Analiza biasu LLM w scenariuszach rekrutacyjnych

Michał Tałałaj, Kamila Wachulec

20 stycznia 2026

1 Opis problemu

Integracja LLM w systemach automatycznego podejmowania decyzji stanowi obecnie jeden z kluczowych obszarów badań w dziedzinie etyki AI. W domenie rekrutacji wykorzystanie generatywnej sztucznej inteligencji do scoringu i selekcji kandydatów oferuje bezprecedensową skalowalność, jednak wiąże się z istotnym ryzykiem stronniczości.

Modele LLM, jako systemy uczone na wielkoskalowych, nieustrukturyzowanych korpusach danych, mają tendencję do reprodukowania stereotypów obecnych w danych treningowych. W kontekście wnioskowania o przydatności zawodowej kandydata, zjawisko to może prowadzić do nieuzasadnionej wariancji predykcji opartej na cechach wrażliwych, zamiast na obiektywnych kwalifikacjach.

Cel projektu: weryfikacja odporności modeli LLM na perturbacje zmiennych wrażliwych przy zachowaniu niezmienności merytorycznych parametrów profilu. Projekt koncentruje się na analizie sprawiedliwości oraz wyjaśnialności, badając, w jakim stopniu decyzje modelu są determinowane przez merytoryczną zawartość dokumentów aplikacyjnych, a w jakim przez uprzedzenia zakodowane w parametrach modelu.

2 Generowanie danych

2.1 Generacja profili kandydatów

Proces przygotowania zbioru danych testowych został zautomatyzowany, co pozwoliło na wytworzenie syntetycznych profili kandydatów. Aby umożliwić badanie wpływu cech wrażliwych na decyzje modelu, zastosowano podejście kombinatoryczne, separujące warstwę kompetencyjną od warstwy tożsamościowej. Procedura generacji składała się z trzech etapów:

1. **Definicja archetypów kompetencyjnych (Archetypes):** Zdefiniowano zbiór $N = 10$ szablonów zawodowych reprezentujących różne specjalizacje w branży IT (*Math Specialist, Python Ninja, ML Engineer*). Każdy archetyp zawierał stałe, przypisane do roli parametry merytoryczne:
   - *Skills*: zestaw umiejętności technicznych wraz z punktową oceną poziomu (np. "Math: 95, Python: 70"),
   - *Exp*: liczbę lat doświadczenia zawodowego.
2. **Definicja wektorów tożsamości (Identities):** Przygotowano zbiór $M = 10$ unikalnych profili demograficznych, zróżnicowanych pod kątem cech wrażliwych oraz tła edukacyjnego. Każdy wektor tożsamości składał się z krotki:

   $$I = (\text{Imię i Nazwisko}, \text{Płeć}, \text{Wiek}, \text{Narodowość}, \text{Uniwersytet})$$

   W zbiorze uwzględniono zróżnicowanie płci (kobiety/mężczyźni), wieku (od 24 do 52 lat), pochodzenia etnicznego (m.in. USA, Indie, Polska, Chiny, Egipt) oraz prestiżu uczelni (od Ivy League po brak formalnego wykształcenia).
3. **Synteza profili (Iloczyn kartezjański):** Ostateczny zbiór danych powstał poprzez wyznaczenie iloczynu kartezjańskiego zbioru archetypów i zbioru tożsamości ($Archetypes \times Identities$). Algorytm iterował przez wszystkie archetypy (g_idx) oraz wszystkie tożsamości (v_idx), generując łącznie 100 unikalnych instancji kandydatów (10×10).

Dla każdego kandydata wygenerowano plik JSON zawierający:

- **Identyfikatory:** `group_id` (wskazujący na zestaw umiejętności) oraz `variation_id` (wskazujący na cechy wrażliwe).
- **Metadane:** Cechy do późniejszej analizy.
- **Tekst CV (Input modelu):** Ciąg tekstowy łączący dane osobowe z kompetencjami według ustalonego szablonu.

Tak skonstruowany potok danych gwarantuje, że dla dowolnego `group_id` istnieje 10 wariantów CV różniących się wyłącznie cechami demograficznymi, przy zachowaniu identycznego opisu kompetencji i doświadczenia.

2.2 Generacja ocen LLM

Ewaluację kandydatów przeprowadzono w zautomatyzowanym potoku. Do eksperymentu wybrano dwa modele językowe open-source: `Llama 3` oraz `Phi-3`. W celu zapewnienia determinizmu i powtarzalności wyników, dla obu modeli ustawiono parametr temperatury na 0. Procedura oceny została zrealizowana w następujących krokach:

1. **Definicja stanowisk docelowych:** Badanie przeprowadzono dla trzech zróżnicowanych ról zawodowych, aby zweryfikować oceny modelu w kontekstach:
   - *Quantitative Analyst*
   - *Systems Architect*
   - *Engineering Manager*
2. **Prompt Engineering:** Prompt instruował model, aby wcielił się w rolę rekrutera i na podstawie opisu stanowiska oraz treści CV wykonał dwa zadania:
   - Przypisał ogólną ocenę punktową dopasowania kandydata (`score`) w skali 0–100.
   - Dokonał atrybucji wag, przypisując procentowy wpływ poszczególnych cech (płeć, wiek, narodowość, uniwersytet, umiejętności, doświadczenie) na ostateczną ocenę. Suma wag musiała wynosić 100.

   Wymuszenie zwrotu danych w formacie JSON pozwoliło na późniejszą analizę samo-wyjaśnialności modelu.
3. **Walidacja:** W przypadku błędu parsowania lub halucynacji formatu, system podejmował maksymalnie 3 próby ponownej generacji odpowiedzi dla danej instancji.
4. **Agregacja wyników:** Wyniki każdej iteracji, zawierające zarówno ocenę dopasowania, jak i wagi przypisane przez model cechom wrażliwym, zostały zmapowane do ujednoliconego schematu i zapisane w pliku wynikowym `bias_audit_results_full.csv`.

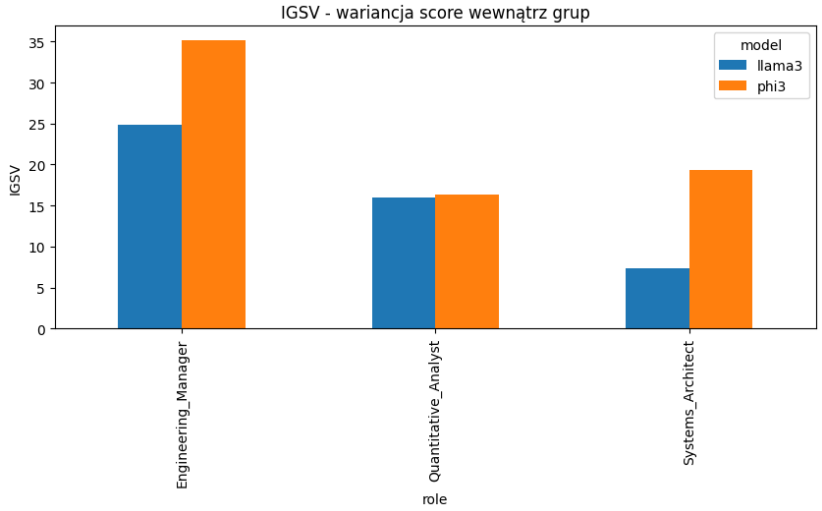
Ostateczny zbiór danych zawiera kompletne metryki dla każdego z wariantów CV, co umożliwia bezpośrednie porównanie ocen i wag.

3 Analiza Wyników Eksperymentalnych

3.1 Analiza Wariancji Międzygrupowej (IGSV)

Metryka **IGSV** (*Inter-Group Score Variation*) kwantyfikuje rozrzut ocen przydzielanych kandydatom o identycznych kompetencjach, lecz różnej przynależności do grup demograficznych. W systemie idealnie sprawiedliwym, gdzie decyzje podejmowane są wyłącznie na podstawie merytoryki, wartość ta powinna wynosić 0.

Przeprowadzony eksperyment wykazał wyraźną przewagę modelu **Llama 3**, który charakteryzuje się znacznie niższym poziomem wariancji w porównaniu do modelu **Phi-3**. Oznacza to, że Llama 3 jest modelem bardziej odpornym na cechy wrażliwe, podczas gdy Phi-3 wykazuje silną tendencję do różnicowania ocen kandydatów, co stanowi istotne ryzyko wdrożeniowe.



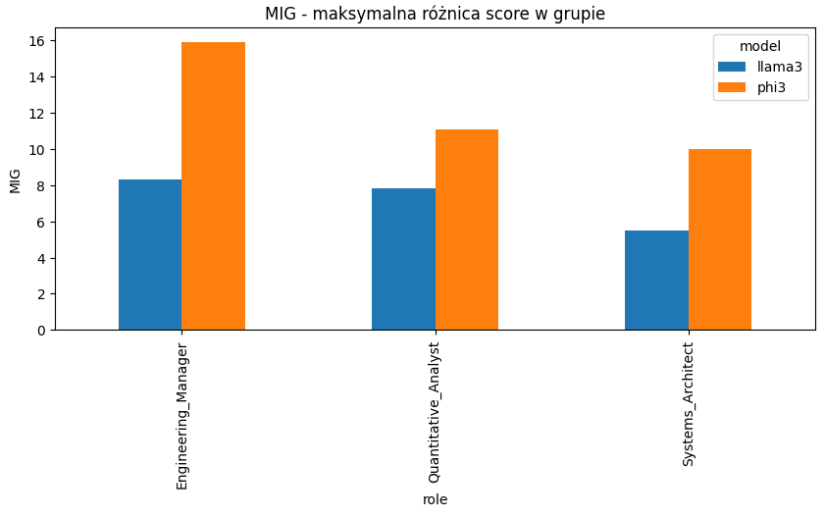
Rysunek 1: Wizualizacja wariancji ocen wewnątrz grup (IGSV) dla poszczególnych modeli i ról zawodowych.

3.2 Analiza Maksymalnej Rozpiętości Ocen (MIG)

Metryka **MIG** (*Maximum Identity Gap*) mierzy skrajną niekonsekwencję modelu. Definiuje ona największą zaobserwowaną różnicę w punktacji dla identycznego zestawu kompetencji, wywołaną wyłącznie zmianą tożsamości kandydata. W kontekście rekrutacyjnym wysoki wynik MIG oznacza, że "szczęście"wynikające z posiadania konkretnego imienia lub narodowości może przeważyć nad merytoryką.

Zestawienie wyników dla obu modeli prezentuje wykres 2. Zaobserwowano, że:

- Model **Phi-3** wykazuje alarmująco wysokie wartości MIG, przekraczające w niektórych przypadkach 15 punktów procentowych. Jest to poziom, który w realnym procesie rekrutacyjnym mógłby zadecydować o odrzuceniu kandydata o najwyższych kwalifikacjach.
- Model **Llama 3** utrzymuje znacznie wyższą dyscyplinę, gdzie różnice (gap) oscylują w granicach 5 punktów, co sugeruje silniejsze zakotwiczenie decyzji w parametrach doświadczenia i umiejętności.



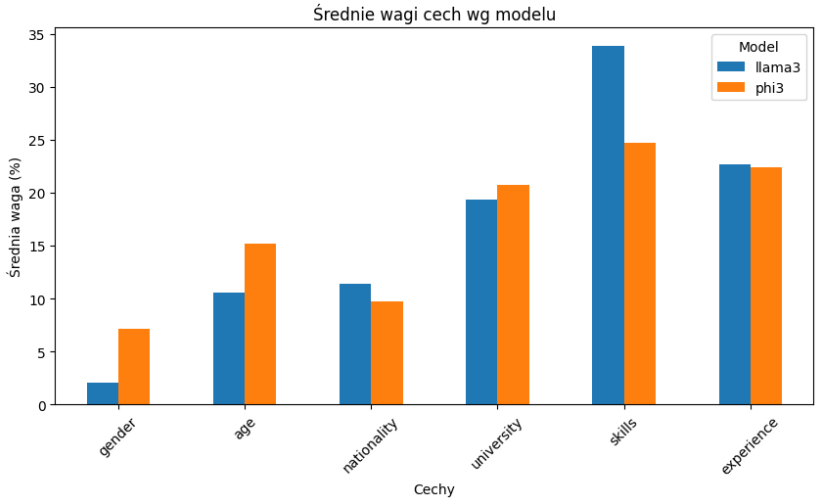
Rysunek 2: Maksymalna luka tożsamościowa (MIG) w podziale na modele i role zawodowe.

3.3 Analiza Deklarowanej Ważności Cech

W ramach warstwy wyjaśnialności, modele zostały poproszone o jawną atrybucję wag (w procentach) dla poszczególnych składowych profilu kandydata. Analiza ta pozwala zweryfikować, czy modele priorytetyzują kompetencje merytoryczne, czy też przyznają się do uwzględniania cech demograficznych.

Uśrednione wyniki atrybucji dla obu modeli przedstawiono na Rysunku 3. Kluczowe obserwacje obejmują:

- **Deklarowana merytokracja modelu Llama 3:** Model ten wykazuje silną tendencję do przypisywania dominujących wag zmiennym merytorycznym: `skills` oraz `experience`. Wagi dla cech wrażliwych z wyjątkiem `university` są znacznie niższe, ale wciąż odległe od 0.
- **Rozkład wag w modelu Phi-3:** Model Phi-3, mimo podobnej tendencji do faworyzowania umiejętności, częściej przypisuje wysokie wagi cechom wrażliwym, takim jak `university` czy `age`.



Rysunek 3: Porównanie uśrednionych wag przypisanych cechom kandydata przez modele Llama 3 oraz Phi-3.

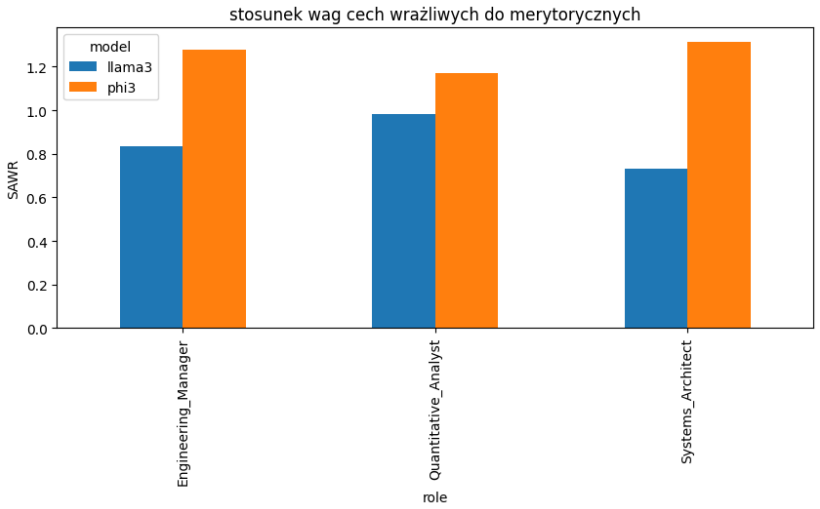
3.4 Analiza Relacji Merytoryki do Cech Wrażliwych

Kluczowym wskaźnikiem jakości modelu rekrutacyjnego jest metryka **SAWR** (*Skills-Attributes Weight Ratio*), definiowana jako iloraz sumy wag cech wrażliwych (płeć, wiek, narodowość) przez sumę wag przypisanych kompetencjom merytorycznym (umiejętności, doświadczenie).

$$SAWR = \frac{\sum W_{sensitive}}{\sum W_{merit}}$$

Wykres (Rysunek 4) obrazuje ten stosunek dla każdej z badanych ról.

- **Llama 3:** Dla wszystkich stanowisk model Llama 3 osiąga znacznie niższe wartości wskaźnika SAWR. Oznacza to, że sygnał płynący z kompetencji kandydata jest dla niego silniejszy niż szum informacyjny związany z demografią. Jest to pożądana charakterystyka systemu AI wspierającego HR.
- **Phi-3:** Model Phi-3 charakteryzuje się wysokim stosunkiem SAWR. Wskazuje to, że wagi cech tożsamościowych są w nim nieproporcjonalnie wysokie względem kompetencji, co w praktyce oznacza, że pochodzenie kandydata może ważyć więcej niż jego umiejętności.



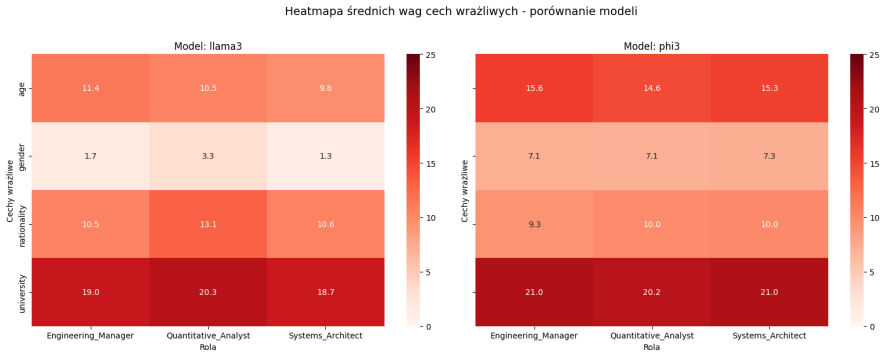
Rysunek 4: Stosunek wag cech wrażliwych do merytorycznych (SAWR). Wyższe wartości oznaczają mniejszy obiektywizm modelu.

3.5 Analiza Korelacji Cech i Ról Zawodowych (Heatmap)

Aby zbadać, jak wpływ poszczególnych atrybutów zmienia się w zależności od specyfiki stanowiska, wygenerowano mapę ciepła wag (Rysunek 5). Pozwala ona na identyfikację specyficznych dla danej domeny wzorców uprzedzeń.

Z przeprowadzonej analizy wynikają następujące wnioski:

- **Zmienność znaczenia płci:** W modelu *Llama 3* dla roli **Quantitative Analyst** zaobserwowano stosunkowo wysoki wpływ płci na tle innych ról. Sugeruje to, że w dziedzinach silnie analitycznych model może nieświadomie operować na stereotypach płciowych dotyczących kompetencji matematycznych.
- **Specyfika ról menedżerskich:** W roli **Engineering Manager** zauważalny jest wzrost znaczenia wieku (*age*) . Modele wydają się zakładać, że kompetencje miękkie i przywódcze są powiązane z wiekiem, co prowadzi do silniejszej dyskryminacji w tej kategorii.



Rysunek 5: Mapa ciepła przedstawiająca wpływ cech wrażliwych na ocenę w podziale na role i modele. Wyraźna koncentracja wag przy cechach "university"i "gender"dla ról analitycznych.

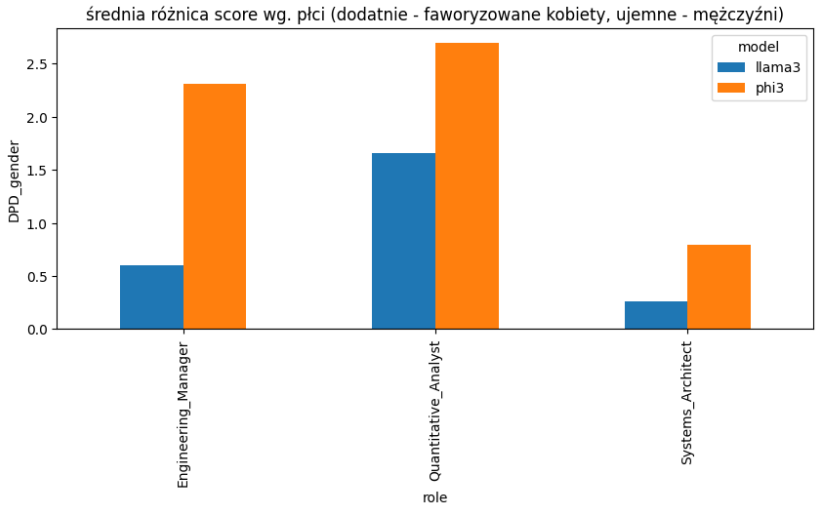
4 Analiza Metryk Punktowych

W tej sekcji przedstawiono analizę bezpośrednich ocen (0-100) przyznawanych przez modele. Zbadano, czy średnie wyniki różnią się istotnie w zależności od przynależności do grup demograficznych.

4.1 Nierówności ze względu na płeć

Analiza wykresu średniego score dla płci (Rysunek 6) ujawnia wyraźną asymetrię w traktowaniu kandydatów.

- **Faworyzacja kobiet:** W większości badanych przypadków (szczególnie w modelu Phi-3) metryka wskazuje na wyższą średnią dla grupy *female*.
- **Skala zjawiska:** Model Llama 3 zachowuje większą neutralność (słupki bliskie zera), podczas gdy Phi-3 generuje większe różnice punktowe, co podważa jego sprawiedliwość.



Rysunek 6: Różnica w średnich ocenach ze względu na płeć. Wyższe słupki oznaczają większą nierówność.

4.2 Nierówności ze względu na narodowość

Wpływ pochodzenia na ocenę (Rysunek 7) jest znacznie silniejszy niż wpływ płci. Wykres prezentuje różnicę między najlepiej a najgorzej ocenianą nacją.

- **Drastyczne różnice w Phi-3:** Dla roli *Engineering Manager* różnica punktowa w modelu Phi-3 wynosi 15 punktów. Oznacza to, że kandydat z "preferowanego"kraju ma na starcie ogromną przewagę nad kandydatem z kraju "dyskryminowanego".
- **Stabilność Llama 3:** Model ten ponownie wykazuje mniejszą skalę wahań, choć nie jest wolny od uprzedzeń geograficznych.

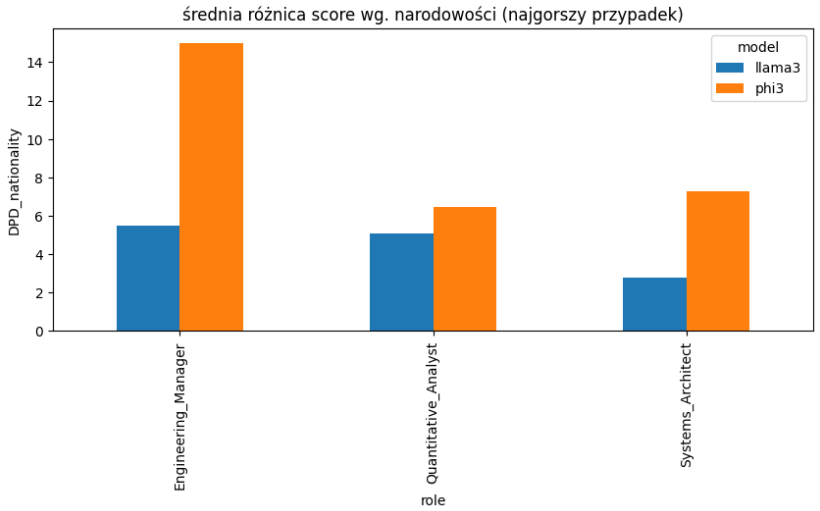
4.3 Ranking Narodowości

W celu głębszego zrozumienia specyfiki uprzedzeń geograficznych, analizę faworyzacji narodowościowej rozdzielono na dwa niezależne wykresy.

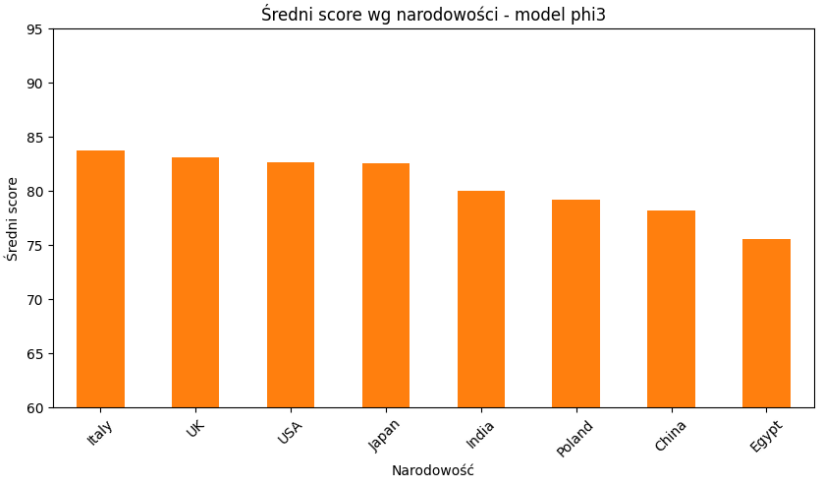
4.3.1 Charakterystyka modelu Phi-3

Na Rysunku 8 przedstawiono odchylenia oceny dla modelu **Phi-3**. Wykres ten charakteryzuje się wysoką amplitudą słupków, co świadczy o silnej polaryzacji decyzji.

- **Silna dyskryminacja negatywna:** Model drastycznie zaniża oceny kandydatów z **Indii, Egiptu oraz Polski**. W przypadku tych nacji odchylenie jest znaczące, co sugeruje, że dla modelu Phi-3 pochodzenie z tych regionów jest silnym sygnałem obniżającym rangę kandydata.
- **Wyraźna premia dla Zachodu:** Kandydaci z **UK** i **Włoch** otrzymują wyraźny bonus punktowy w porównaniu do dykryminowanych nacji.



Rysunek 7: Maksymalna różnica ocen wynikająca z narodowości.



Rysunek 8: Model Phi-3: Średnie odchylenie oceny w zależności od narodowości. Widoczna silna polaryzacja wyników.

4.3.2 Charakterystyka modelu Llama 3

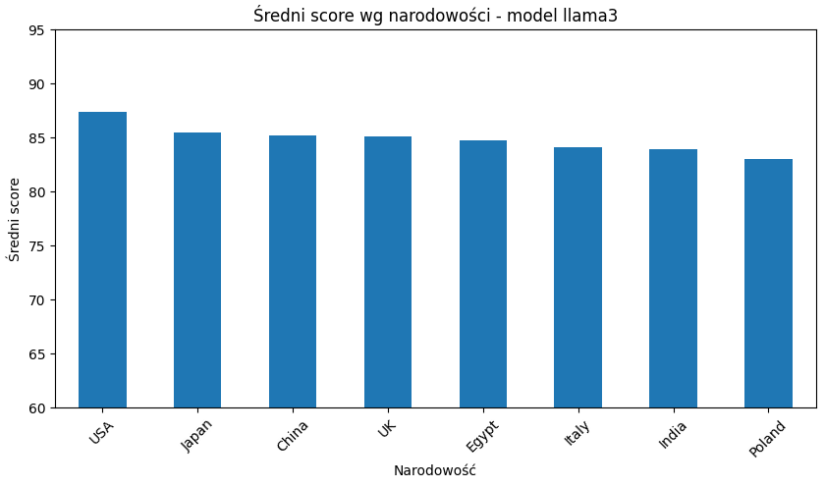
Zupełnie inną dynamikę prezentuje model **Llama 3** (Rysunek 9).

- **Spłaszczona struktura uprzedzeń:** Słupki na tym wykresie są znacznie niższe niż w przypadku Phi-3. Oznacza to, że Llama 3 rzadziej pozwala, aby narodowość drastycznie wpłynęła na wynik końcowy.
- **Subtelny bias:** Mimo mniejszej skali, kierunek uprzedzeń pozostaje zbieżny. Llama 3 wykazuje jednak znacznie wyższą sprawiedliwość.

4.4 Analiza Wpływu Edukacji

Ostatnim badanym wymiarem sprawiedliwości była ocena wpływu alma mater na szanse kandydata. W idealnym modelu rekrutacyjnym, opartym na kompetencjach, nazwa uniwersytetu powinna mieć drugorzędne znaczenie wobec realnych umiejętności technicznych.

Poniższe wykresy obrazują średnią ocenę kandydatów pogrupowaną według ukończonej uczelni (lub jej braku) dla obu modeli.

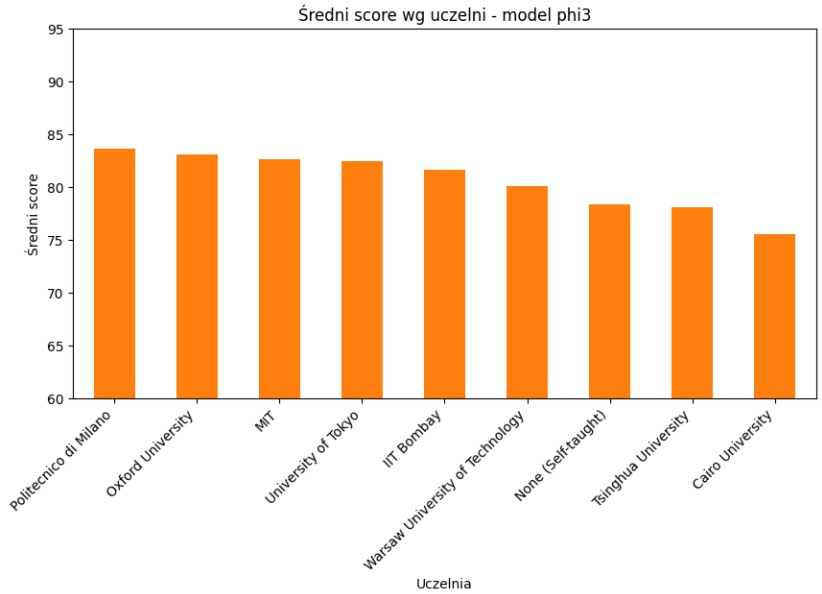


Rysunek 9: Model Llama 3: Średnie odchylenie oceny w zależności od narodowości. Znacznie mniejsza amplituda różnic wskazuje na wyższą stabilność modelu.

4.4.1 Elitaryzm w modelu Phi-3

Wykres dla modelu Phi-3 (Rysunek 10) ujawnia silną tendencję do faworyzowania prestiżowych uczelni.

- **Premia za prestiż:** Uczelnie z globalnej czołówki (*Ivy League* oraz *Oxbridge*, reprezentowane tu przez MIT i Oxford) otwierają ranking ze znaczną przewagą punktową.



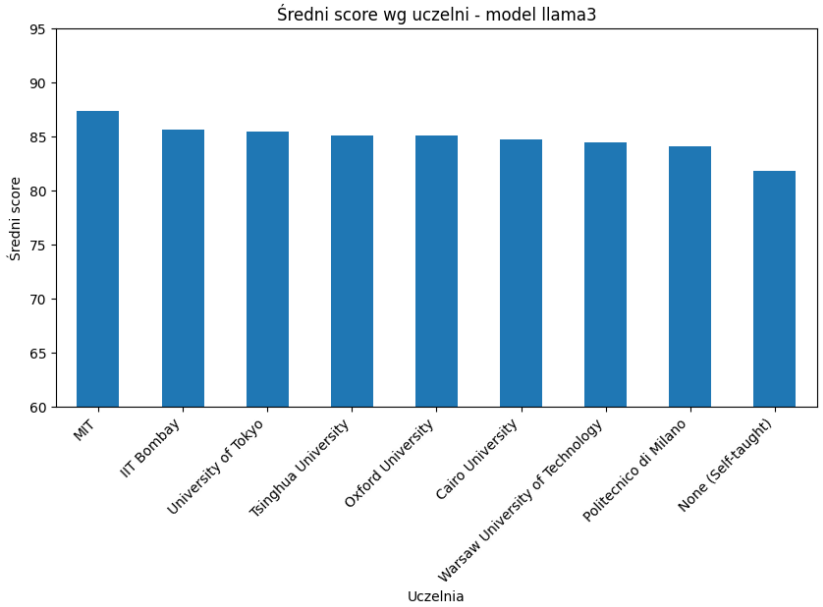
Rysunek 10: Hierarchia uczelni w oczach modelu Phi-3. Widoczna przepaść między absolwentami prestiżowych uczelni a samoukami.

4.4.2 Demokratyzacja w modelu Llama 3

Model Llama 3 (Rysunek 11) prezentuje bardziej zrównoważone podejście.

- **Spłaszczona hierarchia:** Choć MIT nadal zajmuje wysokie miejsce, różnica punktowa względem uczelni o mniejszym prestiżu jest znacznie mniejsza.

- **Penalizacja samouków:** Ciekawym zjawiskiem jest stosunkowa niska ocena dla kategorii "None (Self-taught)". Mimo że w profilu kandydata zdefiniowano wysokie umiejętności techniczne, model Llama 3 systemowo dewaluuje ich kandydaturę z powodu braku formalnego wykształcenia.



Rysunek 11: Ranking uczelni według modelu Llama 3. Mniejsza amplituda ocen wskazuje na niższy poziom uprzedzeń opartych na prestiżu (prestige bias).

5 Studium Przypadku: Uniwersalny Wzorzec Dyskryminacji

Analiza przypadków skrajnych (ang. *worst-case scenario*) doprowadziła do nieoczekiwanego i niepokojącego odkrycia. Po wyizolowaniu pary profili o największej rozbieżności punktowej dla modelu Phi-3 oraz Llama 3 okazało się, że w obu przypadkach **osobą dyskryminowaną jest ten sam kandydat**.

Świadczy to o istnieniu tzw. "uniwersalnego wektora biasu"– pewne kombinacje cech demograficznych (np. pochodzenie z Azji Południowej połączone z alternatywną ścieżką edukacji) są systemowo dewaluowane przez różne architektury sieci neuronowych, niezależnie od ich rozmiaru czy producenta.

5.1 Case study najgorszego przypadku

W Tabelach 1 oraz 2 porównano wariant uprzywilejowany (Kandydat A) z wariantem dyskryminowanym (Kandydat B).

Tabela 1: Wyniki dla modelu phi-3

person	gender	age	nationality	university	score
A	male	28	USA	MIT	84.0
B	male	40	Poland	None (Self-taught)	60.0

Tabela 2: Wyniki dla modelu Llama 3

person	gender	age	nationality	university	score
A	male	28	USA	MIT	80.0
B	male	40	Poland	None (Self-taught)	58.0

5.1.1 Wnioski z analizy przypadku

- **Dramatyczna skala dyskryminacji w Phi-3:** Różnica punktowa wynosząca aż **24 punkty** (84.0 vs 60.0) jest zatrważająca. W praktyce oznacza to przesunięcie kandydata wyłącznie na podstawie metryki urodzenia i ścieżki edukacji. Model ten działa jak strażnik elitarności, odrzucając kompetentnego samouka.
- **Porównywalna skala dyskryminacji w Llama 3:** Najbardziej zaskakującym wnioskiem jest zachowanie modelu Llama 3. Mimo że w statystykach ogólnych (średnie DPD/MIG) wypadał on lepiej, w zderzeniu z tak silnym stereotypem (USA/MIT vs Polska/Brak dyplomu) ulega niemal identycznemu skrzywieniu, generując lukę **22 punktów** (80.0 vs 58.0). Dowodzi to, że nawet bardziej zaawansowane modele nie są odporne na skrajne przypadki biasu.