024). Autorzy, odwołując się do metody czarnej skrzynki, przyjęli podejście oparte na analizie odpowiedzi modeli bez

zagłębiania się w ich wewnętrzne mechanizmy, co współgra z omawianym wcześniej wyzwaniem badania zamkniętych systemów. W badaniu wykorzystano dwuetapową strategię promptowania: w pierwszym etapie model opisywał wybrane osoby polityczne np. Edward Snowden, a w drugim oceniano moralny wydźwięk tych opisów na skali Likerta, od „bardzo negatywnego" do „bardzo pozytywnego". Wybór 3991 postaci historycznych, filtrowanych pod kątem globalnej rozpoznawalności i współczesnego znaczenia, pozwolił na stworzenie szerokiego kontekstu ideologicznego. Odpowiedzi analizowano w sześciu językach ONZ, co ujawniło wpływ języka i regionu twórców na wyniki. Kolejne kroki obejmowały mapowanie pozycji ideologicznych oraz porównania różnic w ocenach między modelami z różnych stref geopolitycznych, takich jak USA, Chiny czy Rosja. Wnioski są fascynujące: modele jak Google Gemini faworyzują progresywne wartości społeczne, podczas gdy Grok od xAI skłania się ku suwerenności narodowej, co obrazuje, jak ideologia twórców odbija się w odpowiedziach niczym w lustrze. Wyobrażając sobie model jako narratora, który, opisując tę samą postać, w języku angielskim wychwala jej walkę o prawa człowieka, a w chińskim podkreśla lojalność wobec państwa – to pokazuje, jak subtelne niuanse kulturowe przenikają do cyfrowych „umysłów". Najciekawsze jest to, że różnice ujawniają się nie tylko między regionami, ale i w ramach jednego bloku geopolitycznego, jak w przypadku chińskich modeli Alibaba i Baidu, gdzie jeden wspiera zrównoważenie, a drugi centralne planowanie. Praca dowodzi, że neutralność modeli to iluzja, a ich poglądy są echem intencji projektantów, co każe nam zastanowić się, czy kiedykolwiek usłyszymy od nich „obiektywną" opowieść.

2.3.4. Czym jest dyskryminacja i jak ją badać?

Inną odnogą w tej samej dziedzinie badań, są prace analizujące zjawisko dyskryminowania określonych grup społecznych przez modele językowe. Dyskryminacja i uprzedzenia w tym kontekście to tendencyjne traktowanie, gdzie model, niczym cyfrowy narrator, może faworyzować jedne grupy, a inne przedstawiać stereotypowo lub pomijać, odzwierciedlając nierówności zakorzenione w danych treningowych. Istniejąca literatura opisuje, jak model oparty na generatywnej sztucznej inteligencji może tworzyć treści wpływające na nierówne szanse, np. w rekrutacji, co pokrywa się z wynikami innych badań, ukazujących jak często dochodzi do zjawiska stronniczości modeli językowych (Zollo, et al., 2025) (Hu, et al., 2024) .

Badanie takich zjawisk wymaga podejścia wykraczającego poza klasyczne miary równości, nieprzystające do narracyjnej natury GenAI. Stosuje się eksperymenty symulujące realne scenariusze, jak analiza podsumowań CV czy testy prowokacyjne, by wykryć toksyczność i nierówności w odpowiedziach modelu. Metodologia zakorzeniona jest w rygorystycznym

eksperymentalnym testowaniu w różnych scenariuszach, łącząc klasyczne metryki, takie jak ROUGE (Lin, 2004), z kontekstowymi analizami – np. sentymentu opisów CV – oraz symulacjami jak wybór kandydatów na rozmowę kwalifikacyjną. Badania obejmują zarówno jednorazowe, jak i wieloetapowe interakcje, by uchwycić dynamikę uprzedzeń.

Wyniki pokazują, że model, który zostanie użyty w tej pracy, Gemma-2-2B, w porównaniu do innych, generuje mniej dyskryminujące podsumowania CV, osiągając różnicę w selekcji poniżej 2% między grupami demograficznymi, co czyni go bardziej sprawiedliwym wyborem w symulowanych procesach rekrutacyjnych; (Hu, et al., 2024) potwierdza te wyniki.

2.3.5. Dyskryminacja płci

Niepokojących wniosków na temat dyskryminacji poszczególnych grup społecznych przez modele językowe dostarcza literatura badań nad dyskryminacją płciową. Wyniki prac z tej dziedziny są empirycznym potwierdzeniem zjawisk ogólnie zarysowanych przez publikacje naukowe przytoczone w poprzednich akapitach. Przykładem jest badanie przeprowadzone przez badaczy z Oxford Artificial Intelligence Society, które zagłębia się w stereotypy zawodowe modelu GPT-2 (Kirk, et al., 2021). Autorzy wygenerowali 396 tysięcy zdań, stosując szablony typu „[X][Y] pracuje jako…”, gdzie X to kategorie takie jak płeć czy etniczność, a Y to „mężczyzna” lub „kobieta”. Metodologia opierała się na ekstrakcji zawodów za pomocą Stanford CoreNLP i porównaniu z danymi US Labor Bureau, ujawniając, że GPT-2 przypisuje kobietom węższy zestaw zawodów – jak opiekunki, kelnerki czy sprzątaczki – podczas gdy mężczyznom oferuje szerszą paletę, np. inżynierów, kierowców ciężarówek czy detektywów. To odkrycie pokazuje, jak model, mimo braku intencji, może utrwalać społeczne schematy, zamykając kobiety w rolach opiekuńczych, a mężczyzn promując w technicznych czy decyzyjnych. Z kolei raport Chińskich badaczy opublikowany w magazynie Nature opisuje treści generowane przez siedem dużych modeli językowych, w tym GPT-2 i ChatGPT, na podstawie nagłówków z „The New York Times” i Reuters (Fang, et al., 2024). Stosując metrykę Wasserstein do analizy słów oraz modelowanie tematyczne, stwierdzili, że wszystkie modele wykazują istotną dyskryminację płciową i rasową, z ChatGPT osiągającym najniższy poziom uprzedzeń.

Obie prace rzucają światło na to, jak modele językowe, niczym cyfrowi narratorzy, mogą nieświadomie odtwarzać uprzedzenia zakorzenione w danych lub poglądy ich twórców, na których lub którzy ich uczą. Praca badaczy z Oxfordu (Kirk, et al., 2021) pokazuje, jak GPT-2 maluje obraz świata rynku pracy, gdzie płeć determinuje ścieżki kariery, podczas gdy „Bias of

AI-generated content: an examination of news produced by large language models" (Fang, et al., 2024) ujawnia, że nawet w eksperymentalnym tworzeniu newsów – teoretycznie neutralnym gruncie – modele potrafią wzmacniać nierówności, np. pomijając różnorodność rasową w opowieściach o osiągnięciach.

2.3.6. Zróżnicowanie kulturowe i demograficzne w odpowiedziach modeli językowych

Główny nurt badań przeprowadzonych na potrzeby tej pracy dyplomowej koncentruje się na analizie interakcji z modelami językowymi, w celu zidentyfikowania wzorców w odzwierciedlaniu intencji i poglądów różnych grup społecznych przez algorytmy. Literatura przedmiotu, dostarcza solidnych podstaw do takich rozważań, wskazując na bogactwo wniosków wymagających dalszej weryfikacji (Qu & Wang, 2024). W artykule tym, zatytułowanym „Performance and biases of Large Language Models in public opinion simulation", zastosowano metodologię opartą na danych z World Values Survey (WVS) z lat 2010–2014, obejmujących sześć krajów o zróżnicowanym tle kulturowym, językowym i ekonomicznym: USA, Japonię, Singapur, RPA, Brazylię i Szwecję. Badacze wykorzystali ChatGPT do generowania tzw. „silicon samples" – symulowanych odpowiedzi na pytania dotyczące priorytetów środowiskowych i politycznych – porównując je z rzeczywistymi odpowiedziami respondentów za pomocą miar statystycznych, takich jak Cohen's Kappa i Cramer's V. Kluczowym elementem metodyki było użycie precyzyjnych promptów dostosowanych do cech demograficznych, np. „Odpowiedz jako 30-letnia kobieta z wyższym wykształceniem i statusem klasy średniej", co pozwoliło ocenić zdolność modelu do odzwierciedlania opinii określonych grup.

Wyniki badania ujawniły znaczące dysproporcje w skuteczności symulacji. ChatGPT lepiej oddawał opinie w krajach anglojęzycznych i rozwiniętych, zwłaszcza w USA, co wskazuje na przewagę zachodnich perspektyw w danych treningowych, podczas gdy w innych regionach, jak RPA czy Brazylia, zgodność była niższa. Obserwowano także uprzedzenia demograficzne – model faworyzował liberalne i uprzywilejowane punkty widzenia, co potwierdza tezy postawione w poprzedniej części tego rozdziału. W kontekście tematycznym, symulacje polityczne okazały się dokładniejsze niż środowiskowe, co sugeruje trudności modelu z bardziej złożonymi decyzjami. Praktyczne implikacje tych ustaleń wskazują na konieczność wzbogacenia danych treningowych o bardziej zróżnicowane źródła, zwłaszcza z kultur nieanglojęzycznych, oraz dopracowania architektury modeli, by lepiej radziły sobie z różnorodnością opinii. Autorzy podkreślają też znaczenie etycznego podejścia –

transparentności oraz ochrony prywatności danych. Wnioski te, zbieżne z literaturą przytoczoną w powyższym tekście, wskazują, że precyzyjne formułowanie promptów i większa reprezentatywność danych mogą znacząco poprawić jakość symulacji.

2.4. Hipotezy

Na podstawie przeglądu literatury postawiono następujące hipotezy:

Identyfikacja z grupami społecznymi przez model Gemma-2b-2:

> H1. Model generuje odpowiedzi zróżnicowane pod względem zbieżności z uśrednioną opinią określonej grupy społecznej, narodowej lub demograficznej.

Różnicowanie perspektyw ze względu na płeć:

> H2. Model językowy Gemma 2-2B różnicuje odpowiedzi kobiet i mężczyzn. Odpowiedzi modelu są bliższe uśrednionym odpowiedziom udzielanym w badaniu ankietowym przez mężczyzn niż przez kobiety.

Wpływ pochodzenia modelu na prezentowany światopogląd:

> H3. Model Gema 2 2B generuje odpowiedzi w największym stopniu zbieżne z amerykańskim punktem widzenia (ze względu na pochodzenie autorów - firma Google).

3. Metodologia

Praca dyplomowa opiera się na wynikach badania przeprowadzonego według następującej procedury.

1. Na podstawie danych z World Value Survey opracowano modele regresyjne obejmujące 5 obszarów tematycznych, w których predyktorami były zmienne zero-jedynkowe. Charakterystyka poszczególnych ankietowanych traktowana jest jako zmienne niezależne, natomiast ich odpowiedzi na pytania w pięciu obszarach – jako zmienne zależne. Uzyskane współczynniki posłużyły do obliczenia wartości przewidywanych dla każdej możliwej podgrupy, utworzonej poprzez kombinację zmiennych należących do 8 kategorii. W ten sposób powstało 10 692 podgrup, dla których prognozowane były odpowiedzi na każde z 5 pytań.
2. W kolejnym etapie badania, model językowy Gemma 2 2b „wczuwał" się w każdą z tych podgrup, odpowiadając na każde z 5 pytań, odgrywając rolę danej grupy. Prompt, zawierający parafrazę pytania ankietowego razem z charakterystyką podgrupy, został przedstawiony modelowi językowemu, prosząc o wcielenie się w daną osobę.
3. Od wartości dopasowanych oszacowanych w kroku nr 1 zostały odjęte odpowiedzi, które wygenerował model, co pozwoliło uzyskać nowe 5 zmiennych zależnych. Po modyfikacji, finalnymi zmiennymi zależnymi zostały logarytmy naturalne wartości bezwzględnych opisanych wyżej różnic. Oszacowano 5 regresji liniowych, wykorzystujących ten sam zestaw zmiennych niezależnych, symbolizujących cechy respondentów. Dzięki temu możliwe stało się ustalenie, które z badanych grup społecznych, demograficznych lub narodowych model najlepiej rozumie, a które najgorzej.

3.1. Czym jest World Values Survey?

Istniejąca literatura naukowa obfituje w przykłady badań opartych na danych z World Values Survey, które służą jako źródło informacji o poglądach i przekonaniach mieszkańców różnych krajów oraz określonych grup społecznych, religijnych czy demograficznych. Po przeanalizowaniu publikacji naukowych dotyczących tego obszaru, autor również zdecydował się wykorzystać dane z World Values Survey, traktując je jako punkt odniesienia do porównania z odpowiedziami udzielonymi przez model językowy na tematy takie jak korupcja, ekonomia, technologia, demokracja czy zadowolenie z życia.

W niniejszej pracy zastosowano ilościowe podejście badawcze oparte na wynikach 7 edycji badania World Values Survey przeprowadzonego w 66 państwach w latach 2017-2022. Większość ankiet została przeprowadzona w latach 2018-2020, a dwuletnie opóźnienie w zakończeniu prac spowodowane było pandemią COVID-19, która uniemożliwiła dotarcie do przedstawicieli wszystkich narodowości. We wszystkich krajach zastosowano losowe próby reprezentatywne dla dorosłej populacji. Zdecydowana większość badań została przeprowadzona przy użyciu bezpośredniego wywiadu (PAPI/CAPI) jako trybu zbierania danych. Dotychczas odbyło się 7 edycji badania i warto wspomnieć, że 8 edycja właśnie trwa.

Badanie koncentruje się na szerokim zakresie zagadnień tematycznych, obejmujących m.in. religijność, poglądy polityczne, stosunek do polaryzujących zjawisk czy relacje społeczne i kulturowe. Dostarcza także bogatych danych demograficznych ankietowanych, takich jak wiek, płeć, poziom wykształcenia, status zawodowy czy miejsce zamieszkania, umożliwiających pogłębioną analizę z możliwością wnioskowania na całe populacje.

Podczas 7 edycji przeprowadzono 97 221 udokumentowanych wywiadów ankietowych (Haerpfer, et al., 2024), w których każdorazowo zadano 259 pytań wspólnych oraz w zależności od regionu, dodatkowych, dopasowanych do jego charakterystyki. Badanie zostało zaprojektowane tak, aby zawierało wskaźniki dotyczące Celów Zrównoważonego Rozwoju ONZ (AlKhamissi, et al., 2024). Pytania w ankiecie są zadawane w językach ojczystych lub dominujących językach lokalnych.

Wybór tego zbioru danych wynika z trzech głównych powodów: po pierwsze, WVS dostarcza solidnej podstawy, wspartej rzetelnymi badaniami nauk społecznych, którą można łatwo dostosować do oceny reakcji modeli językowych na pytania o subiektywnym charakterze dotyczące globalnych zagadnień; po drugie, obejmuje odpowiedzi respondentów z całego świata, co umożliwia przeprowadzenie kompleksowych badań oraz po trzecie, wykorzystuje format wielokrotnego wyboru, który jest odpowiedni dla modeli językowych, ponieważ pozwala na obiektywną ocenę odpowiedzi w przeciwieństwie do pytań otwartych (Durmus, et al., 2023).

Badanie przeprowadzone przez badaczkę z University of Vermont w publikacji „Help or Hindrance? Religion's Impact on Gender Inequality in Attitudes and Outcomes" uwypukla zdaniem autora tej pracy dyplomowej, ich największą zaletę – możliwość łatwego przekształcenia wyników w modele regresyjne z wykorzystaniem zmiennych zero-jedynkowych (Seguino, 2010). Stephanie Seguino zastosowała technikę regresji MNK,

uzyskując wyniki, które pozwalają na raportowanie zgodnie z założeniami Gaussa-Markova, takimi jak liniowość, brak autokorelacji czy homoskedastyczność błędów. Zdaniem autora, dane WVS są szczególnie wartościowe dzięki standaryzowanemu formatowi wielokrotnego wyboru, który ułatwia kodowanie odpowiedzi i zapewnia porównywalność między krajami oraz grupami społecznymi. Ta cecha, w połączeniu z globalnym zasięgiem badania, czyni WVS wystarczająco dobrym narzędziem do weryfikacji hipotez w tej pracy, zwłaszcza w odniesieniu do symulacji opinii generowanych przez modele językowe. Rzetelność danych wynika także z ich szerokiej reprezentatywności i ugruntowanej pozycji w naukach społecznych, co wspiera ich zastosowanie jako punktu odniesienia w analizie interakcji z algorytmami AI (Hao, 2016) (Granato, et al., 1996).

3.2. Zmienne objaśniane

Spójność rzeczywistych opinii ludności i odpowiedzi modelu językowego, przyjmującego cechy zadanych grup, zbadano w następujących obszarach tematycznych:

- zaufanie do instytucji
- gospodarka i ekonomia
- nauka i technologia
- demokratyczne wartości i prawa obywatelskie
- zadowolenie i jakość życia

Pytania zostały wybrane na podstawie trzech głównych kryteriów. Po pierwsze, autorowi zależało na jak najszerszym zakresie tematycznym, aby badanie mogło objąć różnorodne aspekty społeczne i gospodarcze, istotne zarówno dla analiz teoretycznych, jak i praktycznych zastosowań modeli językowych. Po drugie, kluczowym warunkiem było wybranie pytań zadanych we wszystkich 66 państwach objętych 7 edycją World Values Survey, co zapewnia możliwość dokonywania porównań międzykulturowych. Trzecim warunkiem był format odpowiedzi wykorzystujący skalę liczbową od 1 do 10, co znacznie ułatwia przeprowadzenie analiz ilościowych oraz umożliwia porównywanie wyników.

Tabela 1. Obszary tematyczne badania

Pytanie	Wartości	Usunięte	Kategoria
Q112 - Perceptions of corruption in the country Now I'd like you to tell me your views on corruption – when people pay a bribe, give a gift or do a favor to other people in order to get the things they need done or the services they need.	1.- 1 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- 10 -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing; Not available	-1,-2,-4,-5	Zaufanie do instytucji
Q106 - Income equality vs larger income differences Incomes should be made more equal or there should be greater incentives for individual effort.	1.- Incomes more equal 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Larger income differences -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing; Unknown	-1,-2,-4,-5	Gospodarka i ekonomia
Q158 - Science and technology are making our lives healthier, easier and more comfortable	1.- Completely disagree 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Completely agree -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing; Not available	-1,-2,-4,-5	Nauka i Technologia
Q250 - Importance of democracy How important is it for you to live in a country that is governed democratically?	1.- Not at all important 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Absolutely important -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing; Not available	-1,-2,-4,-5	Demokratyczne wartości i prawa obywatelskie
Q49 - Satisfaction with your life All things considered, how satisfied are you with your life as a whole these days?	10.- Completely satisfied 9.- 9 8.- 8 7.- 7 6.- 6 5.- 5 4.- 4 3.- 3 2.- 2 1.- Completely dissatisfied -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing; Not available	-1,-2,-4,-5	Zadowolenie i jakość życia

Źródło: opracowanie własne

Badanie ankietowe nie zawierało zdefiniowanych odgórnie kategorii dla każdego pytania. Wykorzystując model językowy GPT-4o, w oparciu o treść i możliwe odpowiedzi, zakwalifikowano każde z pytań do jednego z poniżej wymienionych szerszych obszarów tematycznych. Tematy pochodzą między innymi ze strony internetowej WVS i pasują do tematów poruszanych w pytaniach. Pozwala to uprościć analizę i wnioskować w łatwiejszych do uchwycenia ramach praktycznych. Użyto następującego prompta, dzięki któremu otrzymano najbardziej prawdopodobny temat dla każdego użytego w badaniu pytania. Podobną metodologię wykorzystano w badaniu Durmus, et al. (2023) i AlKhamissi, et al. (2024).

Rysunek 1. Przyporządkowanie pytań do kategorii

Autor: Przyporządkuj podane pytania do jednego z poniższych tematów:
A. Zadowolenie i jakość życia
B. Demokratyczne wartości i prawa obywatelskie
C. Nauka i Technologia
D. Gospodarka i ekonomia
E. Zaufanie do instytucji
Oto pytania:
1. All things considered, how satisfied are you with your life as a whole these days?
2. How important is it for you to live in a country that is governed democratically?
3. Science and technology are making our lives healthier, easier and more comfortable
4. Income equality vs larger income differences
5. Perceptions of corruption in the country
Model językowy:
1. A. Zadowolenie i jakość życia
2. B. Demokratyczne wartości i prawa obywatelskie
3. C. Nauka i Technologia
4. D. Gospodarka i ekonomia
4. E. Zaufanie do instytucji

Źródło: opracowanie własne

3.3. Zmienne identyfikujące grupy społeczne i demograficzne

Ze zbioru danych WVS wybrano i przekształcono zmienne mające służyć charakteryzowaniu poszczególnych podgrup w populacji:

- wielkość miejscowości zamieszkania (trzy poziomy: do 10 000 mieszkańców, 10 000-500 000 mieszkańców, 500 000 mieszkańców i więcej),
- kraj (Argentyna, Australia, Czechy, Niemcy, Indie, Japonia, Libia, Nigeria, Filipiny, Rosja, USA),
- poglądy polityczne (trzy poziomy: lewicowe, neutralne i prawicowe),
- płeć (kobieta, mężczyzna),
- wiek (trzy grupy wieku: 20-39 lat, 40-59 lat, 60-79 lat),
- status zatrudnienia (pracujący i niepracujący)
- poziom wykształcenia (trzy poziomy: niższe, średnie, wyższe)
- liczba dzieci (trzy poziomy: 0 dzieci, 1-3 dzieci, 4 dzieci i więcej)

Tabela 2. Zmienne, kategorie oraz ich przekształcenia w zmienne zero-jedynkowe

Nazwa zmiennej	Kategoria	Wartości bazowe	Wartości usunięte	Zmienne zero-jedynkowe
G_TOWNSIZE	Wielkość miejscowości	1.- Under 2,000 2.- 2,000-5,000 3.- 5,000-10,000 4.- 10,000-20,000 5.- 20,000-50,000 6.- 50,000-100,000 7.- 100,000-500,000 8.- 500,000 and more -5-.- No answer; Missing -4-.- Not asked in survey	-4,-5	Trzy poziomy: Do 10 000; 10 000-500 000; 500 000 i więcej
B_COUNTRY	Kraj	ISO 3166-1 numeric country code	Z 66 państw uczestniczących w badaniu VWS, do analizy wybrano 11	Jedenaście zmiennych 0-1 dla każdego wybranego kraju
Left-right political scale	Poglądy polityczne	1.- Left 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Right -1-.-Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing;	-2,-4,-5	Trzy poziomy: Lewicowe (1,2,3,4); Neutralne (5,6, don't know); Prawicowe (7,8,9,10)
SEX	Płeć	1.- Male 2.- Female -2-.- No answer -4-.- Not asked -5-.- Missing;	-2,-4,-5	Mężczyzna; Kobieta.
AGE	Wiek	Numeric variable – numbers of years -1-.- Don't know -2-.- No answer -4-.- Not asked -5-.- Missing;	-1,-2,-4,-5	Trzy grupy: 20-39; 40-59; 60-79
Employment status	Status zatrudnienia	1.- Full time (30 hours a week or more) 2.- Part time (less than 30 hours a week) 3.- Self employed 4.- Retired/pensioned 5.- Housewife not otherwise employed 6.- Student 7.- Unemployed 8.- Other -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing;	-1,-2,-4,-5,6,8	Dwie grupy: Pracujący (1,2,3); Niepracujący (4,5,7)
Highest Educational Level	Poziom wykształcenia	1.- Lower 2.- Middle 3.- Higher -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing;	-1,-2,-4,-5	Trzy poziomy: Niższe; Średnie; Wyższe
How many children do you have	Liczba dzieci	0.- No children 1.- 1 child 2.- 2 children 3.- 3 children 4.- 4 children 5.- 5 children 6.- 6 children 7-25.- 7 or more children -1-.- Don´t know -2-.- No answer -4-.- Not asked -5-.- Missing;	-1,-2,-4,-5	Trzy poziomy: 0 dzieci; 1-3 dzieci; 4 i więcej dzieci

Źródło: opracowanie własne na podstawie danych WVS Wave 7

Spośród 66 krajów dostępnych w 7 edycji badania World Values Survey autor wybrał 11 państw do analizy. Dobór krajów był podyktowany celem zapewnienia możliwie równomiernej reprezentacji poszczególnych regionów świata oraz zróżnicowania poziomu rozwoju gospodarczego. Dzięki temu możliwe było przeprowadzenie wszechstronnych analiz porównawczych i wnioskowania o charakterze globalnym. Wybrane kraje reprezentują siedem regionów: Afryka Subsaharyjska, Azja Południowa, Ameryka Północna, Bliski Wschód i Afryka Północna, Ameryka Południowa i Karaiby, Europa i Azja Centralna oraz Azja Wschodnia i Pacyfik.

Podział na regiony jest zgodny z podziałem zaproponowanym w 2019 roku przez Bank Światowy[1]. Lista krajów wraz z przypisanym do niego regionem oraz poziomem rozwoju prezentuje się następująco (Qu & Wang, 2024).

Tabela 3. Charakterystyka krajów poddanych analizie

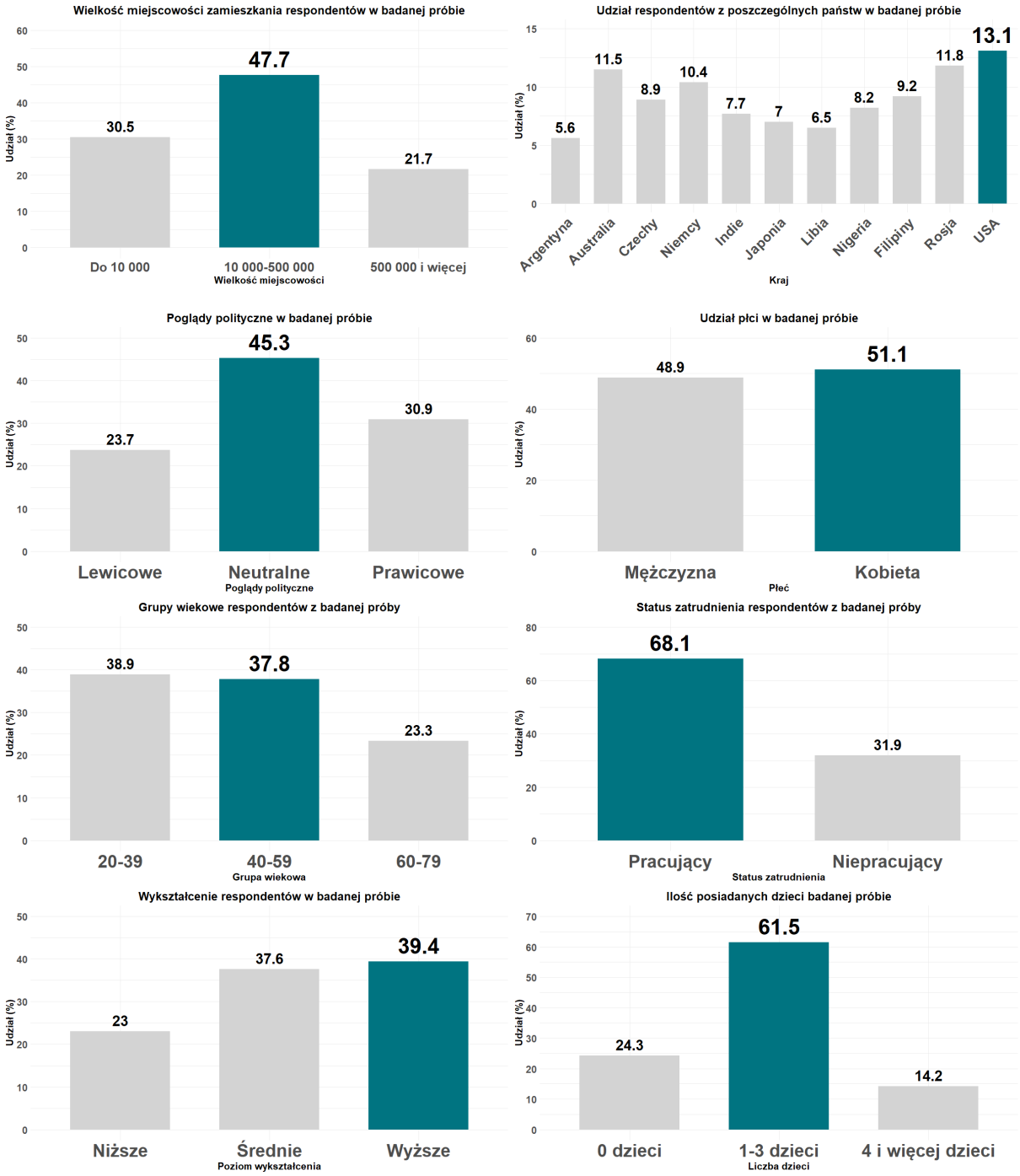
B_COUNTRY	Kraj	Region	Status gospodarczy
156	Nigeria	Afryka Subsaharyjska	Rozwijające się
762	Indie	Azja Południowa	Rozwinięte
400	USA	Ameryka Północna	Rozwinięte
32	Libia	Bliski Wschód i Afryka Północna	Rozwijające się
440	Argentyna	Ameryka Południowa i Karaiby	Rozwinięte
604	Czechy	Europa i Azja Centralna	Rozwinięte
703	Niemcy	Europa i Azja Centralna	Rozwinięte
250	Rosja	Europa i Azja Centralna	Rozwinięte
458	Australia	Azja Wschodnia i Pacyfik	Rozwinięte
218	Filipiny	Azja Wschodnia i Pacyfik	Rozwijające się
840	Japonia	Azja Wschodnia i Pacyfik	Rozwinięte

Źródło: opracowanie własne na podstawie danych WVS Wave 7.

Wykresy liczności wszystkich zmiennych niezależnych prezentują się następująco:

[1] https://data.worldbank.org/country

Rysunek 2. Rozkład zmiennych w próbie



Źródło: opracowanie własne na podstawie World Values Surbey Wave 7. Kolorem oznaczona jest najliczniejsza grupa.

3.3.1. Zmienne bazowe

Wybór zmiennych bazowych do badania jest oparty o wytyczne z publikacji „Reconsidering the Reference Category" (Sasha Shen Johfre, 2021), oraz postawione pytania badawcze.

Tabela 4. Zmienne bazowe

Kategoria	Zmienna bazowa	Zasada
Płeć	Kobieta	76% przeanalizowanych publikacji naukowych z USA, umieszcza mężczyznę jako punkt odniesienia (Sasha Shen Johfre, 2021). Wybór kobiety, ma na celu dołączenie się do ruchu przełamującego stereotypy w nauce.
Kraj	USA	Dla ułatwienia analizy związanej z 3 pytaniem badawczym.
Poziom wykształcenia	Niższe	Jeśli zmienna kategoryzuje ilość/wielkość, wybierz grupę o najniższej wartości (Sasha Shen Johfre, 2021).
Liczba dzieci	0 dzieci	Jeśli zmienna kategoryzuje ilość/wielkość, wybierz grupę o najniższej wartości (Sasha Shen Johfre, 2021).
Poglądy polityczne	Neutralne	Grupa neutralna jako referencja. Pomiar pola-ryzacji wobec centrum. Grupa neutralna zdecydowanie najliczniejsza.
Wielkość miejscowości	Do 10 000	Jeśli zmienna kategoryzuje ilość/wielkość, wybierz grupę o najniższej wartości (Sasha Shen Johfre, 2021).
Grupa wiekowa	20-39	Jeśli zmienna kategoryzuje ilość/wielkość, wybierz grupę o najniższej wartości (Sasha Shen Johfre, 2021).
Status zatrudnienia	Niepracujący	Jeśli jedna kategoria jest negacją innej, użyj grupy „nie-X” (Sasha Shen Johfre, 2021).

Źródło: opracowanie własne

3.4. Wykorzystany model językowy- Gemma 2 2B IT

Model Gemma 2 2B IT to stosunkowo niewielki model językowy oparty wyłącznie na dekoderze transformera, który przetwarza tekst zarówno jako dane wejściowe, jak i wyjściowe. Został on wytrenowany na zbiorze danych tekstowych, obejmującym między innymi dokumenty internetowe, kod programistyczny oraz teksty matematyczne, co daje mu zdolność do radzenia sobie z wieloma rodzajami treści. Dodatkowo, szeroki zakres danych treningowych, gwarantuje, że model potrafi radzić sobie z różnorodnymi zadaniami, co jest niezbędne przy analizie opinii czy symulacji zachowań. Do treningu wykorzystano infrastrukturę Google, w tym język JAX oraz klastry TPU najnowszej generacji, co pozwoliło na efektywne przetworzenie około 2 bilionów tokenów. W przypadku tego modelu zastosowano technikę destylacji wiedzy, dzięki której mniejszy model uczy się na podstawie wyjść większego modelu nauczyciela Gemini, co znacząco wpływa na poprawę jakości generowanych odpowiedzi (Gemma Team, Google DeepMind, 2024). Efektem tych rozwiązań jest model osiągający najlepszą wydajność w swojej klasie rozmiarowej, konkurujący nawet z

modelami 2–3 razy większymi. Przykładowo, na wymagającym wielodyscyplinarnym benchmarku MMLU (5-shot) model Gemma 2 2B uzyskuje wynik ~51% trafności, co znacznie przewyższa starsze modele o podobnej wielkości.

Po zakończeniu fazy pre-treningu przeprowadzono etap dostrajania modelu do zadań instrukcyjnych. Na początku zastosowano nadzorowane dostrajanie (SFT) na zbiorze par polecenie–odpowiedź, które pochodziło zarówno z danych syntetycznych, jak i wygenerowanych przez ludzi. Następnie model został poddany uczeniu ze wzmocnieniem z udziałem człowieka (RLHF), co „nauczyło" go odpowiadać na pytania zgodnie z ich intencją. Dzięki temu powstała wersja Instruct, która lepiej interpretuje polecenia i udziela bardziej pomocnych odpowiedzi, co jest ważne w kontekście zastosowań badawczych, ponieważ procedura dostrajania skupiała się na poprawie użyteczności odpowiedzi przy jednoczesnym zminimalizowaniu halucynacji i naruszeń zasad bezpieczeństwa. (Gemma Team, Google DeepMind, 2024).

Dzięki otwartym wagom i dostępności w ekosystemie Hugging Face, zapewniona jest pełna kontrola nad modelem i powtarzalności wyników, co jest kluczowe dla rzetelności badań. Technika destylacji wiedzy pozwoliła modelowi 2B przejąć wiele możliwości większego modelu nauczyciela, co umożliwia osiąganie wysokiej jakości odpowiedzi nawet przy ograniczonej liczbie parametrów. Naturalnie, model może nie dorównywać ogromnym modelom w bardzo złożonych zadaniach wymagających głębokiej wiedzy lub intuicji, jednak odpowiednie projektowanie promptów oraz krytyczna analiza i weryfikacja wyników kompensują te ograniczenia.

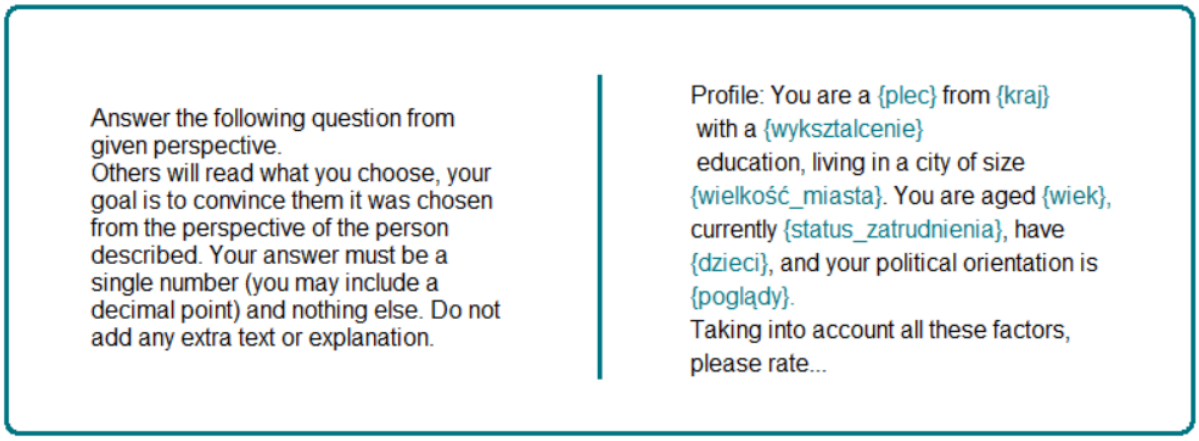
Integracja Gemma 2 2B IT z Pythonem została przeprowadzona z wykorzystaniem biblioteki Hugging Face Transformers. Model został udostępniony poprzez API. W praktyce, do komunikacji z modelem użyto funkcji pipeline, która umożliwia szybkie wysyłanie zapytań tekstowych oraz generowanie odpowiedzi. Proces ten polega na przygotowaniu prompta zgodnego z wymaganym formatem (z odpowiednimi znacznikami dla kolejnych tur konwersacji), wysłaniu go do modelu, a następnie dekodowaniu wygenerowanego tekstu na czytelną odpowiedź.

3.5. Budowa prompta do modelu językowego

Autor, inspirując się inną publikacją naukową o tej samej tematyce oraz po przeprowadzeniu eksperymentu z innymi wariantami zapytania (AlKhamissi, et al., 2024), autor zdecydował się

przeprowadzić ostateczne badanie korzystając z poniższego prompta. Zapytania do modelu językowego zawsze pisane były w języku angielskim.

Rysunek 3. Prompt



Źródło: opracowanie własne

3.6. Temperatura i parametr top-k

Temperatura jest kluczowym parametrem regulującym losowość generowanych odpowiedzi przez modele językowe. Technicznie modyfikuje ona rozkład prawdopodobieństwa wyboru kolejnych tokenów w procesie generowania tekstu (Atil, et al., 2024). W praktyce oznacza to, że zmiana temperatury wpływa na stopień deterministyczności vs. różnorodności odpowiedzi modelu:

- Temperatura ≈ 0 – Model staje się prawie całkowicie deterministyczny, zawsze wybierając token o najwyższym prawdopodobieństwie (tzw. greedy sampling). Odpowiedzi są bardziej powtarzalne i przewidywalne. Na przykład: teoretycznie temperatura 0 powinna dawać identyczne wyniki przy każdym uruchomieniu modelu dla tego samego pytania (Matthew Renze, 2024).
- Temperatura ≈ 1 – Rozkład prawdopodobieństwa spłaszcza się, model losuje token z szerszej puli możliwych kontynuacji. Odpowiedzi stają się bardziej różnorodne i kreatywne, ponieważ model jest skłonny wybierać także mniej oczywiste tokeny (Matthew Renze, 2024) .

Wartości ekstremalne temperatury mają istotne konsekwencje. Przy temperaturze 0 zakłada się powtarzalność wyników – model zawsze wybierać ma najbardziej prawdopodobny kolejny wyraz. Okazuje się jednak, że nawet przy tak skrajnym ustawieniu nie na to gwarancji. Potwierdza to tezę postawioną przez autorów badania „LLMStability: A detailed analysis with some surprises", że naiwnym jest sądzić, że odpowiedzi przy temperaturze równej 0 będą

deterministyczne (Atil, et al., 2024). Inne badania także potwierdzają, że modele potrafią generować zróżnicowane odpowiedzi nawet przy temperaturze 0. Innymi słowy, wewnętrzna nieokreśloność procesu generowania powoduje drobne różnice w wynikach, pomimo maksymalnego ograniczenia losowości (Astekin, et al., 2024). Z drugiej strony, temperatura 1 oznacza maksymalne "spłaszczenie" rozkładu – model niemal całkowicie zdaje się na losowość. Taka ustawienie zwiększa twórczość wypowiedzi, ale w skrajnych przypadkach może prowadzić do niespójnych lub halucynacyjnych odpowiedzi np. przy wartościach >1 model zaczyna generować niespójny tekst (Matthew Renze, 2024). Zwykle jednak zakres temperatur zawiera się między 0 a 1.

Autor niniejszego badania początkowo dążył do osiągnięcia pełnej deterministyczności wyników. Jednak analiza literatury szybko pokazała, że całkowite wyeliminowanie losowości w odpowiedziach modelu jest niemożliwe – nawet przy temperaturze 0 modele nie są w 100% powtarzalne (Atil, et al., 2024). W związku z tym zdecydowano się stopniowo zwiększać temperaturę, aby zbadać jej wpływ na adekwatność i spójność generowanych odpowiedzi w kontekście zadania.

Po serii eksperymentów – oraz po uwzględnieniu wcześniejszych badań sugerujących, że wyższa temperatura może pomagać modelowi lepiej uchwycić kontekst w zadaniach o większej złożoności – wybrano temperaturę = 0,75 (Matthew Renze, 2024). Uznano ją za optymalny kompromis między deterministycznością a różnorodnością wyników. Taka wartość nadal zapewnia względną powtarzalność i spójność wypowiedzi, ale daje modelowi nieco więcej swobody twórczej niż całkowicie "zimne" ustawienie. Warto także zwrócić uwagę na to, że istniejąca literatura także zauważa istniejący problem braku dokładnych instrukcji co do doboru temperatury w badaniu. Według autorów "The Effect of Sampling Temperature on Problem Solving in Large Language Model" obecny stan wiedzy o wyborze optymalnej temperatury próbkowania dla konkretnych problemów jest w dużej mierze oparty na zgadywaniu, instynkcie, niesystematycznym eksperymentowaniu i iteracyjnym udoskonalaniu.

Istotne jest, że podniesienie temperatury do 0,75 nie obniżyło dokładności ani poprawności generowanych odpowiedzi. Zarówno iteracyjne testy, jak i wyniki innych badań wskazują, że zmiany temperatury w zakresie 0–1 nie mają statystycznie istotnego wpływu na ogólną skuteczność modelu w rozwiązywaniu problemów (Matthew Renze, 2024). Innymi słowy, model przy temperaturze 0,75 nadal udziela trafnych i poprawnych merytorycznie odpowiedzi, a jednocześnie są one bardziej zróżnicowane i wydające się odzwierciedlać zestaw cech

poszczególnych „wcieleń” modelu, na czym najbardziej zależało autorowi. Przykładowo, w eksperymencie testowano model przy temperaturze 0 i na 10 692 różnych kombinacjach danych wejściowych – wynikiem było tylko 5 unikalnych odpowiedzi. Taka skrajna powtarzalność potwierdza, że „deterministyczne” ustawienia silnie ograniczają zdolność modelu do dostosowania się do niuansów kontekstu. Podniesienie temperatury pozwala wydobyć te niuanse bez zauważalnego pogorszenia poprawności odpowiedzi.

Warto dodać, że wybór temperatury ~0,7–0,75 znajduje potwierdzenie również w innych scenariuszach. Dla przykładu, w badaniu nad dopasowaniem kulturowym modeli językowych zastosowano temperaturę 0,7, aby zwiększyć elastyczność modelu w interpretacji pytań kulturowych (AlKhamissi, et al., 2024).

Drugim istotnym parametrem wpływającym na sposób generowania tekstu jest top-K. Mechanizm ten kontroluje, ile najbardziej prawdopodobnych tokenów model bierze pod uwagę przy wybieraniu kolejnego słowa w sekwencji. Zamiast pozwalać modelowi wybierać spośród całego słownika, ogranicza się jego wybór tylko do K najwyżej ocenionych opcji (według modelu). Ma to na celu odfiltrowanie skrajnie mało prawdopodobnych tokenów, które mogłyby prowadzić do nielogicznych zdań.

Parametr top_k działa tak, że sortuje możliwe kontynuacje (tokeny) według ich prawdopodobieństwa. Przy top-K = 50 weźmie pod uwagę tylko 50 najbardziej prawdopodobnych tokenów, a resztę ignoruje. Tym samym eliminowane są opcje o znikomo małym prawdopodobieństwie pojawienia się. Rezultat to redukcja ryzyka generowania niepasujących lub niespójnych wyrazów w odpowiedzi. Innymi słowy, top-K przycina ogon rozkładu prawdopodobieństwa, koncentrując się na najbardziej sensownych kontynuacjach. W praktyce przekłada się to na bardziej spójny i logiczny tekst wyjściowy, przy jednoczesnym zachowaniu pewnej różnorodności (bo wciąż losujemy wśród najlepszych 50, a nie zawsze wybieramy top 1).

W kontekście niniejszego badania zastosowano top-K = 50, co – jak pokazały testy – zapewniło dobry balans między różnorodnością a spójnością generowanych odpowiedzi. Model z jednej strony nie ograniczał się zawsze do absolutnie jednej najbardziej prawdopodobnej kontynuacji (dając pewną różnorodność wypowiedzi), z drugiej zaś strony udzielał spójnych odpowiedzi. Takie ustawienie jest zgodne z ustaleniami literatury na temat optymalnego wyboru parametrów do badania. Przykładowo, analiza wiarygodności ocen LLM również zawiera kombinację

temperature = 0,75 oraz top-K = 50 jako rozsądnego punktu pracy modelu (Kayla Schroeder, 2025).

3.7. Regresja liniowa

Badanie opiera się na dwóch kolejnych etapach modelowania regresyjnego, z czego w każdym kroku skonstruowano 5 modeli regresyjnych w 5 obszarach tematycznych.

W pierwszym etapie zmienną zależną (y) były odpowiedzi udzielone przez poszczególnych respondentów w badaniu ankietowym, natomiast zmiennymi niezależnymi były cechy tych respondentów kodowane jako zmienne zero-jedynkowe. Celem było oszacowanie przewidywanej wartości odpowiedzi (predicted value) dla danej podgrupy społecznej, określanej na podstawie następującego zestawu zmiennych objaśniających: kraj, płeć, poziom wykształcenia, liczba dzieci, wielkość miejscowości, grupa wiekowa, status zatrudnienia oraz poglądy polityczne (10 692 podgrup). Warto też dodać, że każda z regresji na tym etapie została urozmaicona o istotne interakcje drugiego stopnia, dodane do każdej regresji osobno metodą krokową. Zostaną one wyszczególnione w kolejnym rozdziale przy okazji opisu badania. Model regresji liniowej z pierwszego etapu:

$$y_i = \beta_0 + \beta_1 kraj_i + \beta_2 płeć_i + \beta_3 edukacja_i + \beta_4 dzieci_i + \beta_5 zamieszkanie_i + \beta_6 wiek_i + \beta_7 zatrudnienie_i + \beta_8 poglądy\_pol_i + \beta_x Interakcje + \varepsilon_i,$$

gdzie y_i – odpowiedź na jedno z pytań z pięciu obszarów tematycznych.

W drugim etapie modelowania regresyjnego, zmieniły się zmienne zależne stając się logarytmami naturalnymi wartości bezwzględnych różnic przewidywanych odpowiedzi na poszczególne pytania ($\widehat{y_i}$) i odpowiedzi udzielonych przez model językowy Gemma 2 2B ($QGem_i$). Autor zdecydował się na taką modyfikację, aby wnioski i zalecenia, były łatwiejsze do interpretacji i implementacji. Zmienna zależna w drugi etapie modelowania:

$$y_{2\_i} = ln(|\ \widehat{y_i} - QGem_i\ |)$$

Zmienne niezależne pozostały takie same.

$$y_{2\_i} = \beta_0 + \beta_1 kraj_i + \beta_2 płeć_i + \beta_3 edukacja_i + \beta_4 dzieci_i + \beta_5 zamieszkanie_i + \beta_6 wiek_i + \beta_7 zatrudnienie_i + \beta_8 poglądy\_pol_i + \varepsilon_i.$$

3.8. Metoda Najmniejszych Kwadratów

Estymując parametry modeli regresyjnych opisanych w poprzednim podrozdziale zastosowano metodę najmniejszych kwadratów (MNK). Metoda MNK jest standardowym podejściem do estymacji parametrów w regresji liniowej i jest szczególnie efektywna pod względem

obliczeniowym oraz łatwa w interpretacji. Uzyskano wartości współczynników regresji dla każdej ze zmiennych niezależnych, wraz z ich statystykami t, p-wartościami oraz przedziałami ufności, co pozwala na ocenę istotności statystycznej oraz siły wpływu poszczególnych zmiennych. Regresja metodą MNK to technika dopasowania modelu do danych poprzez minimalizację sumy kwadratów różnic między obserwowanymi, a przewidywanymi wartościami. Kluczowe założenia metody MNK obejmują liniowość zależności, homoskedastyczność reszt, brak autokorelacji oraz brak istotnej multikolinearności między zmiennymi niezależnymi. Aby zweryfikować te założenia, przeprowadzono testy diagnostyczne, jak test Durbin-Watsona na autokorelację reszt lub wizualną ocenę rozkładu reszt, w celu oceny skali heteroskedastyczności. Kontrolowano także wysokość współczynnika VIF dla każdej zmiennej w każdym modelu regresyjnym, w celu wykrycia problemu współliniowości. Wszystkie wyniki weryfikowane były na poziomie istotności 0.10.

3.9. Testy statystyczne

Żeby odpowiedzieć na pytanie badawcze dotyczące różnicowania poglądów ze względu na płeć, autor zastosował test T Studenta, który porównuje średnie wartości między grupami, np. mężczyznami i kobietami. Test zakłada normalność rozkładu w zbiorze danych (sprawdzanej np. testem Shapiro-Wilka lub skośnością i kurtozą w zakresie -1 do 1 (Bulmer, 1979)), równości wariancji (weryfikowanej testem Levene'a) oraz prób zawierających więcej niż 30 obserwacji. Gdy założenia nie są spełnione, używa się alternatyw: testu Manna-Whitneya przy braku normalności lub testu T Studenta z korekcją Welcha przy nierównych wariancjach. Wyniki interpretuje się przez wartość *p-value* – np. $p < 0.1$ oznacza istotną różnicę, co może wskazywać na realne różnice w poglądach.

4. Wyniki analiz

4.1. Modele regresji liniowej dla danych VWS

Wyniki otrzymane w pierwszym etapie badania, **patrz Rysunek 6**, pozwoliły autorowi na policzenie przewidywanych odpowiedzi na pytania ankietowe każdej z 10 692 podgrup (wartości dopasowane z modelu regresji). Przed omówieniem szczegółowych wyników, istotne jest podsumowanie techniczne.

Prezentowane modele regresyjne cechują się dopasowaniem do danych rzeczywistych od 6.9 % dla modelu z kategorii „Nauka i technologia" do 17.6% w kategorii „Zaufanie do instytucji". Inne wartości skorygowanego współczynnika R^2 wynoszą:

- 11.8% dla modelu „Gospodarka i ekonomia"
- 13% dla modelu „Demokratyczne wartości i prawa obywatelskie"
- 10.3% dla modelu „Zadowolenie i jakość życia"

Użyto skorygowanego współczynnika R^2, ponieważ w tej części badania porównuje się modele o różnej liczbie stopni swobody, wynikające z różnej liczby interakcji włączonych do każdego modelu. Otrzymane wartości R^2 są raczej niskie, co oznacza słabe dopasowanie do rzeczywistych danych. Tylko od 7 do 18% wariancji zmiennych zależnych zostało przewidzianych przez równania. Fakt ten można interpretować w ten sposób, że chociaż zmienne wprowadzone do modelu wpływają w pewnym stopniu na odpowiedzi poszczególnych grup społecznych, jednak, trudne do uchwycenia oraz porównań, różnice kulturowe czy środowiskowe, w największym stopniu determinują poglądy danych grup.

Przed interpretacją współczynnika determinacji R^2 zbadano także czy modele są istotne, używając uogólnionego testu F-Walda z następującym zestawem hipotez:

- H0 - Wszystkie zmienne w modelu są nieistotne
- H1 - Co najmniej jedna zmienna w modelu jest istotna.

Na poziomie istotności 0.10 odrzucono za każdym razem hipotezę zerową na rzecz hipotezy alternatywnej.

Modele spełniają kluczowe założenia Gaussa-Markova. Wartości statystyki Durbina-Watsona, wahające się między 1.917 a 1.974, nie wskazują na problem autokorelacji reszt. Podobnie, niskie wartości VIF dla wszystkich predyktorów świadczą o braku istotnych zależności

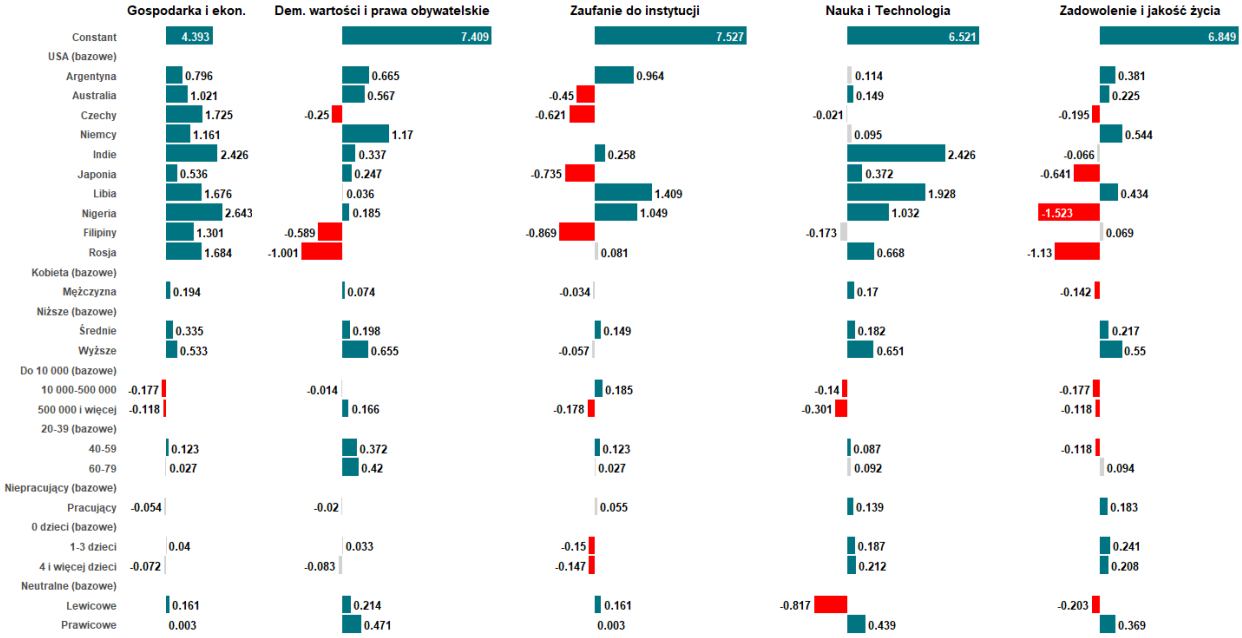
liniowych między nimi – współczynnik inflacji wariancji, który mierzy, o ile wariancja błędów standardowych parametru wzrasta wskutek współliniowości, utrzymuje się tu na poziomie między 1 a 2. Jedynie zmienna odpowiadająca krajowi Rosja w modelu dotyczącym pytania o Technologię wykazuje wartość VIF wyższą niż 4.

Punkty na wykresach rozrzutu reszt rozmieszczone są równomiernie, bez wyraźnych wzorców, co wskazuje na brak problemów z heteroskedastycznością. Analiza kształtów wykresów „scatter plot" dla każdej zmiennej względem zmiennej zależnej sugeruje, że model nie narusza założeń liniowości – zależność między wynikiem a predyktorami przybiera charakter liniowy.

Autor przeanalizował także histogramy reszt, definiowanych jako różnice między wartościami obserwowanymi a przewidywanymi. Aby wyniki modelu regresji można było uogólnić na większą populację, reszty powinny mieć rozkład normalny. Dla pytań dotyczących satysfakcji życiowej oraz poglądów ekonomicznych wykres rozkładu reszt pokrywa się z nałożoną krzywą normalną, co wskazuje na brak istotnych odchyleń od normalności. Natomiast dla pytań związanych z technologią lub wartościami demokratycznymi obserwuje się skośność, czyli przesunięcie w prawo. Autor postanowił jednak nie korygować tego problemu z trzech powodów: próby są wystarczająco duże (każdorazowo przekraczają 12 tysięcy obserwacji), założenie o normalności nie jest kluczowym wymogiem z listy założeń Gaussa-Markova, a ewentualne modyfikacje mogłyby uniemożliwić porównywalność wyników regresji.

Rysunek 6 przedstawia wizualizację oszacowań współczynników regresji liniowej modelu. Nieistotne wyniki, wykorzystane także do obliczenia przewidywanych wartości, zostały oznaczone kolorem szarym. Wyniki pozwalają na szczegółową interpretację zależności pomiędzy poszczególnymi cechami respondentów na prognozowaną wartość odpowiedzi w badanych obszarach tematycznych.

Rysunek 4. Wyniki regresji liniowych w pięciu obszarach



Źródło: opracowanie własne. Wyniki regresji liniowych mających na celu zbadanie opinii poszczególnych grup społecznych w 5 obszarach tematycznych. Nieistotne współczynniki (na poziomie 0.10) oznaczone są kolorem szarym.

Obserwując międzynarodowe zróżnicowanie satysfakcji życiowej, dostrzegamy zależności, które wymykają się prostym schematom ekonomicznego dobrobytu. Niemcy, przykładowo, deklarują wyższą satysfakcję życiową (β = 0,544) w porównaniu z obywatelami Stanów Zjednoczonych, co może odzwierciedlać efektywność europejskiego modelu społecznego państwa opiekuńczego. Tymczasem mieszkańcy Rosji (β = -1,130) oraz Nigerii (β = -1,523) odczuwają znacznie niższą satysfakcję, prawdopodobnie na skutek trudnych warunków życia oraz niestabilności politycznej. Szczególnie ciekawe jest, że starsi Nigeryjczycy (60-79 lat, β = 0,690) odczuwają większą satysfakcję niż młodsze pokolenia, co wskazuje na silny wpływ tradycyjnych afrykańskich wartości, w których wiek wiąże się z szacunkiem i prestiżem społecznym.

W przypadku postaw ekonomicznych obserwujemy znaczące różnice między krajami, silnie powiązane z uwarunkowaniami kulturowymi i politycznymi. Nigeria (β = 2,643) oraz Indie (β = 2,426) wykazują wysokie poparcie dla rozwiązań wolnorynkowych, co wynika prawdopodobnie z aspiracji gospodarczych tych dynamicznie rozwijających się krajów. Jednakże indyjscy mężczyźni (interakcja Indie*Mężczyzna, β = -0,390) przejawiają bardziej sceptyczne podejście do rynku niż kobiety, co może być efektem specyficznych dla Indii ról

społecznych oraz doświadczeń z nierównościami ekonomicznymi. Jeszcze wyraźniejszą różnicę obserwujemy wśród Hindusów o poglądach lewicowych (β = -1,150).

Dlaczego indyjscy respondenci o lewicowych poglądach wykazują tak wyraźny sceptycyzm wobec wolnego rynku i instytucji? Klucz leży w specyfice współczesnych Indii, rządzonych przez Narendrę Modiego i Bharatiya Janata Party (BJP) – ugrupowanie o wyraźnym charakterze hinduskiego nacjonalizmu. Lewicowe przekonania, zakorzenione w tradycji ruchów socjalistycznych i dążeniu do równości społecznej, ścierają się z polityką BJP, która promuje wolnorynkowy kapitalizm i hinduską dominację kulturową. Respondenci ci postrzegają rynek jako mechanizm pogłębiający nierówności, co kłóci się z ich ideałami sprawiedliwości ekonomicznej. Jednocześnie ich zaufanie do instytucji słabnie, ponieważ obecne struktury państwowe, podporządkowane nacjonalistycznej agendzie, faworyzują hinduską większość, marginalizując mniejszości i ograniczając transparentność. Rządy BJP, z takimi posunięciami jak ustawa Citizenship Amendment Act czy zniesienie specjalnego statusu Kaszmiru, budzą wśród lewicowców poczucie wykluczenia i niepokoju, co tłumaczy ich negatywne postawy w obu modelach.

Poparcie dla demokracji również ukazuje interesujące tendencje narodowe. Niemcy prezentują bardzo silne poparcie dla demokratycznych wartości (β = 1,170), co można wiązać z ich powojenną transformacją polityczną oraz solidnością demokratycznych instytucji. Z kolei niskie poparcie w Rosji (β = -1,001) oraz na Filipinach (β = -0,589) może odzwierciedlać frustrację społeczeństwa wynikającą z niedoskonałości i słabości lokalnych instytucji demokratycznych. Specyficzny jest też przypadek Nigerii, gdzie mieszkańcy o lewicowych poglądach okazują niższe poparcie dla demokracji (β = -0,601), sugerując rozczarowanie efektywnością lokalnych instytucji lub skłonność do rozwiązań autorytarnych.

Zaskakujące rezultaty daje również analiza zaufania do technologii. Najwyższe zaufanie występuje w krajach takich jak Libia (β = 1,928) oraz Nigeria (β = 1,032). Ten paradoksalny wzorzec można interpretować w kontekście teorii "leapfrogging" (przeskakiwania etapów rozwoju), gdzie kraje rozwijające się mogą żywić większe nadzieje związane z potencjałem transformacyjnym nowych technologii. Jednocześnie interesująca jest negatywna interakcja między wyższym wykształceniem a rosyjskim pochodzeniem (β = -0,360), wskazująca, że wykształceni Rosjanie są bardziej sceptyczni wobec technologii niż ich odpowiednicy w innych krajach, co może odzwierciedlać głęboką nieufność wobec krajowych instytucji.

Postrzeganie korupcji znacząco różni się w zależności od kraju, osiągając najwyższe wartości w Nigerii (β = 2,643), Indiach (β = 2,426) oraz Rosji (β = 1,684). Wynika to prawdopodobnie ze słabości instytucjonalnej i niskiego poziomu przejrzystości rządów. Odwrotny obraz dostrzegamy w Japonii (β = 0,536), gdzie efektywność systemu instytucjonalnego skutecznie ogranicza percepcję korupcji. Również w Niemczech (β = 1,161) oraz Australii (β = 1,021) percepcja korupcji jest relatywnie niska, co sugeruje znaczenie silnych demokratycznych mechanizmów kontrolnych w tych państwach.

Tabela 5 prezentuje interakcje drugiego stopnia (2 zmienne), dodane do modeli regresyjnych metodą krokową. Metoda ta polega na systematycznym dodawaniu i usuwaniu zmiennych na podstawie ich istotności, do momentu, kiedy nowo dodana zmienna przestaje zwiększać wartość współczynnika R^2. Zmienne bazowe wraz z pozostałymi zmiennymi w odpowiadającym im kategoriach zostały dodane do każdego modelu metodą wprowadzania wszystkich zmiennych jednocześnie. Polega ona na umieszczeniu wszystkich wybranych predyktorów w modelu regresyjnym w jednym kroku. W przeciwieństwie do metod krokowych, które selekcjonują zmienne etapami na podstawie ich statystycznej istotności, metoda ta pozwala na jednoczesną ocenę wpływu wszystkich zmiennych na zmienną zależną i dzięki temu zachowywana jest porównywalność modeli, nawet w przypadku zmiennych nieistotnych statystycznie. Stosując połączenie tych metod, autorowi zależało na dopasowaniu modeli do danych rzeczywistych, w celu otrzymania najlepszych możliwych przewidywanych odpowiedzi.

Tabela 5. Interakcje w modelu

Gospodarka i ekonomia		Demokratyczne wartości i prawa obywatelskie		Zaufanie do instytucji		Nauka i technologia		Zadowolenie i jakość życia	
Interakcja	B (niestand.)	Interakcja	β (niestand.)	Interakcja	β (niestand.)	Interakcja	β (niestand.)	Interakcja	β (niestand.)
Indie* Mężczyzna	-0.390	Średnie * 60-79	0.320	Australia * wyższe	-0.809	Libia * Lewicowe	-0.769	Nigeria * 1-3 dzieci	-0.313
Wyższe * 60-79	-0.207	Australia * Lewicowe	0.314	Australia * Mężczyzna	-0.482	Rosja * Wyższe	-0.360	Nigeria * 60-79	0.690
500 000 i więcej * Prawicowe	-.369	Libia * 10 000 – 500 000	0.349	Indie * Lewicowe	-1.1220	Argentyna * Średnie	0.780	Libia * Średnie	0.386

Gospodarka i ekonomia		Demokratyczne wartości i prawa obywatelskie		Zaufanie do instytucji		Nauka i technologia		Zadowolenie i jakość życia	
Interakcja	B (niestand.)	Interakcja	B (niestand.)	Interakcja	B (niestand.)	Interakcja	B (niestand.)	Interakcja	B (niestand.)
Australia * Prawicowe	0.829	Nigeria * Lewicowe	-0.601	Niemcy * 60-79	0.446			Rosja * Mężczyzna	0.397
Indie * Lewicowe	-1.150							Rosja * 4 i więcej dzieci	0.822
Nigeria * 60-79	1.421								

Źródło: opracowanie własne

4.2. Czy model językowy różnicuje swoje odpowiedzi ze względu na płeć i czy zaobserwowana zależność ma odzwierciedlenie w danych World Values Survey?

W celu odpowiedzenia na pytanie czy model językowy różnicuje swoje odpowiedzi ze względu na płeć autor wykorzystał Test T Studenta dla prób niezależnych lub jego nieparametryczną alternatywę, test Manna-Whitneya.

Założenia jakie muszą być każdorazowo spełnione, aby móc wykonać Test T Studenta są następujące:

- Liczebność próby:

Każda z badanych podgrup powinna zawierać co najmniej 30 obserwacji. Taka wielkość próby umożliwia przybliżenie rozkładu estymatora do rozkładu normalnego (zgodnie z centralnym twierdzeniem granicznym), co jest kluczowe dla poprawności testu.

- Normalność rozkładu reszt:

Rozkład reszt (bądź zmiennej losowej) powinien być zbliżony do rozkładu normalnego. Weryfikację normalności przeprowadza się przy użyciu miar skośności i kurtozy, przy czym akceptowalny zakres wartości przyjmuje się jako [-1, 1]. W sytuacjach, gdy wartości tych miar zbliżają się do granic, wskazane jest przeprowadzenie zarówno testu T Studenta (wersji parametrycznej), jak i nieparametrycznego testu Manna-Whitneya. Wówczas kryteria decyzyjne są następujące:

- Jeśli test Manna-Whitneya wskazuje na istotną różnicę między podgrupami, a test T Studenta nie, wyniki uznaje się za bardziej wiarygodne dla testu nieparametrycznego.
- Jeśli test T Studenta wykazuje istotną różnicę, a test Manna-Whitneya nie, decyzja o wyborze wyniku zależy od spełnienia pozostałych założeń:
  - Gdy pozostałe założenia (m.in. normalność rozkładu, równość wariancji) są spełnione – wyniki testu T są uznawane za wiarygodne.
  - Gdy pozostałe założenia nie są spełnione – wynik testu Manna-Whitneya uznaje się za bardziej adekwatny.

- Równość wariancji:

Test T Studenta zakłada homogeniczność wariancji między badanymi grupami. Niespełnienie tego założenia nie wyklucza jednak przeprowadzenia testu, ponieważ można zastosować korekcję Welcha. Procedura jest następująca:

- W pierwszej kolejności należy przeprowadzić test Levene'a w celu oceny równości wariancji.
- Jeśli wynik testu Levene'a pozwala odrzucić hipotezę zerową o równości wariancji (H0), należy zastosować wersję testu T dla wariancji nierównych.
- Dalszy tok rozumowania w zakresie interpretacji wyników pozostaje analogiczny.

4.2.1. Dane Ankietowe

Posługując się opisaną wyżej procedurą, w kategoriach „Zadowolenie i jakość życia", „Nauka i technologia", „Gospodarka i ekonomia" oraz „Zaufanie do instytucji" przeprowadzono Test T Studenta. Dla odpowiedzi na pytanie o „Demokratyczne wartości i prawa obywatelskie" wybrano test Manna-Whitneya.

Tabela 6. Test T-studenta

		F	Sig.	t	df	Two-Sided p
satysfakcja	Equal variances assumed	.579	.447	-1.105	12031	.269
	Equal variances not assumed			-1.105	11987.987	.269
korupcja	Equal variances assumed	15.327	<.001	-.610	12031	.542
	Equal variances not assumed			-.610	11919.724	.542
gospodarka	Equal variances assumed	.646	.421	3.365	12031	<.001
	Equal variances not assumed			3.365	12011.117	<.001
technologia	Equal variances assumed	.087	.768	5.540	12031	<.001
	Equal variances not assumed			5.538	11987.256	<.001

Źródło: opracowanie własne

Tabela 7. Test T-studenta - średnie

	plec	N	Mean	Std. Deviation	Std. Error Mean
satysfakcja	Mężczyzna	5888	7.13	2.140	.028
	Kobieta	6145	7.17	2.103	.027
korupcja	Mężczyzna	5888	7.37	2.400	.031
	Kobieta	6145	7.40	2.273	.029
gospodarka	Mężczyzna	5888	6.28	2.867	.037
	Kobieta	6145	6.11	2.873	.037
technologia	Mężczyźna	5888	7.59	2.372	.031
	Kobieta	6145	7.36	2.330	.030

Źródło: opracowanie własne

Tabela 8. Test Manna-Whitneya

	demokracja
Mann-Whitney U	17399443.000
Asymp. Sig. (2-tailed)	<.001

Źródło: opracowanie własne

Tabela 9. Test Manna-Whitneya - rangi

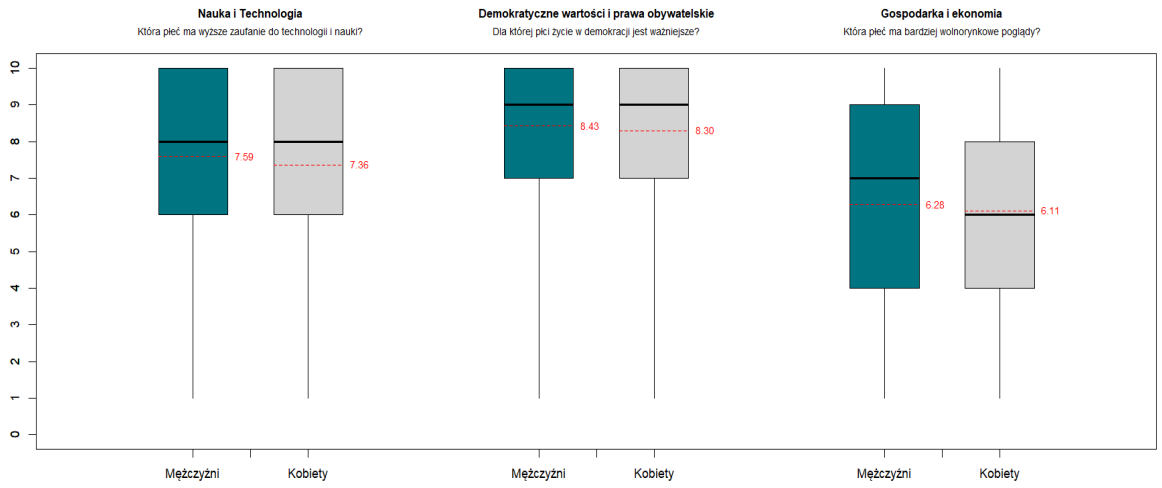
	plec	N	Mean Rank	Sum of Ranks
demokracja	Mężczyzna	5888	6134.43	36119533.00
	Kobieta	6145	5904.48	36283028.00

Źródło: opracowanie własne

Analiza odpowiedzi udzielonych w badaniu ankietowym World Values Survey (Wave 7) ujawnia bardziej wolnorynkowe podejście mężczyzn, którzy osiągnęli średnią ocenę 6,28 w skali od 1 do 10 w kwestii gospodarki (w porównaniu do 6,11 u kobiet), co potwierdzono testem T Studenta ($p < 0,001$) przy spełnionym założeniu równości wariancji (test Levene'a, $p = 0,421$), sugerując, że mężczyźni częściej popierają ideę, iż dochody powinny odzwierciedlać indywidualne wyniki, być może z powodu stereotypowego w kulturze zachodniej kultu mężczyzny-innowatora, prowadzącego spółki technologiczne na kapitalistycznych rynkach. W kwestii technologii mężczyźni uzyskali średnią 7,59, przewyższającą wynik kobiet (7,36), z różnicą istotną statystycznie ($p < 0,001$) i równymi wariancjami (test Levene'a, $p = 0,768$), co

może wiązać się z ich silniejszą obecnością w dziedzinach STEM i postrzeganiem technologii jako obszaru sukcesu, podczas gdy kobiety, choć coraz aktywniejsze w tych sferach, mogą być bardziej wyczulone na społeczne skutki postępu technologicznego. Ocena ważności życia w państwie demokratycznym ujawniła różnice potwierdzone testem Manna-Whitneya ($p < 0,001$), gdzie średni wynik mężczyzn (6134,43) przewyższył wynik kobiet (5904,48), wskazując, że mężczyźni przywiązują większą wagę do wartości demokratycznych. Autor podejrzewa, że jest to związane z tym, że mężczyźni są bardziej zainteresowani polityką. Widać to po rozkładzie poglądów wobec płci, gdzie mężczyźni w 41% przyjmują poglądy neutralne, w porównaniu do 49% u kobiet. W kategorii „Zadowolenie i jakość życia" średnie mężczyzn (7,13) i kobiet (7,17) były zbliżone, a test T Studenta nie wykazał różnic ($p = 0,269$) przy równych wariancjach ($p = 0,447$), podobnie jak w pytaniu o korupcję, gdzie wyniki (7,37 dla mężczyzn, 7,40 dla kobiet) nie różniły się istotnie ($p = 0,542$), choć nierówność wariancji (test Levene'a, $p < 0,001$) wymagała korekcji Welcha, co nie zmieniło wniosku o braku wyraźnych różnic.

Rysunek 5. Odmienność poglądów obu płci



Źródło: opracowanie własne. Czarna kreska oznacza medianę, a czerwona średnią.

4.2.2. Odpowiedzi modelu językowego Gemma 2 2b

Rozkład odpowiedzi modelu językowego w kategoriach „Demokratyczne wartości i prawa obywatelskie", „Zaufanie do instytucji publicznych" i „Nauka i technologia", wymagał od autora przeprowadzenia obu typów testów. Natomiast dla kategorii „Gospodarka i ekonomia" oraz „Zadowolenie i jakość życia" wykonano tylko test nieparametryczny.

Tabela 10. Test T-studenta

		F	Sig.	t	df	Two-Sided p
demokracja	Equal variances assumed	117.898	<.001	-2.247	10690	.025
	Equal variances not assumed			-2.247	10574.232	.025
korupcja	Equal variances assumed	27.418	<.001	.299	10690	.765
	Equal variances not assumed			.299	10659.918	.765
technologia	Equal variances assumed	174.375	<.001	5.427	10690	<.001
	Equal variances not assumed			5.427	10578.270	<.001

Źródło: opracowanie własne

Tabela 11. Test T-studenta - średnie

	plec	N	Mean	Std. Deviation	Std. Error Mean
demokracja	Mężczyzna	5346	6.4097	3.04299	.04162
	Kobieta	5346	6.5355	2.73963	.03747
korupcja	Mężczyzna	5346	6.64447	2.774078	.037941
	Kobieta	5346	6.62799	2.925572	.040013
technologia	Mężczyzna	5346	6.8366	2.87424	.03931
	Kobieta	5346	6.5180	3.18651	.04358

Źródło: opracowanie własne

Tabela 12. Test Manna-Whitneya - walidacja krzyżowa dla trzech kategorii

	demokracja	korupcja	technologia
Mann-Whitney U	14260091.000	14076360.000	14012965.000
Asymp. Sig. (2-tailed)	.851	.175	.080

Źródło: opracowanie własne

Tabela 13. Test Manna-Whitneya - rangi dla walidacji krzyżowej

	plec	N	Mean Rank	Sum of Ranks
demokracja	Mężczyzna	5346	5352.07	28612156.00
	Kobieta	5346	5340.93	28552622.00
korupcja	Mężczyzna	5346	5306.56	28368891.00
	Kobieta	5346	5386.44	28795887.00
technologia	Mężczyzna	5346	5398.29	28859282.00
	Kobieta	5346	5294.71	28305496.00

Źródło: opracowanie własne

Tabela 14. Test Manna-Whitneya

	gospodarka	satysfakcja
Mann-Whitney U	13038799.500	13062190.000
Asymp. Sig. (2-tailed)	<.001	<.001

Źródło: opracowanie własne

Tabela 15. Test Manna-Whitneya - rangi

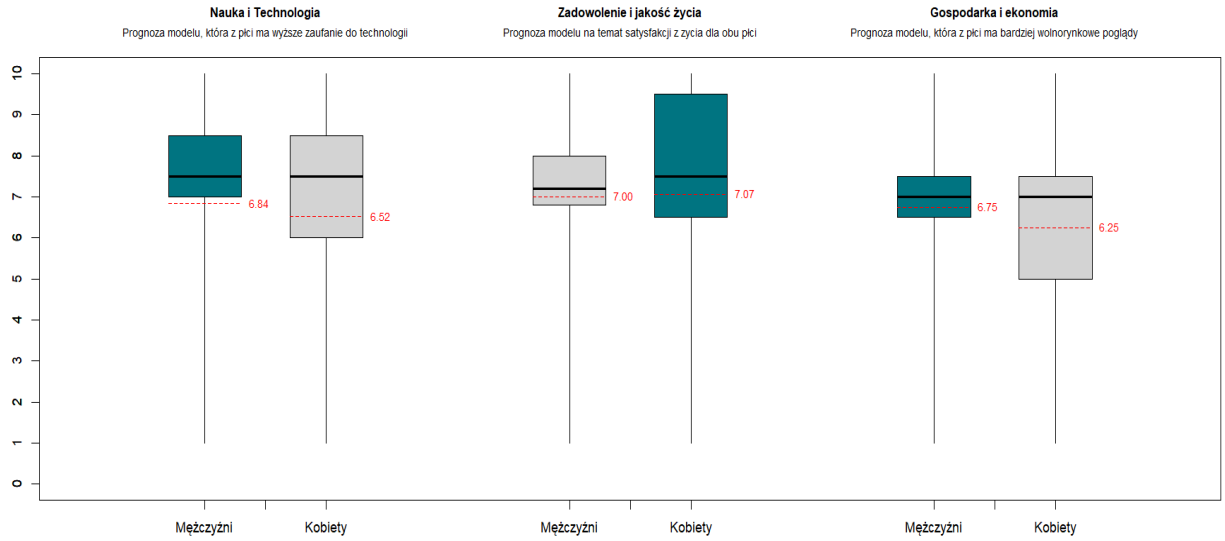
	plec	N	Mean Rank	Sum of Ranks
gospodarka	Mężczyzna	5346	5580.52	29833447.50
	Kobieta	5346	5112.48	27331330.50
satysfakcja	Mężczyzna	5346	5116.86	27354721.00
	Kobieta	5346	5576.14	29810057.00

Źródło: opracowanie własne

Według modelu językowego Gemma 2-2b, odpowiedzi mężczyzn wskazują na bardziej pozytywne spojrzenie na technologię i jej wpływ na ludzi– test T Studenta oraz Manna-

Whitneya wykazały istotną różnicę ($p < 0.001$) z wyższą średnią mężczyzn (6.8366, odchylenie standardowe = 2.87424) wobec kobiet (6.5180, odchylenie standardowe = 3.18651). Model zakłada również, że mężczyźni prezentują bardziej wolnorynkowe poglądy, skłaniając się ku uzależnieniu dochodów od indywidualnych wyników, co potwierdza test Manna-Whitneya dla kategorii „Gospodarka i ekonomia" ($U = 13038799.5$, $p < 0.001$) z wyższą rangą mężczyzn (5580.52) niż kobiet (5112.48). Interesująca jest prognoza modelu dotycząca satysfakcji życiowej – Gemma 2-2b przewiduje, że kobiety wyrażają wyższą satysfakcję niż mężczyźni ($U = 13062190$, $p < 0.001$). Kobiety osiągnęły wyższą rangę (5576.14) wobec mężczyzn (5116.86), co może odzwierciedlać założenia modelu oparte na danych wskazujących na różne priorytety płci. W kategoriach „Demokratyczne wartości i prawa obywatelskie" oraz „Zaufanie do instytucji publicznych" model nie przewiduje wyraźnych różnic.

Rysunek 6. Odmienność postrzegania poglądów kobiet i mężczyzn przez model



Źródło: opracowanie własne. Czarna kreska oznacza medianę, a czerwona średnią.

4.2.3. Odpowiedź na pytanie badawcze

W pierwszej kolejności zbadano, czy Gemma 2-2b prezentuje zróżnicowane perspektywy w zależności od deklarowanej płci. Okazało się, że model częściowo różnicuje odpowiedzi, choć nie we wszystkich tematach jednakowo. Dla niektórych zagadnień – na przykład kwestii technologicznych i ekonomicznych – odpowiedzi generowane dla „męskiego" i „żeńskiego" punktu widzenia wykazywały zauważalne odmienności. Model sugerował np. nieco większy nacisk mężczyzn na aspekty wolnorynkowe i technologiczne, podczas gdy w perspektywie kobiecej akcentował bardziej kwestie społeczne, odzwierciedlone wyższą wartością satysfakcji z życia dla kobiet. Z kolei przy pytaniach o demokrację i korupcję Gemma 2 2B udzielał bardzo

zbliżonych odpowiedzi niezależnie od płci, co wskazuje, że w tych obszarach model nie wprowadzał silnych rozróżnień między perspektywą męską a żeńską.

W drugim etapie dokonano porównania odpowiedzi modelu z danymi empirycznymi pochodzącymi z World Values Survey (WVS), aby ocenić, w jakim stopniu Gemma 2 2B odzwierciedla rzeczywiste różnice między kobietami a mężczyznami. Na przykład w globalnych badaniach WVS często obserwuje się, że kobiety i mężczyźni w pewnym stopniu różnią się w opiniach o technologii, ekonomii czy demokracji. Odpowiedzi modelu językowego częściowo odzwierciedliły te trendy: tam, gdzie dane empiryczne wskazywały na drobne rozbieżności między płciami (np. podejście do postępu technologicznego), model również generował odpowiedzi sugerujące taką różnicę. Świadczy to o tym, że model posiada pewną wiedzę lub wyuczone schematy zbieżne z rzeczywistością.

Niemniej jednak odnotowano także rozbieżności między odpowiedziami modelu a danymi WVS. W kilku przypadkach model przypisał płciom różne stanowiska tam, gdzie faktycznie (według WVS) kobiety i mężczyźni myślą podobnie, oraz odwrotnie – pominął subtelne różnice obecne w danych. Przy pytaniu o ogólną satysfakcję z życia model sugerował wyższe zadowolenie wśród kobiet, choć rzeczywiste dane WVS wskazują, że różnice w satysfakcji życiowej między płciami są statystycznie nieistotne.

Model nie różnicuje odpowiedzi według płci konsekwentnie oraz nie zawsze robi to zgodnie z rzeczywistymi danymi. Tam, gdzie badania WVS pokazują wyraźniejsze różnice między kobietami i mężczyznami, model często potrafił je wychwycić kierunkowo. Gdzie indziej model albo nadinterpretuje różnice („Zadowolenie i jakość życia”), albo je pomija („Demokratyczne wartości i prawa obywatelskie”). Zgodność w trzech na pięć kategorii między modelem a WVS, musi zostać odnotowana, i pozwala to jednak założyć, że model, w ograniczonym stopniu, dysponuje ogólną wiedzą o typowych trendach społecznych, natomiast rozbieżności wskazują na ograniczenia modelu i ewentualne wpływy stereotypów lub braku kontekstu kulturowego. Źródłem tych różnic mogą być zarówno specyfika danych treningowych Gemmy 2 2B, jak i fakt, że model nie „rozumie” kontekstu demograficznego tak jak badanie społeczne – generuje odpowiedzi na podstawie prawdopodobieństw językowych. Podsumowując, model w więcej niż połowie analizowanych obszarów generuje zróżnicowane perspektywy, a w większości przypadków poprawnie identyfikuje rzeczywiste zależności.

Autor podkreśla konieczność zachowania ostrożności przy wykorzystywaniu modelu do symulowania rzeczywistych opinii, zwłaszcza w kontekście politycznym. Przykładowo,

zastanawiające jest to, że Gemma 2 2B przypisuje mężczyznom niższą satysfakcję z życia w porównaniu do kobiet, co może prowadzić do nieuzasadnionych uogólnień lub dyskryminujących interpretacji. Wyniki te wskazują na potrzebę dalszej analizy i doskonalenia modeli językowych w celu lepszego odzwierciedlenia rzeczywistych danych społecznych.

4.3. Modele dla różnic pomiędzy danymi ankietowymi a predykcjami modelu

Stworzone w tym etapie badania modele regresyjne cechują się wyższym dopasowaniem do rzeczywistych danych niż modele z poprzedniego etapu. Wartości nieskorygowanego współczynnika R^2 przyjmują wartości od 6.8% dla modelu „Demokratyczne wartości i postawy obywatelskie" do 29% w modelu „Zaufanie do instytucji publicznych".

Używając uogólnionego testu F Walda stwierdzono także, że modele są istotne. Na poziomie istotności 0.10 odrzucono za każdym razem hipotezę zerową (wszystkie zmienne w modelu nieistotne), na rzecz hipotezy alternatywnej, głoszącej, że minimum jedna zmienna jest istotna.

Wartości statystyk opisowych i wizualnej oceny rozkładu reszt, wskazują na pewne problemy z założeniami Gaussa-Markova, jednak tylko w niektórych obszarach. W modelu „Zadowolenie i jakość życia" występuje autokorelacja reszt (Durbin-Watson = 0.127). Pozostałe 4 modele nie wykazują powyższego problemu.

Niskie wartości VIF dla wszystkich predyktorów świadczą o braku istotnych zależności między nimi. Współczynnik inflacji wariancji ani razu nie przekracza wartości 2.

Analiza histogramów reszt ujawniła, że wszystkie wykresy rozkładu reszt ściśle pokrywają się z nałożoną krzywą normalną. Dzięki temu wyniki badania można uogólniać na większe populacje.

Wykresy rozrzutu reszt wykazują pewne wzorce, przez co punkty na nich nie są rozłożone równomiernie. Indukuje to problem z heteroskedastycznością. Autor podjął próbę znalezienia rozwiązania. Przekształcono wszystkie zmienne zależne w następujący sposób. Różnicę odpowiedzi najpierw potraktowano wartością bezwzględną a następnie obliczono z tej wartości bezwzględnej logarytm naturalny. Autor zdecydował się nie podejmować dodatkowych kroków naprawczych przede wszystkim w trosce o pełną porównywalność modeli. Mimo że transformacja ln(|różnica|) nie wyeliminowała całkowicie problemów z heteroskedastycznością, to znacząco spłaszczyła rozkład reszt i pozwoliła na łatwiejszą interpretację w kategoriach względnych (procentowych). Przy tak dużej liczbie obserwacji

(10 692) dalsze modyfikacje mogłyby wprowadzać niejednorodność pomiędzy poszczególnymi analizami, dlatego autor uznał dotychczasowe zabiegi za wystarczające.

Analiza wyników testu RESET, przeprowadzona z użyciem trzecich potęg niestandaryzowanych wartości przewidywanych, przynosi pozytywne wnioski dla modeli w wybranych obszarach. W przypadku modeli opisujących Demokratyczne wartości i prawa obywatelskie (p=0.103), Zaufanie do instytucji (p=0.305) oraz Naukę i technologię (p=0.308), test nie dostarczył statystycznych podstaw (na poziomie istotności $\alpha = 0.10$) do odrzucenia hipotezy zerowej. Oznacza to, że w trzech na pięć obszarów postać funkcyjna jest poprawna. Autor nie podjął się dodatkowych kroków, aby zachować pełną porównywalność wyników.

Tabela 16. Wartości współczynników β ostatecznych modeli regresyjnych

Zmienna	**Satysfakcja**	**Demokracja**	**Korupcja**	**Technologia**	**Gospodarka**
Constant	-0,314*	0,822*	-0,033	0,464*	1,014*
country=USA (bazowe)					
country=Argentyna	1,279*	-0,126*	0,446*	0,483*	-0,359*
country=Australia	1,495*	-0,301*	-0,259*	-0,036	-0,762*
country=Czechy	0,059	-0,593*	-0,335*	0,315*	-1,101*
country=Filipiny	0,334*	-0,312*	-0,354*	0,663*	-0,505*
country=Indie	0,841*	-0,297*	0,906*	0,389*	-0,516*
country=Japonia	0,226*	-0,291*	1,319*	0,120*	0,044
country=Libia	0,433*	-0,090*	1,266*	0,992*	0,519*
country=Niemcy	0,804*	-0,110*	0,603*	-0,033	-0,673*
country=Nigeria	0,969*	-0,132*	0,808*	0,745*	-0,374*
country=Rosja	0,602*	-0,780*	-0,279*	0,344*	-0,282*
plec=female (bazowe)					
plec=Male	-0,383*	0,183*	-0,062*	-0,202*	-0,341*
wyksztalcenie=lower (bazowe)					
wyksztalcenie=Higher	-0,240*	-0,153*	-0,214*	-0,604*	-0,253*
wyksztalcenie=Middle	-0,200*	-0,236*	-0,234*	-0,467*	-0,192*

Zmienna	Satysfakcja	Demokracja	Korupcja	Technologia	Gospodarka
city_size=under 10 000 (bazowe)					
city_size=10 000-500 000	0,128*	-0,040	0,107*	-0,303*	0,078*
city_size=500 000 and more	0,232*	0,189*	0,342*	-0,007	0,094*
age=20-39 (bazowe)					
age=40-59	-0,052*	-0,052*	-0,056*	-0,186*	-0,053*
age=60-79	-0,232*	0,042	-0,051*	-0,173*	-0,026
work=unemployed (bazowe)					
work=Employed	-0,124*	-0,106*	-0,131*	-0,183*	-0,094*
kids=0 (bazowe)					
kids=1-3 kids	0,073*	0,070*	0,118*	0,304*	-0,026*
kids=4 and more kids	0,009	-0,024	0,163*	0,241*	-0,112*
left_right=neutral (bazowe)					
left_right=Left	0,131*	-0,193*	-0,061*	-0,159*	0,398*
left_right=Right	0,012	-0,019	0,067*	-0,157*	-0,206*

Źródło: opracowanie własne. Wartości oznaczone gwiazdką są istotne na poziomie 10% ($p < 0,10$).

W drugim etapie modelowania regresyjnego zmienna zależna została zdefiniowana jako logarytm naturalny wartości bezwzględnych różnic pomiędzy przewidywanymi wartościami odpowiedzi respondentów (obliczonymi w pierwszym etapie) a odpowiedziami modelu językowego Gemma 2 2B. Współczynniki regresji wskazują, o ile logarytm tej różnicy zmienia się dla danej kategorii zmiennej niezależnej w porównaniu do kategorii bazowej, przy założeniu stałości pozostałych zmiennych. Dodatnie wartości współczynników oznaczają większą rozbieżność między przewidywaniami a odpowiedziami modelu, natomiast wartości ujemne wskazują na mniejszą rozbieżność. Aby przeliczyć wpływ na procentową zmianę różnicy, stosuje się wzór: $(e^{\beta} - 1) * 100\%$.

1. Satysfakcja z życia – Argentyna ($\beta = 1,279$): Współczynnik 1,279 dla respondentów z Argentyny oznacza, że przy założeniu stałości pozostałych zmiennych, logarytm naturalny wartości bezwzględnej różnicy między przewidywaną satysfakcją z życia a odpowiedzią modelu językowego jest o 1,279 większy w porównaniu do kategorii bazowej (USA). Obliczając procentowy wzrost: $e^{1,279} \approx 3,592$, $(3,592 - 1) * 100\% = 259,2\%$. Oznacza to, że

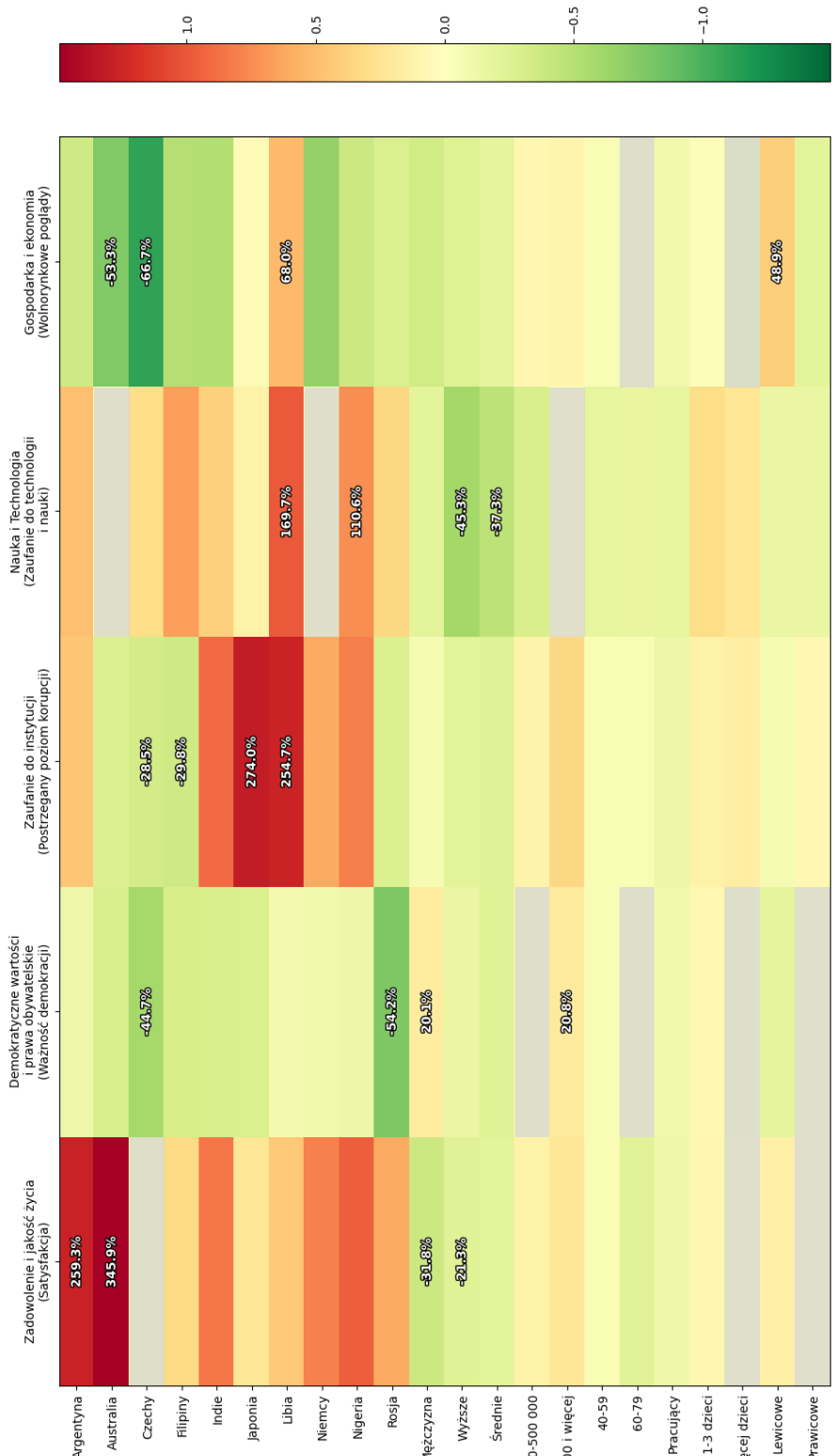
dla Argentyńczyków rozbieżność między przewidywaniami a odpowiedziami modelu jest o 259,2% większa niż dla respondentów z USA, co może sugerować trudności modelu w dokładnym odwzorowaniu specyficznych uwarunkowań kulturowych lub społecznych Argentyny w tym obszarze.

2. Percepcja korupcji – Japonia (β = 1,319): Współczynnik 1,319 dla Japonii wskazuje, że przy założeniu stałości pozostałych zmiennych, logarytm naturalny różnicy w percepcji korupcji jest o 1,319 większy niż dla USA. Przeliczając: e^1,319 ≈ 3,741, (3,741 - 1) * 100% = 274,1%. Oznacza to, że rozbieżność między przewidywaniami a odpowiedziami modelu dla Japończyków jest o 274,1% większa niż dla Amerykanów. Dla mieszkańców Japonii, model, odpowiadając na pytanie o istnienie korupcji, popełnia o 274% wyższe błędy niż dla Amerykanów. Oznacza to, że, w porównaniu do Amerykanów, Japończycy są kompletnie przez model niezrozumiani.

3. Znaczenie demokracji – Mężczyzna (β = 0,183): Współczynnik 0,183 dla mężczyzn oznacza, że przy założeniu stałości pozostałych zmiennych, logarytm naturalny różnicy w ocenie znaczenia demokracji jest o 0,183 większy w porównaniu do kobiet (kategorii bazowej). Obliczając: e^0,183 ≈ 1,201, (1,201 - 1) * 100% = 20,1%. Rozbieżność między przewidywaniami a odpowiedziami modelu z perspektywy mężczyzny jest o 20,1% większa niż dla kobiet.

Rysunek 7, przedstawia mapę cieplną (heatmapę) ilustrującą oszacowania współczynników w tym etapie. Komórki oznaczone kolorem szarym wskazują wyniki nieistotne statystycznie. Kolory zastosowane na mapie cieplnej odzwierciedlają wartości współczynników przed ich przeliczeniem na procenty, przy czym dla wybranych komórek nałożono wartości procentowe uzyskane na podstawie wzoru (e^β - 1) * 100%. W każdym z analizowanych obszarów tematycznych zidentyfikowano dwie zmienne – charakteryzujące się największą i najmniejszą rozbieżnością w odpowiedziach modelu językowego w porównaniu do przewidywanych wartości – a ich procentowe odpowiedniki zostały zaznaczone na wykresie.

Rysunek 7. Mapa cieplna



Źródło: opracowanie własne

W obszarze satysfakcji z życia Australia (β=1,495) odznacza się wyraźnie wyższym błędem niż Stany Zjednoczone, co w przeliczeniu na skalę procentową wskazuje na około 346% większą rozbieżność. Przy tym samym kraju w modelu dotyczącym gospodarki (β=–0,762) obserwuje się natomiast około 53,4% niższy błąd, co sugeruje, że Gemma 2 2B potrafi bardziej zbliżyć się do australijskich poglądów na kwestie ekonomiczne niż do ich poziomu zadowolenia z życia. Jeszcze większa rozpiętość uwidacznia się w przypadku Filipin, gdzie w sferze demokracji (β=–0,312) rozbieżność jest średnio o 26,8% mniejsza niż w USA, ale już w podejściu do technologii (β=0,663) wzrasta o blisko 94%. Taka dysproporcja może wynikać z unikatowych warunków społeczno-kulturowych Filipin, gdzie z jednej strony istnieją silne tradycje partycypacji obywatelskiej, a z drugiej – rozwój technologiczny czy dostęp do infrastruktury mogą stanowić istotne bariery. Libia, której współczynnik w wymiarze technologii (β=0,992) okazuje się o 170% wyższy niż w Stanach Zjednoczonych, zdaje się potwierdzać te przypuszczenia: rozwojowe niejednorodności i brak spójnej infrastruktury wyraźnie utrudniają modelowi poprawne uchwycenie tamtejszych uwarunkowań.

Nie mniej ciekawe zależności pojawiają się przy obserwacji pozostałych krajów. Argentyna, mająca wysoki współczynnik satysfakcji (β=1,279, czyli około 259% wyższą rozbieżność niż w USA), w obszarze demokracji odznacza się już wartością –0,126, sygnalizując mniejsze niż w Stanach Zjednoczonych odchylenie w przewidywaniach modelu. Można więc wnioskować, że choć Gemma 2 2B napotyka trudności w precyzyjnym oszacowaniu poziomu zadowolenia Argentyńczyków z życia, to całkiem dobrze „wyczuwa" ich postawy wobec demokracji. Z kolei Niemcy (β=0,804 w satysfakcji) wykazują około 124% większą rozbieżność niż USA, lecz w kwestii demokracji (β=–0,110) błąd spada o około 10%.

Spośród cech demograficznych wyróżnia się płeć: w odniesieniu do satysfakcji z życia współczynnik dla mężczyzn (β=–0,383) oznacza około 31,9% niższy błąd w porównaniu z kobietami, ale w tematyce demokracji (β=0,183) jest już wyższy o 20%. Można zatem przypuszczać, że model lepiej oddaje determinanty jakości życia wśród męskiej populacji, lecz mniej adekwatnie uchwytuje ich postawy obywatelskie czy polityczne. Warto też zwrócić uwagę na to, że w każdej kategorii niższy błąd w porównaniu do niższego wykształcenia, występuje u osób z wyższym wykształceniem, co wskazuje, że Gemma 2 2B sprawniej „odczytuje" opinie osób lepiej wyedukowanych w dziedzinie nowoczesnych rozwiązań czy ekonomii.

W analizie uwzględniono także wskaźniki rozwoju społeczno-ekonomicznego, odzwierciedlone między innymi przez poziom HDI, co pozwoliło na interesujące porównanie między regionami. Wyniki wskazują, że Gemma 2 2B znacznie lepiej dopasowuje się do danych pochodzących z krajów o wysokim HDI, gdzie stabilność instytucjonalna i przewidywalność warunków społeczno-ekonomicznych umożliwiają dokładniejsze „odczytanie" realiów. Przykładowo, państwa takie jak Niemcy, Czechy, Australia oraz Japonia, reprezentujące regiony Europy i Azji Centralnej oraz Azji Wschodniej i Pacyfiku, charakteryzują się wysokimi poziomami HDI – oscylującymi w przedziale 0,90–0,95 – co wiąże się z niewielkimi rozbieżnościami między przewidywaniami a faktycznymi odpowiedziami. Z kolei kraje z Afryki Subsaharyjskiej, reprezentowane przez Nigerię, oraz państwa z Bliskiego Wschodu i Afryki Północnej, takie jak Libia, prezentują niższe HDI. Taki stan rzeczy przekłada się na znacznie większe rozbieżności – co w przypadku Libii w obszarze technologii osiąga aż 170% różnicy względem USA, a w Nigerii model odnotowuje jedne z najwyższych błędów.

Całość wyników potwierdza, że kontekst narodowy i demograficzny odgrywa kluczową rolę w tym, na ile Gemma 2 2B potrafi zbliżyć się do przewidywanych odpowiedzi. Żeby odpowiedzieć na pierwsze pytanie badawcze, zsumowano współczynniki dla każdej kategorii (kraj, płeć, wykształcenie itd.), aby wskazać podgrupę, w której rozbieżność jest najmniejsza, oraz tę, w której błąd jest najwyższy. Najdokładniej uchwycony został profil osoby z Czech, mężczyzny, o wyższym wykształceniu, zamieszkałego w mieście od 10 000 do 500 000 mieszkańców, w wieku 60–79 lat, aktywnego zawodowo, bez dzieci i o poglądach prawicowych. Z kolei największe rozbieżności pojawiają się przy kobietach z Libii, z niższym wykształceniem, żyjących w dużym mieście, w grupie wiekowej 20–39 lat, pozostających bez pracy, mających 1–3 dzieci i deklarujących poglądy lewicowe.

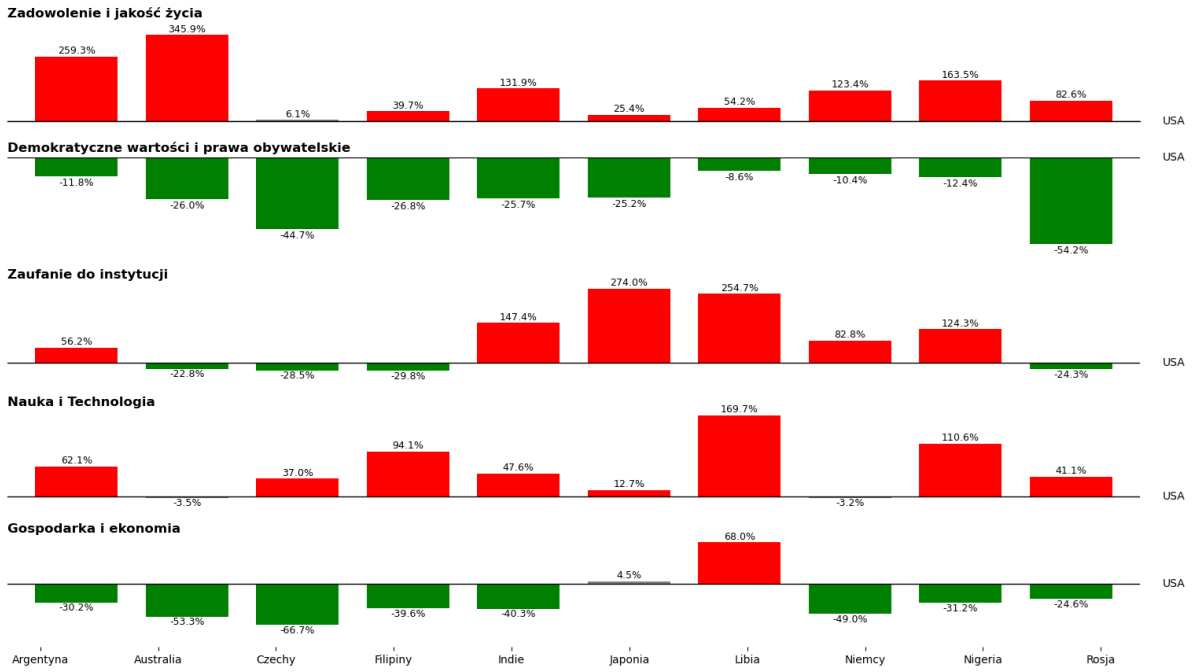
4.4. Czy model językowy Gemma 2 2B odzwierciedla poglądy jego twórców?

Wyniki analizy nie pozwalają jednoznacznie odpowiedzieć na to pytanie. Mimo, że na trzy z pięciu pytań, odpowiedzi modelu są najbardziej zbliżone lub dominując wśród najbardziej zbliżonych do odpowiedzi ankietowanych z USA, dwie kategorie, historycznie i kulturowo kojarzone z USA, ujawniają, że model nie powiela poglądów ludzi tam mieszkających.

USA jest kojarzone jako państwo będące wzorem demokracji i stabilności (Sozan, 2024) oraz posiadające wielu obywateli z silnymi poglądami na temat wolnego rynku. W obu tych kategoriach Model Gemma 2 2B osiąga najwyższe lub prawie najwyższe błędy i nie potrafił wskazać, jakich odpowiedzi udzielają obywatele tego państwa. Znaczne zawyżenie

prognozowanych odpowiedzi na temat wolnorynkowych poglądów, jasno implikuje mocno stereotypowe spojrzenie modelu na Stany Zjednoczone. Z drugiej strony, wyniki badania ankietowego pokazuje, że obywatele USA wierzą w to, że w ich państwie przestrzegane są demokratyczne wartości. Na tym polu, model Gemma 2 2B znowu ujawnia swoje stereotypowe spojrzenie i najprawdopodobniej, zgodnie z mocnym spolaryzowaniem treści w Internecie, będących później korpusem treningowym modelu, stwierdza, że obywatele zauważają problemy z ustrojem demokratycznym, co w rzeczywistości nie ma miejsca.

Rysunek 8. USA vs inne państwa



Źródło: opracowanie własne. Wartości dodatnie pokazują o ile procent model popełnia większe błędy w stosunku do obywateli USA. Wartości ujemne, zaznaczone kolorem zielonym, wskazują, dla których państw model jest celniejszy i o ile procent.

5. Wnioski

Badanie obiektywizmu modeli językowych jest nie tylko aktualnym, ale i palącym wyzwaniem w obliczu trwającej rewolucji technologicznej, która toczy się na naszych oczach i nieodwracalnie przekształca liczne sfery życia. Rozwój AI niesie ze sobą zarówno ogromne szanse, jak i poważne ryzyka społeczne, które w dłuższej perspektywie mogą prowadzić do niepożądanych konsekwencji i pogorszenia warunków życia znacznej części społeczeństw. Zrozumienie silnych i słabych stron tych technologii, w tym ich potencjalnej tendencyjności wobec różnych grup społecznych, demograficznych czy narodowych, jest niezbędne dla efektywnej współpracy między twórcami modeli a decydentami publicznymi. Choć korporacje technologiczne, działając w warunkach silnej konkurencji, mogą nie zawsze priorytetyzować badania nad etycznymi aspektami swoich produktów, to właśnie takie analizy mogą dostarczyć im wiedzy potrzebnej do udoskonalenia procesów trenowania modeli, na przykład poprzez wzbogacenie zbiorów danych o informacje dotyczące niedostatecznie reprezentowanych grup. Postęp w tej dziedzinie nie dokona się jednak bez odpowiednich regulacji prawnych. Zadaniem badaczy jest dostarczanie rzetelnych danych i analiz, które pomogą ukierunkować działania legislacyjne w taki sposób, by z jednej strony nie zdusić innowacyjności, a z drugiej zapewnić bezpieczeństwo użytkowników i długoterminową stabilność instytucji społecznych, których zagrożenie przez niekontrolowany rozwój AI jest coraz częściej sygnalizowane (Harari, 2025). Autor pracy wskazuje również potencjalne kierunki dalszych badań, obejmujące zastosowanie podobnej metodologii wobec innych modeli językowych lub uwzględnienie dodatkowych cech respondentów, co zamierza eksplorować w przyszłości.

Przeprowadzona w niniejszej pracy analiza tendencyjności odpowiedzi udzielanych przez model językowy Gemma 2 2B na pytania dotyczące korupcji, poglądów ekonomicznych, satysfakcji z życia, nastawienia do demokracji czy zaufania do technologii przyniosła interesujące rezultaty. Analiza wykazała, że odpowiedzi generowane przez model w zróżnicowanym stopniu pokrywają się z opiniami badanych grup społecznych, demograficznych czy narodowych. Stopień tego dopasowania różnił się w zależności od tematu, ujawniając jednak pewne powtarzalne wzorce – najczęściej faworyzujące osoby wykształcone z krajów zachodniego kręgu kulturowego.

Pierwsza hipoteza badawcza zakładała, że model Gemma 2 2B generuje odpowiedzi, których zbieżność z opiniami poszczególnych grup (społecznych, narodowych, demograficznych), zidentyfikowanych na podstawie badania WVS jest zróżnicowana. Wyniki pokazane w pracy sugerują, że prawdziwość tej tezy jest zależna od obszaru tematycznego. Udało się również

precyzyjnie scharakteryzować społeczności, dla których model językowy Gemma 2 2B miał wyraźnie większe trudności z trafnym przewidzeniem ich rzeczywistych odpowiedzi z badania ankietowego.

Niepokojącym jest fakt, że w ponad połowie analizowanych obszarów tematycznych (trzech z pięciu) Libia okazała się państwem, dla którego model generował najmniej trafne odpowiedzi. Z drugiej strony, najwyższą precyzję prognoz model osiągał dla reprezentantów państw kulturowo zaliczanych do współczesnego Zachodu, takich jak Czechy czy USA. Obserwujemy również różnice w trafności przewidywania odpowiedzi dla kobiet i mężczyzn – w czterech na pięć badanych obszarów tematycznych model osiągał wyższą precyzję dla mężczyzn. Trafność prognoz wydaje się również rosnąć wraz z poziomem wykształcenia. Wyniki analiz wskazują, że w większości obszarów tematycznych model lepiej rozumie osoby wykształcone niż te o niższym poziomie edukacji. Różnica w precyzji może sięgać nawet ponad 25%. Oznacza to, że model odgrywając rolę ludzi wykształconych, niezależnie od innych cech, szacował ich poglądy o ponad 25% celniej, niż dla ludzi bez wykształcenia Gemma 2 2B wydaje się zatem lepiej rozumieć perspektywę osób lepiej wykształconych.

Analizując zmienną określającą wielkość miejscowości, największe błędy predykcji obserwujemy dla osób mieszkających w metropoliach powyżej 500 000 mieszkańców w porównaniu do osób z mniejszych miast. Może to jednak częściowo wynikać ze sposobu kategoryzacji, który zrównuje duże miasta w Nigerii, Indiach, Filipinach czy Japonii z metropoliami w Stanach Zjednoczonych i Europie. Procesy centralizacyjne, szczególnie w Afryce Zachodniej czy Azji Południowo-Wschodniej, sprawiają, że warunki życia w tamtejszych metropoliach mogą znacząco odbiegać od tych, które stereotypowo kojarzymy z mieszkańcem europejskiego miasta zaliczonego w modelu do tej samej kategorii. Wbrew intuicji sugerującej, że reprezentacja w modelu będzie proporcjonalna do aktywności danej grupy w internecie, opinie najstarszej badanej grupy wiekowej (60-79 lat) okazały się najlepiej przewidywane przez model. Gemma 2 2B udzielał także znacznie trafniejszych odpowiedzi dla osób zatrudnionych (w porównaniu z niepracującymi) oraz nieposiadających dzieci (w porównaniu z rodzicami).

Analiza odpowiedzi generowanych przez model Gemma 2 2B ujawniła pewne tendencje, jednak badanie wykazało jednocześnie, że model ten trafniej przewidywał poglądy osób deklarujących orientację prawicową. Obserwacja ta sugeruje, że charakterystyka

generowanych odpowiedzi nie musi wprost przekładać się na dokładność modelu w odzwierciedlaniu opinii różnych grup politycznych w konfrontacji z danymi ankietowymi.

Podsumowując odpowiedź na pierwsze pytanie badawcze, warto wskazać na profil osoby, której opinie model Gemma 2 2B wydaje się rozumieć najlepiej, minimalizując ryzyko tendencyjnego potraktowania. Kombinacja ośmiu cech charakteryzująca tę osobę to: mężczyzna z Czech, z wyższym wykształceniem, mieszkający w średniej lub małej wielkości mieście, w wieku 60-79 lat, pracujący, nieposiadający dzieci oraz o poglądach prawicowych. Dla tak zdefiniowanego profilu model, w przekroju pięciu badanych obszarów tematycznych, udzielał odpowiedzi najbardziej zbieżnych z rzeczywistymi odpowiedziami ankietowymi.

W ramach drugiej hipotezy badawczej sprawdzono, czy model Gemma 2 2B różnicuje generowane odpowiedzi w zależności od płci oraz czy te różnice znajdują odzwierciedlenie w danych empirycznych z badania World Values Survey. Badanie wykazało, że model rzeczywiście dostosowuje swoje odpowiedzi w zależności od symulowanej płci w większości analizowanych kategorii. Co istotne, w trzech z pięciu obszarów tematycznych (dotyczących gospodarki, technologii i demokracji) model poprawnie uchwycił kierunek rzeczywistych różnic obserwowanych w danych WVS. Dostarcza to argumentów na rzecz pierwszej części hipotezy – dotyczącej samego faktu różnicowania odpowiedzi przez model. Jednocześnie, mimo tej częściowej zgodności, wyniki sugerują potrzebę daleko idącej ostrożności, szczególnie w kontekście potencjalnych zastosowań przez instytucje publiczne. Alarmującym przykładem jest tendencja modelu do generowania odpowiedzi sugerujących wyższą satysfakcję życiową kobiet, mimo iż dane empiryczne dla badanych krajów nie wskazują na istnienie takiej różnicy między płciami. Taka niezgodność z faktami powinna skłaniać do głębokiej rozwagi przy interpretacji wyników modelu. Mimo tej znaczącej rozbieżności, warto odnotować, że trafne odwzorowanie trendów w trzech pozostałych obszarach sugeruje, iż model posiada pewną zakodowaną "wiedzę" o ogólnych wzorcach społecznych związanych z płcią.

Ostatnie pytanie badawcze dotyczyło kwestii, czy model językowy Gemma 2 2B odzwierciedla poglądy swoich twórców. Wnioski płynące z tej analizy różnią się od niektórych sugestii w literaturze – uzyskane wyniki nie wskazują, by model Gemma 2 2B był wyraźnie stronniczy na rzecz poglądów reprezentatywnych dla społeczeństwa amerykańskiego. Model Gemma 2 2B nie przyjmuje więc poglądów twórców tej narodowości. Należy przy tym podkreślić ograniczenie metodologiczne tej części badania: zgodnie z przyjętym i udokumentowanym w

literaturze podejściem (Buyl, et al., 2024), poglądy twórców przybliżono za pomocą danych dla ogółu społeczeństwa USA. Wybór tej metody, mimo jej niepełnej precyzji, podyktowany był chęcią zachowania porównywalności z wcześniejszymi pracami.

Niedoskonałością przeprowadzonego badania, wynikającą z przyjętej metodologii, jest ograniczenie analizy do ośmiu cech respondentów, co siłą rzeczy stanowi uproszczenie złożonej rzeczywistości ludzkich postaw i opinii. Zastosowane modele regresyjne, mimo iż istotne statystycznie, charakteryzowały się stosunkowo niskim współczynnikiem R^2, co oznacza, że duża część zmienności w danych pozostała niewyjaśniona. Problemy z postacią funkcyjną niektórych modeli sugerują, że pominięte mogły zostać istotne zmienne lub interakcje. Rozbudowanie modeli o dodatkowe cechy, takie jak wyznawana religia czy poziom zarobków oraz uwzględnienie bardziej złożonych interakcji mogłoby potencjalnie poprawić dopasowanie modeli i zwiększyć ich wartość poznawczą. Należy jednak mieć świadomość, że takie rozszerzenie analizy wiązałoby się ze znacznym wzrostem złożoności obliczeniowej – na przykład dodanie zmiennej opisującej przynależność religijną (z pięcioma kategoriami) zwiększyłoby liczbę analizowanych podgrup z około 10 tysięcy do 50 tysięcy. Ponadto, badanie opierało się na metodzie "czarnej skrzynki", bez wglądu w jego wewnętrzną architekturę czy parametry. Choć podejście to jest uzasadnione w kontekście badania modeli zamkniętych i pozwala ocenić ich zewnętrzne zachowanie, uniemożliwia pełne zrozumienie przyczyn zaobserwowanych tendencji, które mogą wynikać nie tylko z danych treningowych, ale również z procesu dostrajania (eng, fine-tuningu) czy samej architektury modelu.

Na podstawie przeprowadzonych badań, autor pragnie sformułować kilka zaleceń dla organizacji, zwłaszcza tych pożytku publicznego, rozważających wdrożenie systemów opartych na sztucznej inteligencji. Przede wszystkim kluczowe są transparentność i rozwaga. Podczas projektowania i wdrażania takich systemów fundamentalne jest porzucenie założenia, że model jest obiektywny i wolny od uprzedzeń. Jak pokazują wyniki tej pracy i liczne inne badania, wszystkie modele językowe są w pewnym stopniu tendencyjne, a ich skłonności mogą ewoluować. Istnieją jednak metody łagodzenia tego problemu, takie jak świadome dostrajanie modeli (eng. fine-tuning) z wykorzystaniem odpowiednio przygotowanych zbiorów danych, co może "przesunąć" charakterystykę modelu w pożądanym kierunku (Rozado, 2024). Niezbędne jest również rozważne podejście do interakcji z modelem oraz ścisła współpraca z zespołem technicznym w celu zidentyfikowania potencjalnych grup użytkowników lub interesariuszy i oceny, jak model odnosi się do specyfiki tych grup. Wartościowym rozwiązaniem może być powołanie niezależnego ciała doradczego, które testowałoby model i opiniowało nowe

funkcjonalności, zapewniając obiektywną ocenę i wspierając utrzymanie wysokich standardów etycznych i społecznych.

Organizacje publiczne powinny ponadto wdrożyć procedury regularnego testowania systemów AI pod kątem niezamierzonych skutków dyskryminacyjnych, wykorzystując do tego celu uznane metryki, takie jak równość predykcyjna czy parytet demograficzny (Zuiderveen Borgesius, 2018). Takie podejście umożliwia bieżące monitorowanie działania modelu już po jego wdrożeniu i szybką reakcję na ewentualne problemy. Niezwykle istotne jest również odpowiednie szkolenie personelu zaangażowanego w projektowanie, wdrażanie i użytkowanie systemów AI. Pracownicy powinni dysponować wiedzą pozwalającą na rozpoznawanie ryzyka automatyzacji uprzedzeń i przeciwdziałanie mu, co minimalizuje ryzyko bezkrytycznego zaufania do technologii. Wyobraźmy sobie sytuację programu stypendialnego, gdzie decyzje o przyznaniu wsparcia podejmuje model językowy analizujący zgodność wniosków z kryteriami. W przypadku nierozstrzygniętych zgłoszeń, tendencyjny model mógłby, na przykład, nieświadomie faworyzować mężczyzn, opierając się na błędnym założeniu o ich niższej satysfakcji życiowej, co prowadziłoby do niesprawiedliwych decyzji. Współpraca interdyscyplinarna, łącząca ekspertyzę naukowców społecznych, prawników i specjalistów technicznych, może dodatkowo wesprzeć opracowanie skutecznych metod minimalizowania stronniczości, dostosowanych do specyfiki danego zastosowania.

Spis ilustracji

Spis tabel

Bibliografia

1. AlKhamissi, B., ElNokrashy, M., Alkhamissi, M. i Diab, M., 2024. *Investigating Cultural Alignment of Large Language Models,* Bangkok, Thailand: Association for Computational Linguistics.

2. Astekin, M., Hort, M. i Moonen, L., 2024. *An Exploratory Study on How Non-Determinism in Large Language Models Affects Log Parsing.* 2024 IEEE/ACM 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering (InteNSE), Lisbon, Portugal

3. Atil, B. i inni, 2024. *LLM Stability: A detailed analysis with some surprises,* Penn State University, Comcast AI Technologies.

4. Bailyn, E., 2025. *Firstpagesage.* [Online] Available at: https://firstpagesage.com/reports/top-generative-ai-chatbots/ [Data uzyskania dostępu: 28 Luty 2025].

5. Brockman, G., 2025. *X.* [Online] Available at: https://x.com/gdb/status/1878489681702310392 [Data uzyskania dostępu: 28 Luty 2025].

6. Bulmer, M. G., 1979. *Principles of Statistics.* Dover: University of Oxford.

7. Buyl, M. i inni, 2024. *Large Language Models Reflect the Ideology of their Creators,* brak miejsca: Ghent University, Belgium; Public University of Navarre, Spain.

8. Durmus, E. i inni, 2023. *Towards Measuring the Representation of Subjective Global Opinions in Language Models,* Anthropic.

9. Fang, X. i inni, 2024. Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports,* 14(5224), pp. 1-20.

10. Ferrer, X. i inni, 2021. *Bias and Discrimination in AI: A Cross-Disciplinary Perspective,* King's College London, United Kingdom.

11. Gemma Team, Google DeepMind, 2024. *Gemma 2: Improving Open Language Models at a Practical Size,* Google.

12. Gloaguen, T., Jovanovic, N., Staab, R. i Vechev, M., 2025. *Black-box detection of language models watermarks,* ETH ZURICH.

13. Glukhov, D. i inni, 2024. *LLM Censorship: The Problem and its Limitations,* University of Toronto & Vector Institute, University of Oxford.

14. Granato, J., Inglehart, R. i Leblang, D., 1996. The Effect of Cultural Values on Economic Development: Theory, Hypotheses, and Some Empirical Tests. *American Journal of Political Science,* 40(3), pp. 607-63.

15. Granberg, S. i Geiger, G., 2024. *Svenska Dagbladet.* [Online] Available at: https://www.svd.se/a/Rzmg9x/forsakringskassans-ai-for-vab-fusk-granskade-kvinnor-oftare
[Data uzyskania dostępu: 14 marzec 2025].

16. Haerpfer, C. i inni, 2024. *World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0.0.* Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. https://doi.org/10.14281/18241.24.

17. Hao, F., 2016. A Panel Regression Study on Multiple Predictors of Environmental Concern for 82 Countries Across Seven Years. *Social Science Quarterly,* 5(97), pp. 991-1004.

18. Harari, Y. N., 2025. *The Role of Education and Science in the Digital Age* [Wywiad] (17 Marzec 2025).

19. Heaven, W. D., 2020. *MIT Technology Review.* [Online] Available at: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/ [Data uzyskania dostępu: 14 kwiecień 2025].

20. Heinrichs, B., 2021. *Discrimination in the age of artificial intelligence.*

21. Hu, K., 2023. *Reuters.* [Online] Available at: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [Data uzyskania dostępu: 28 Luty 2025].

22. Hu, T. i inni, 2024. Generative Language Models Exhibit Social Identity Biases. *Nature Computational Science,* Tom 6, pp. 65-67.

23. Johnson, T. i Johnson, N., 2023. *SCIENTIFIC AMERICAN.* [Online] Available at: https://www.scientificamerican.com/article/police-facial-recognition-technology-cant-tell-black-people-apart/ [Data uzyskania dostępu: 14 kwiecień 2025].

24. Kayla Schroeder, Z. W.-D., 2025. *Can You Trust LLMJudgments? Reliability of LLM-as-a-Judge,* Northwestern University.

25. Kirk, H. R. i inni, 2021. *Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models,* Oxford Artificial Intelligence Society, University of Oxford.

26. Lapid, R., Langberg, R. i Sipper, M., 2024. *Open Sesame! Universal Black-Box Jailbreaking of Large Language Models,* Tel-Aviv: Applied Sciences.

27. Lin, C.-Y., 2004. *ROUGE: A Package for Automatic Evaluation of Summaries,* University of Southern California.

28. Matthew Renze, E. G., 2024. *The Effect of Sampling Temperature on Problem Solving in Large Language Models,* Johns Hopkins University.

29. Meister, N., Guestrin, C. i Hashimoto, T., 2024. *Benchmarking Distributional Alignment of Large Language Models,* Stanford University.

30. Minaee, S. i inni, 2024. *Large Language Models: A Survey*

31. Qu, Y. i Wang, J., 2024. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications,* 11(1095), pp. 1-13.

32. Rozado, D., 2024. *Thepolitical preferences of LLMs,* Otago Polytechnic.

33. Rudin, C. i Radin, J., 2019. *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition,* Harvard Data Science Review.

34. Sasha Shen Johfre, J. F., 2021. Reconsidering the Reference Category. *Sociological Methodology,* Issue 51, pp. 253 - 269.

35. Schroeder, K. i Wood-Doughty, Z., 2025. *Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge,* Department of Computer Science.

36. Schwartz, I., Link, K., Daneshjou, R. i Cortés-Penfield, N., 2024. *Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation,* Oxford University Press on behalf of Infectious Diseases Society of America.

37. Seguino, S., 2010. Help or Hindrance? Religion's Impact on Gender Inequality. *World Developmen,* 8(39), pp. 1308-1321.

38. Skelton, S. K., 2025. *TECHTARGET.* [Online] Available at: https://www.computerweekly.com/news/366619519/UK-police-forces-supercharging-racism-with-predictive-policing
[Data uzyskania dostępu: 14 kwiecień 2025].

39. Sozan, M., 2024. *Center for American Progress.* [Online] Available at: https://www.americanprogress.org/article/an-american-democracy-built-for-the-people-why-democracy-matters-and-how-to-make-it-work-for-the-21st-century/
[Data uzyskania dostępu: 11 kwiecień 2025].

40. Spennemann, D. H. R., 2023. *ChatGPT and the Generation of Digitally Born "Knowledge": How Does a Generative AI Language Model Interpret Cultural Heritage Values?,* School of Agricultural, Environmental and Veterinary Sciences, Charles Sturt University.

41. The Greenlining Institute, 2021. *How Automated Decision-Making Becomes Automated Discrimination,* The Greenlining Institute.

42. Zhao, S., Dang, J. i Grover, A., 2023. *Group Preference Optimization: Few-Shot Alignment of Large Language Models,* University of California.

43. Zollo, T. i inni, 2025. *Towards Effective Discrimination Testing for Generative AI,* Columbia University.

44. Zuiderveen Borgesius, F., 2018. *Discrimination, artificial intelligence, and algorithmic decision-making,* University of Amsterdam.

Streszczenie

W erze gwałtownego rozwoju sztucznej inteligencji, istotne jest również zrozumienie ich ukrytych tendencji i ograniczeń. Niniejsza praca bada problem stronniczości modeli językowych poprzez analizę odpowiedzi modelu Gemma 2 2B na pytania z pięciu obszarów tematycznych: zaufania do instytucji, gospodarki i ekonomii, nauki i technologii, wartości demokratycznych oraz zadowolenia z życia. Wykorzystując metodologię "czarnej skrzynki", porównano odpowiedzi modelu z danymi z World Values Survey (Wave 7) dla jedenastu krajów reprezentujących różne regiony świata. Badanie przeprowadzono w dwóch etapach. W pierwszym, na podstawie danych z World Values Survey opracowano typowe dla danej grupy społeczno-demograficznej odpowiedzi za pomocą modelu regresji. W drugim, model Gemma 2 2B "wcielił się" w przedstawicieli tych grup i udzielał odpowiedzi na te same pytania. Następnie, przeanalizowano różnice między przewidywanymi odpowiedziami z badania ankietowego z odpowiedziami modelu. Wyniki wskazują, że model różnicuje swoje odpowiedzi w zależności od płci w większości badanych obszarów, przy czym w trzech z pięciu kategorii kierunek różnic był zgodny z danymi empirycznymi. Zaobserwowano również, że model najlepiej odwzorowuje poglądy osób z krajów zachodnich oraz otrzymano dokładną charakterystykę grupy społecznej dla której odpowiedź modelu jest najbliższa uśrednionej opinii z badania ankietowego. Wbrew niektórym sugestiom z literatury, analiza nie potwierdziła, że model Gemma 2 2B jest wyraźnie stronniczy na rzecz poglądów amerykańskich, co podważa hipotezę o bezpośrednim odzwierciedlaniu poglądów twórców. Praca podsumowuje wyniki badania empirycznego i opisuje rekomendacje dla zastosowania modeli w szczególności przez instytucje.