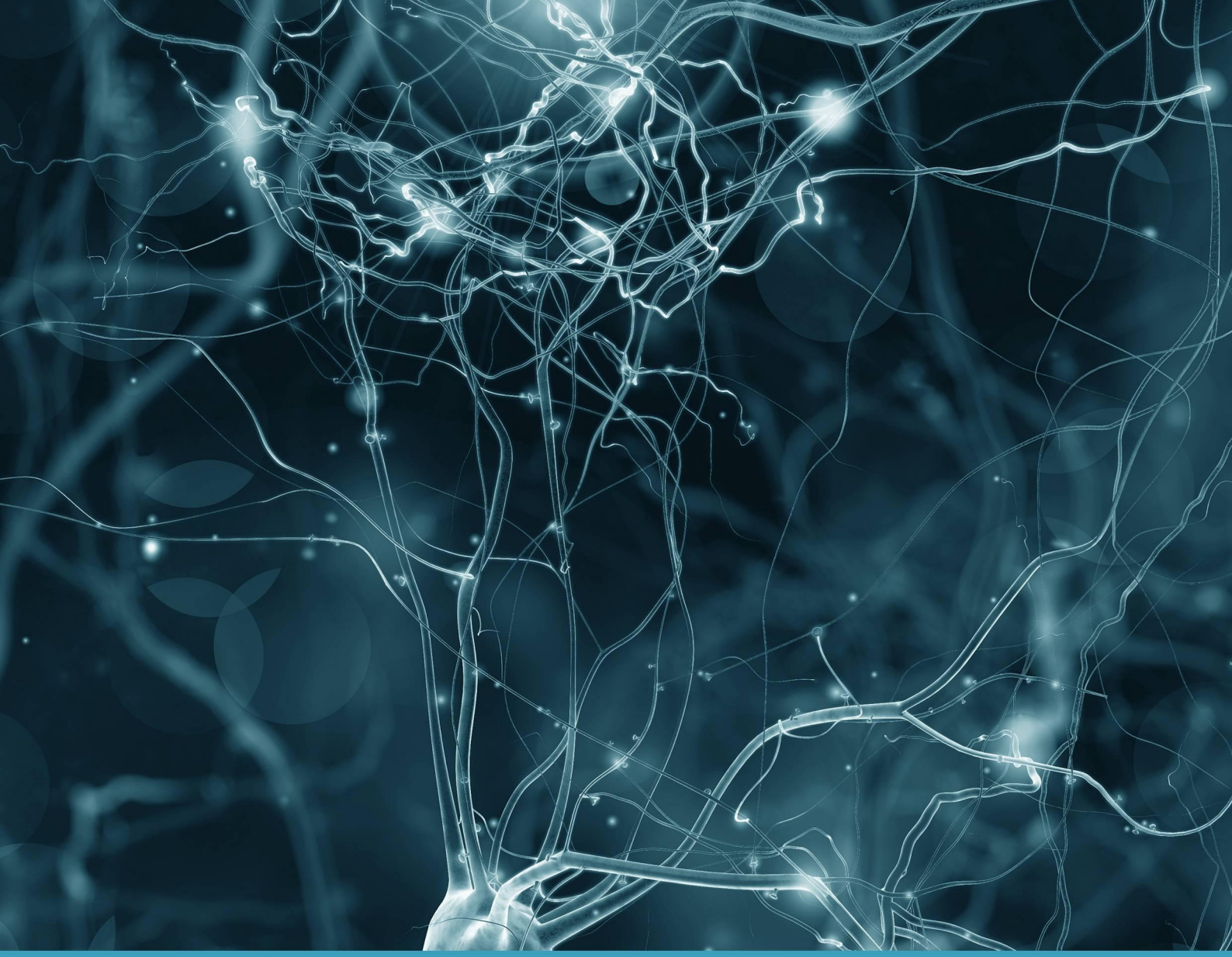


AGiDOT

Sztuczna inteligencja w biznesie

Stawiamy na transformację biznesu poprzez wykorzystanie sztucznej inteligencji

Równanie Płci 2

Bias płciowy w modelach AI — tym razem w liczbach. 8 baterii testów, 18 modeli, dane oficjalne.

Pierwszy raport „Równanie Płci" był zapowiedzią: na kilku przykładach pokazał, że modele językowe kodują stereotypy płciowe, i zakończył się apelem o dalsze badania. **To jest ta kontynuacja — oparta już nie na przykładach, lecz na twardych, mierzalnych danych.**

Przeprowadziliśmy **osiem baterii testów** na **18 modelach AI** — od klasycznych osadzeń słów sprzed lat, przez nowoczesne modele polskie i chmurowe, po duże modele czatowe. Zmierzyliśmy siłę uprzedzeń liczbą (test WEAT), zestawiliśmy je z **oficjalnymi danymi Eurostat** o realnym udziale kobiet w zawodach, sprawdziliśmy uprzedzenia w generowaniu tekstu i — co najważniejsze dla firm — **przetestowaliśmy, czy AI oceniające CV przyznaje te same punkty identycznym kwalifikacjom niezależnie od płci kandydata.**

Wyniki bywają zaskakujące. Największym mitem, który obalamy, jest przekonanie, że „nowszy i większy model = mniej uprzedzeń".

AGIDOT · 2026-07-28

Spis treści

Wprowadzenie

Abstrakt

Modele sztucznej inteligencji uczą się z ogromnych zbiorów tekstu — a wraz z językiem przejmują ukryte w nim stereotypy. Pierwszy raport „Równanie Płci" pokazał to zjawisko na pojedynczych przykładach arytmetyki na embeddingach (słynne **„praca − mężczyzna + kobieta"**) i zakończył się apelem o dalsze, systematyczne badania.

Ten raport odpowiada na ten apel **liczbami**. Zamiast kilku ręcznych przykładów przeprowadziliśmy **osiem standaryzowanych baterii testów (E1–E8)** na **18 modelach**: sześciu nowoczesnych modelach embeddingowych (polskich, wielojęzycznych i chmurowych), jednym klasycznym modelu sprzed lat oraz jedenastu modelach czatowych — od małych 3-miliardowych po duży model klasy GPT-4o. Każdy test jest powtarzalny, deterministyczny i zakończony konkretną liczbą.

Co nowego względem raportu 1

- **Miara zamiast wrażenia.** Siłę uprzedzeń mierzymy testem WEAT (Word Embedding Association Test) — dającym wielkość efektu i istotność statystyczną, a nie tylko „widać, że model kojarzy".
- **Porównanie wielu modeli.** Najmocniejszy element: ten sam test na kilkunastu modelach naraz, ustawionych w ranking uprzedzeń.
- **Zderzenie z rzeczywistością.** Uprzedzenia modelu zestawiamy z **oficjalnymi danymi Eurostat** o realnym udziale kobiet w zawodach.
- **Poza embeddingami.** Sprawdzamy uprzedzenia także w **generowaniu tekstu** przez chatboty i w **automatycznej ocenie CV** — czyli tam, gdzie AI już dziś podejmuje decyzje dotyczące ludzi.
- **Dowód, że da się to naprawić.** Pokazujemy, że uprzedzenie można nie tylko zmierzyć, ale i zredukować.

Dlaczego to ważne dla firm

AI coraz częściej wspiera **rekrutację**, selekcję CV, personalizację treści i analizę kandydatów. Jeśli model nosi w sobie nieuświadomione uprzedzenia, przenosi je wprost na decyzje dotyczące realnych ludzi — często niezauważenie, bo „to przecież tylko algorytm". Ten raport pokazuje, gdzie dokładnie te uprzedzenia się kryją, jak duże są w konkretnych modelach i co z nimi zrobić.

Jak czytać ten dokument

Piszemy popularnonaukowo — kolejne rozdziały tłumaczą pojęcia (embeddingi, oś płci, WEAT) prostym językiem, zanim pokażą liczby. Każdy rozdział kończy się praktycznym wnioskiem. Osoby techniczne znajdą metodykę i pełne wyniki w rozdziale końcowym oraz w otwartym repozytorium projektu.

> *Ważne zastrzeżenie. Bias modelu to nie „zła wola" ani intencja maszyny — to statystyczne odbicie wzorców obecnych w danych treningowych. Mówiąc „model jest stronniczy", opisujemy mierzalną właściwość, a nie przypisujemy maszynie poglądów.*

Jak w ogóle mierzy się uprzedzenia AI

Słowa jako liczby: embeddingi

Model językowy nie „rozumie" słów tak jak człowiek. Zamienia każde słowo na ciąg liczb — **wektor**, zwany embeddingiem — tak dobrany, by słowa o podobnym znaczeniu miały podobne wektory. Dzięki temu „lekarz" i „szpital" leżą blisko siebie, a „lekarz" i „banan" — daleko.

Ta reprezentacja ma zaskakującą własność: pozwala na **arytmetykę na znaczeniach**. To stąd wziął się przykład z raportu 1 — „praca − mężczyzna + kobieta" prowadzi do innego rejonu przestrzeni niż „praca − kobieta + mężczyzna". Jeśli model nauczył się, że pewne zawody są „męskie", a inne „żeńskie", widać to w geometrii jego wektorów.

Oś płci

Skoro słowa są punktami w przestrzeni, możemy wyznaczyć w niej **kierunek płci**. Bierzemy pary definicyjne — on/ona, mężczyzna/kobieta, ojciec/matka, brat/siostra — i uśredniamy różnicę między wektorem męskim a żeńskim. Powstaje strzałka: jeden jej koniec to „biegun męski", drugi „żeński".

Teraz dowolne słowo — na przykład nazwę zawodu — możemy **rzutować** na tę oś i odczytać liczbę: jak bardzo „męskie" albo „żeńskie" jest to słowo w oczach modelu. To ilościowa wersja intuicji z raportu 1.

WEAT — termometr uprzedzeń

Najważniejszym narzędziem raportu jest **WEAT** (Word Embedding Association Test), uznana metoda z badań nad uczciwością AI. Działa jak test skojarzeń: sprawdza, czy model **silniej kojarzy** jedną grupę słów (np. męskie) z jednym zbiorem pojęć (np. kariera, nauka), a drugą grupę (żeńskie) z innym (rodzina, sztuka).

Wynik WEAT to jedna liczba — **wielkość efektu** *d*:

- $d \approx 0$ — brak uprzedzenia (model kojarzy obie płcie tak samo),
- $d \approx 0{,}5$ — efekt średni,
- $d \approx 0{,}8$ i więcej — efekt duży,
- znak mówi o kierunku: dodatni *d* = klasyczny stereotyp (męskie ↔ kariera/ nauka/ kompetencja, żeńskie ↔ rodzina/sztuka/ciepło).

Dodatkowo liczymy **istotność statystyczną** (test permutacyjny) — czy wynik nie jest dziełem przypadku. Efekt oznaczony gwiazdką (*) to $p < 0{,}05$, czyli wynik, któremu można zaufać.

Trzy pytania, które zadajemy każdemu modelowi

We wszystkich testach embeddingowych używamy trzech klasycznych „baterii" skojarzeń, zaadaptowanych na język polski:

1. **Kariera vs Rodzina** — czy męskie słowa kojarzą się z pracą, a żeńskie z domem.
2. **Nauka vs Sztuka** — czy męskie z naukami ścisłymi, żeńskie z humanistyką.
3. **Kompetencja vs Ciepło** — czy mężczyzna kojarzy się ze sprawczością i decyzyjnością, a kobieta z opiekuńczością i empatią.

Co i jak testujemy

Osiem baterii testów pokrywa trzy warstwy, w których uprzedzenie może się ujawnić:

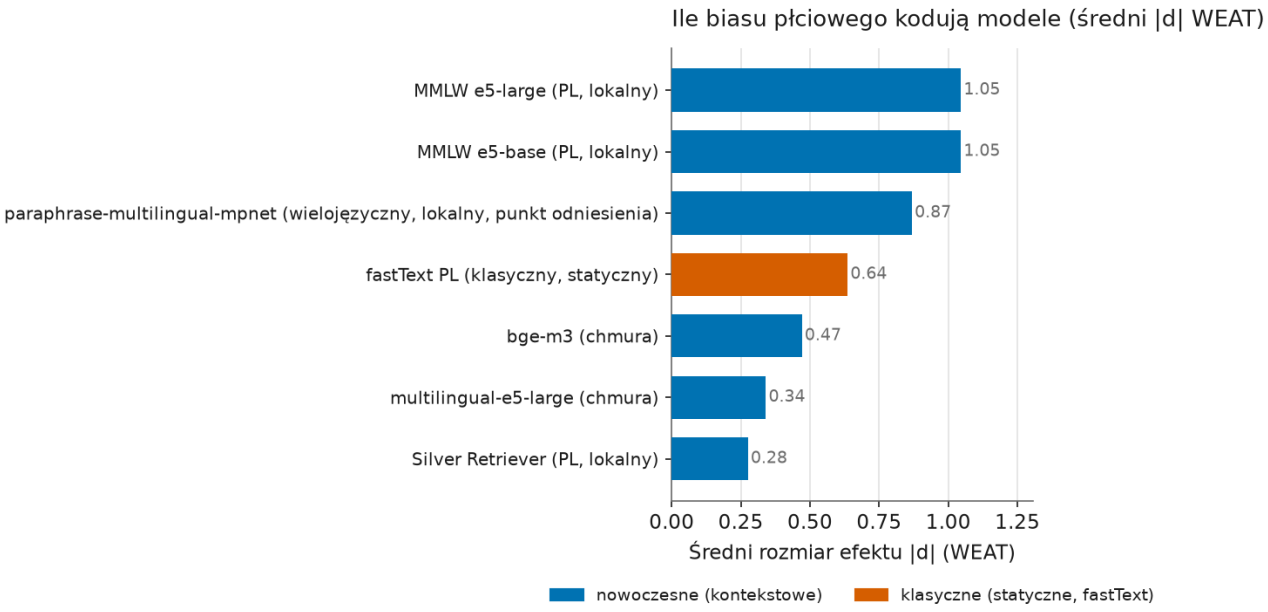
- **W statycznych embeddingach** (E1–E5, E7) — geometria skojarzeń między słowami.
- **W generowaniu tekstu** (E6) — jaką formę rodzajową wybiera chatbot, opisując zawód.
- **W decyzjach o ludziach** (E8) — jak model ocenia to samo CV w zależności od płci kandydata.

Wszystkie biegi są deterministyczne (ten sam wynik przy każdym uruchomieniu) i zapisane, aby każdy mógł je odtworzyć.

Ile uprzedzeń mają modele — i który jest najgorszy

Ranking uprzedzeń

To jest główny wynik raportu. Przepuściliśmy trzy baterie WEAT przez **siedem modeli embeddingowych** i uszeregowaliśmy je według średniej wielkości efektu *|d|* — im wyżej, tym silniejsze uprzedzenia płciowe.



Wynik jest przewrotny. **Najbardziej stronnicze okazały się nowoczesne, wyspecjalizowane modele polskie MMLW** (średnie *|d|* ≈ 1,05) — i to one dają efekty istotne statystycznie:

- MMLW w baterii **Nauka vs Sztuka**: *d* = +1,47 (*), efekt bardzo duży,
- MMLW w baterii **Kompetencja vs Ciepło**: *d* = +1,45 (*).

Dla porównania duże, chmurowe modele wielojęzyczne (multilingual-e5-large, bge-m3) mają uprzedzenia **słabsze i nieistotne statystycznie** (*|d|* 0,34–0,47).

Dlaczego lepszy model bywa bardziej stronniczy

To nie błąd — to skutek specjalizacji. Modele MMLW są trenowane głównie na **polskim** tekście i lepiej rozumieją niuanse naszego języka. Wraz z tą biegłością wchłaniają jednak

także **polskie stereotypy** obecne w danych. Model „gorszy" w polszczyźnie (np. wielojęzyczny gigant chmurowy) po prostu słabiej wyłapuje te zależności — także te stereotypowe.

Uwaga ważna dla praktyki: **nie ma prostej zależności „większy/nowszy = mniej stronniczy".** Silver Retriever, również polski model, ma najniższe uprzedzenia w całej stawce (0,28). O sile uprzedzeń decyduje sposób trenowania i dane, a nie sam rozmiar czy data premiery.

Stare kontra nowe

Do porównania dołączyliśmy **klasyczny model fastText** — statyczne osadzenia słów w stylu sprzed lat, reprezentujące „starą szkołę" AI (na wykresie zaznaczone innym kolorem). Gdyby postęp automatycznie usuwał uprzedzenia, klasyk powinien wypaść najgorzej.

Tak się nie stało. fastText plasuje się **w środku stawki** (|*d*| = 0,64) — poniżej najnowszych modeli MMLW. A w baterii Nauka vs Sztuka jego uprzedzenie (*d* = +1,40, *) jest praktycznie **na poziomie najnowszych modeli polskich**.

To podwójny cios w mit „nowsze = lepsze":

1. przez lata rozwoju uprzedzenie w wymiarze nauka/sztuka **nie zniknęło**,
2. najnowsze wyspecjalizowane modele polskie są *bardziej* stronnicze niż wiekowy klasyk.

Ciekawostka na marginesie: w klasycznym fastText brakowało w ogóle kilku słów — świeżych **feminatywów** jak „górniczka", „strażaczka" czy „weterynarka". W czasach, gdy powstawał ten model, te formy nie były jeszcze w powszechnym użyciu. Sam ich brak to mały pomnik tego, jak szybko zmienia się język.

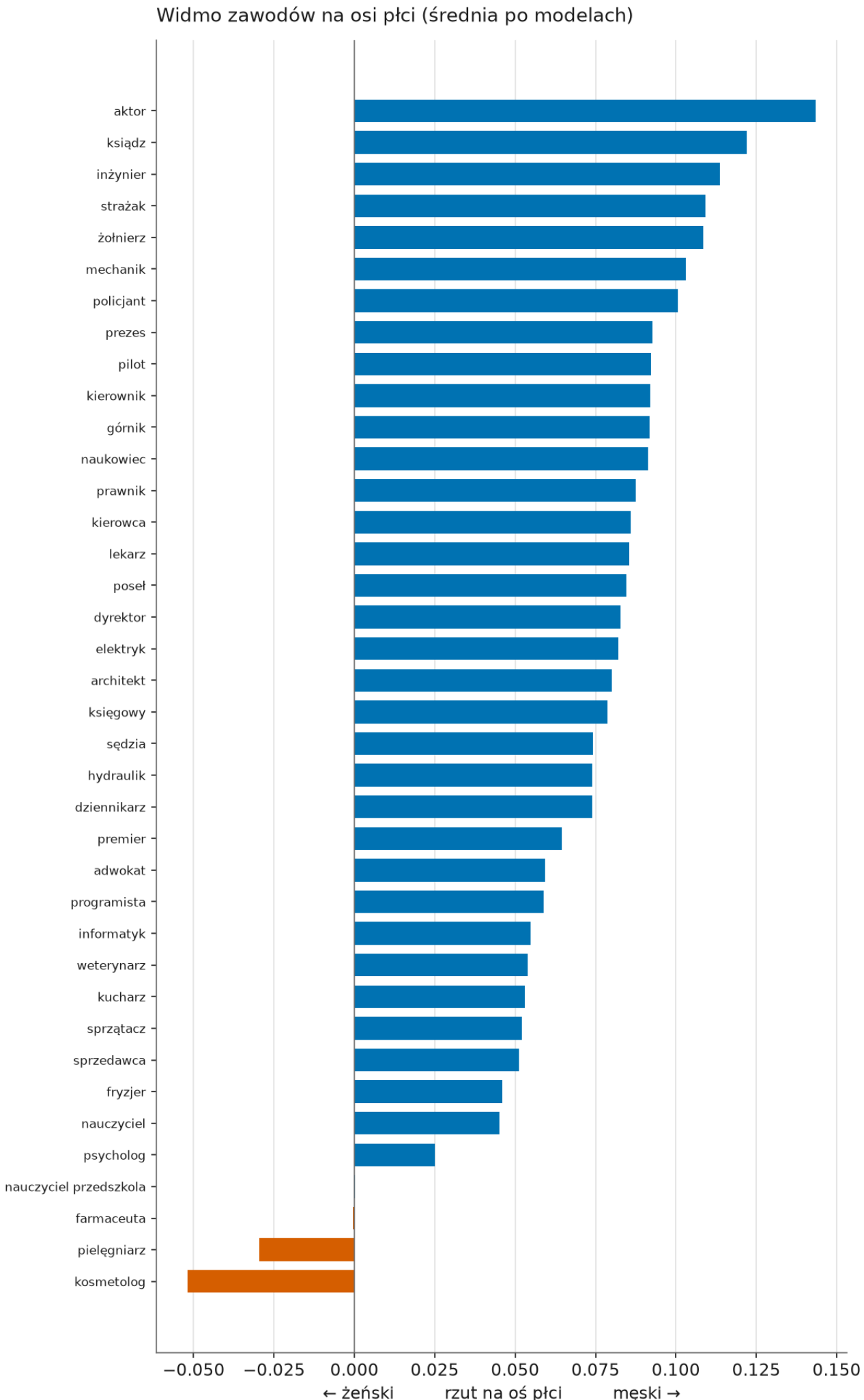
Co to znaczy w praktyce

Wybierając model do zastosowania wrażliwego na płeć (rekrutacja, analiza kandydatów, personalizacja), **nie kieruj się nowością ani rozmiarem**. Zmierz uprzedzenia konkretnego modelu na własnym zadaniu — różnice między modelami są czterokrotne (0,28 vs 1,05) i nieprzewidywalne z samej metryki.

Widmo zawodów: od „męskich" do „żeńskich"

Zawody na osi płci

Skoro potrafimy wyznaczyć oś płci (rozdział 2), możemy zrzutować na nią listę zawodów i zobaczyć, jak model je rozkłada — od najbardziej „męskich" po najbardziej „żeńskie". Poniższe widmo to średnia z wszystkich testowanych modeli.


Widmo zawodów na osi płci (średnia po modelach)
aktor
ksiądz
inżynier
strażak
żołnierz
mechanik
policjant
prezes
pilot
kierownik
górnik
naukowiec
prawnik
kierowca
lekarz
poseł
dyrektor
elektryk
architekt
księgowy
sędzia
hydraulik
dziennikarz
premier
adwokat
programista
informatyk
weterynarz
kucharz
sprzątacz
sprzedawca
fryzjer
nauczyciel
psycholog
nauczyciel przedszkola
farmaceuta
pielęgniarz
kosmetolog
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
0.125
0.150
← żeński
rzut na oś płci
męski →

Rozkład jest zgodny z potocznymi stereotypami — co samo w sobie jest wynikiem: model odtwarza społeczne wyobrażenia o „męskości" i „żeńskości" zawodów.

- **Biegun męski:** inżynier, ksiądz, żołnierz, strażak, górnik, prezes, mechanik, pilot, policjant.
- **Biegun żeński:** pielęgniarz, kosmetolog, nauczyciel przedszkola, farmaceuta.

Na samym szczycie „męskości" pojawia się **aktor** — to artefakt formy gramatycznej (męskoosobowa forma bazowa), dobry przykład, jak sam język wpływa na pomiar.

Analogie: echo przykładu z raportu 1

W raporcie 1 pokazaliśmy operację „zawód − mężczyzna + kobieta". Tutaj zbadaliśmy ją systematycznie: dla każdego zawodu sprawdziliśmy, czy taka analogia „przeskakuje" ku żeńskiej formie zawodu.

Efekt jest realny, ale słabszy i bardziej chwiejny niż w klasycznych embeddingach word2vec sprzed lat — nowoczesne modele kontekstowe budują przestrzeń znaczeń inaczej. Najsilniej ku formie żeńskiej przesuwa się **aktor → aktorka** (+0,059); dla większości zawodów przesunięcie jest niewielkie. To ważna lekcja metodyczna: **efektowna arytmetyka na wektorach działa najlepiej w prostych, statycznych modelach** i nie należy jej nadinterpretować w nowych.

Co to znaczy w praktyce

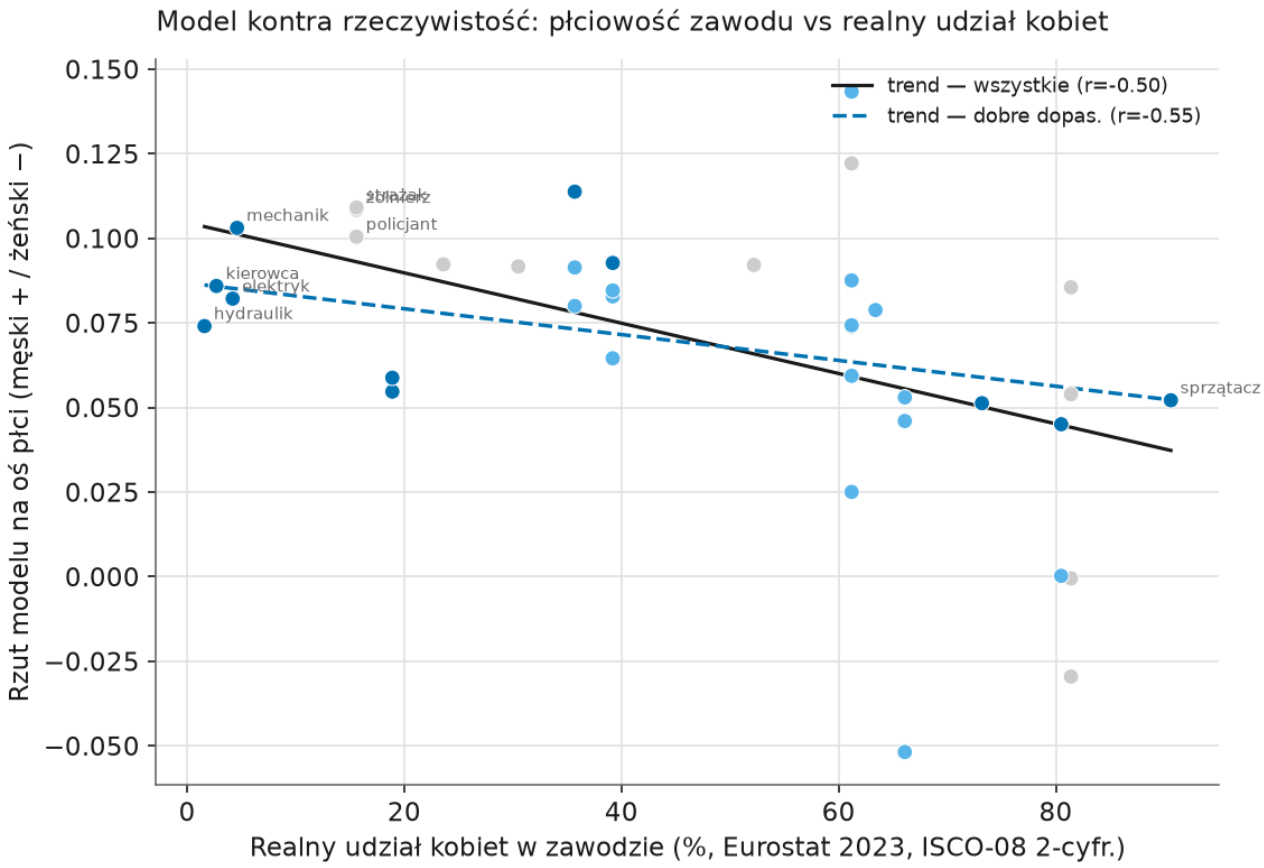
Model nosi w sobie gotową „mapę" tego, które zawody są męskie, a które żeńskie. Gdy taki model wspiera np. targetowanie ogłoszeń o pracę albo podpowiada ilustracje do stanowisk, ta mapa może **cicho zawężać** grono odbiorców — pokazując ofertę dla inżyniera głównie mężczyznom, a dla pielęgniarki głównie kobietom. Widmo z tego rozdziału to dokładnie ta mapa, wydobyta na światło.

Model kontra rzeczywistość

Czy model tylko odbija świat, czy go zniekształca

Naturalne pytanie: skoro model uważa strażaka za „męski", a pielęgniarkę za „żeńską" — może po prostu **wiernie odbija rzeczywistość**? Przecież w tych zawodach faktycznie pracuje więcej mężczyzn lub kobiet.

Żeby to sprawdzić, zestawiliśmy „płciowość" zawodu według modelu (rzut na oś płci z poprzedniego rozdziału) z **realnym udziałem kobiet** w tym zawodzie. Tym razem nie z przybliżenia, lecz z **oficjalnych danych Eurostatu** (badanie aktywności ekonomicznej BAEL/LFS, Polska, 2023, klasyfikacja zawodów ISCO-08).



Wynik: model naprawdę śledzi rzeczywistość

Korelacja jest wyraźna i we właściwym kierunku — im więcej kobiet realnie pracuje w zawodzie, tym bardziej „żeński" jest ten zawód dla modelu:

Zbiór zawodów	Korelacja (Pearson *r*)
Wszystkie zawody	**−0,50**
Zawody dobrze dopasowane do danych (test odporności)	**−0,55**

Druga liczba wymaga słowa wyjaśnienia. Oficjalne dane Eurostatu są dostępne na poziomie **grup zawodów**, nie pojedynczych stanowisk — dlatego strażak, policjant i żołnierz trafiają do jednej wspólnej wartości, a ksiądz do grupy „zawodów prawnych, społecznych i kulturalnych". Dla części zawodów to grube przybliżenie. Gdy jednak ograniczymy analizę do zawodów, które **dobrze** odpowiadają swojej grupie, **korelacja rośnie** (−0,55) — dowód, że sygnał jest realny, a szum bierze się z gruboziarnistości statystyki, nie z modelu.

Odbicie, nie wyrocznia — ale odbicie utrwalające

Model odzwierciedla realny rozkład płci w zawodach. To brzmi neutralnie, ale ma poważną konsekwencję: **AI utrwala stan obecny.** Jeśli dziś w zawodzie inżyniera jest 20% kobiet, model traktuje „inżyniera" jako męskiego — i gdy wesprze rekrutację czy doradztwo zawodowe, będzie tę proporcję **reprodukować**, zamiast dać równe szanse. Uprzedzenie nie musi „wyolbrzymiać" rzeczywistości, by szkodzić — wystarczy, że ją zamraża.

> *Uczciwie o danych. Eurostat/LFS podaje feminizację na poziomie 2-cyfrowej klasyfikacji ISCO-08 (37 grup), więc konkretny zawód dziedziczy wartość swojej grupy. Ograniczenie to adresujemy jawnie — oceną dopasowania każdego zawodu i testem odporności na podzbiorze dobrze dopasowanych. Pełną precyzję (poziom 4-cyfrowy) daje badanie GUS „Struktura wynagrodzeń według zawodów" — to naturalny kolejny krok.*

Co to znaczy w praktyce

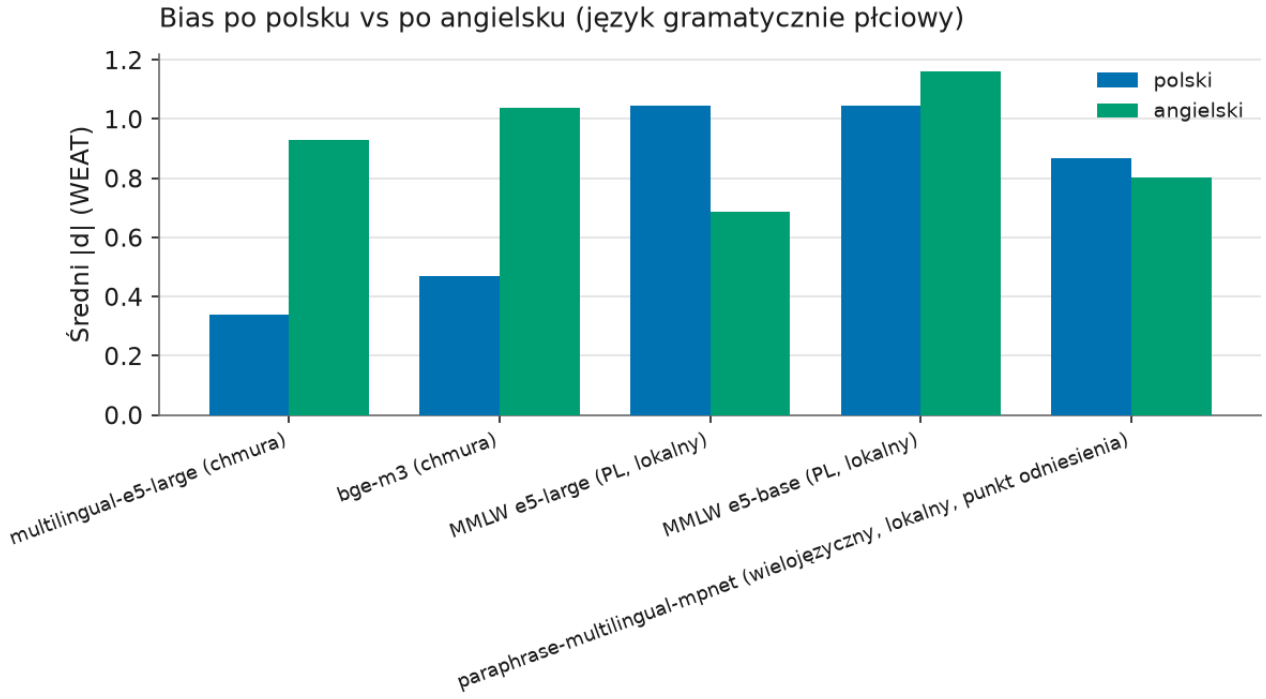
AI wykorzystywana w procesach kadrowych **nie jest neutralnym lustrem** — jest lustrem, które utrwala. Jeśli celem organizacji jest wyrównywanie szans, model zostawiony sam sobie działa przeciwko temu celowi, bo z definicji odtwarza istniejące proporcje. To argument nie za rezygnacją z AI, lecz za świadomym jej korygowaniem (rozdział 9).

Polski kontra angielski

Język gramatycznie płciowy

Polszczyzna jest „przesiąknięta" płcią bardziej niż angielski: lekarz kontra lekarka, inżynier kontra inżynierka, a nawet czasowniki w czasie przeszłym zdradzają rodzaj (zrobił / zrobiła). Angielskie „doctor" czy „engineer" są rodzajowo neutralne. Czy to znaczy, że polskie modele muszą kodować więcej uprzedzeń?

Sprawdziliśmy to, uruchamiając te same testy WEAT na **listach polskich i angielskich**, dla modeli wielojęzycznych.



Odpowiedź: to zależy od modelu

Wbrew prostej intuicji nie ma jednej reguły — **kierunek zależy od tego, na czym model był trenowany**:

- **Modele chmurowe, wielojęzyczne** (multilingual-e5-large, bge-m3) mają uprzedzenia **silniejsze po angielsku**. Przykładowo bateria Kompetencja vs Ciepło dla e5-large: *d* = +1,35 (*) po angielsku wobec +0,68 (nieistotne) po polsku. Te modele „myślą" głównie po angielsku i tam ich stereotypy są najostrzejsze.
- **Polski model MMLW** — odwrotnie: uprzedzenia **silniejsze po polsku**, zgodnie ze swoją specjalizacją.

Co to znaczy w praktyce

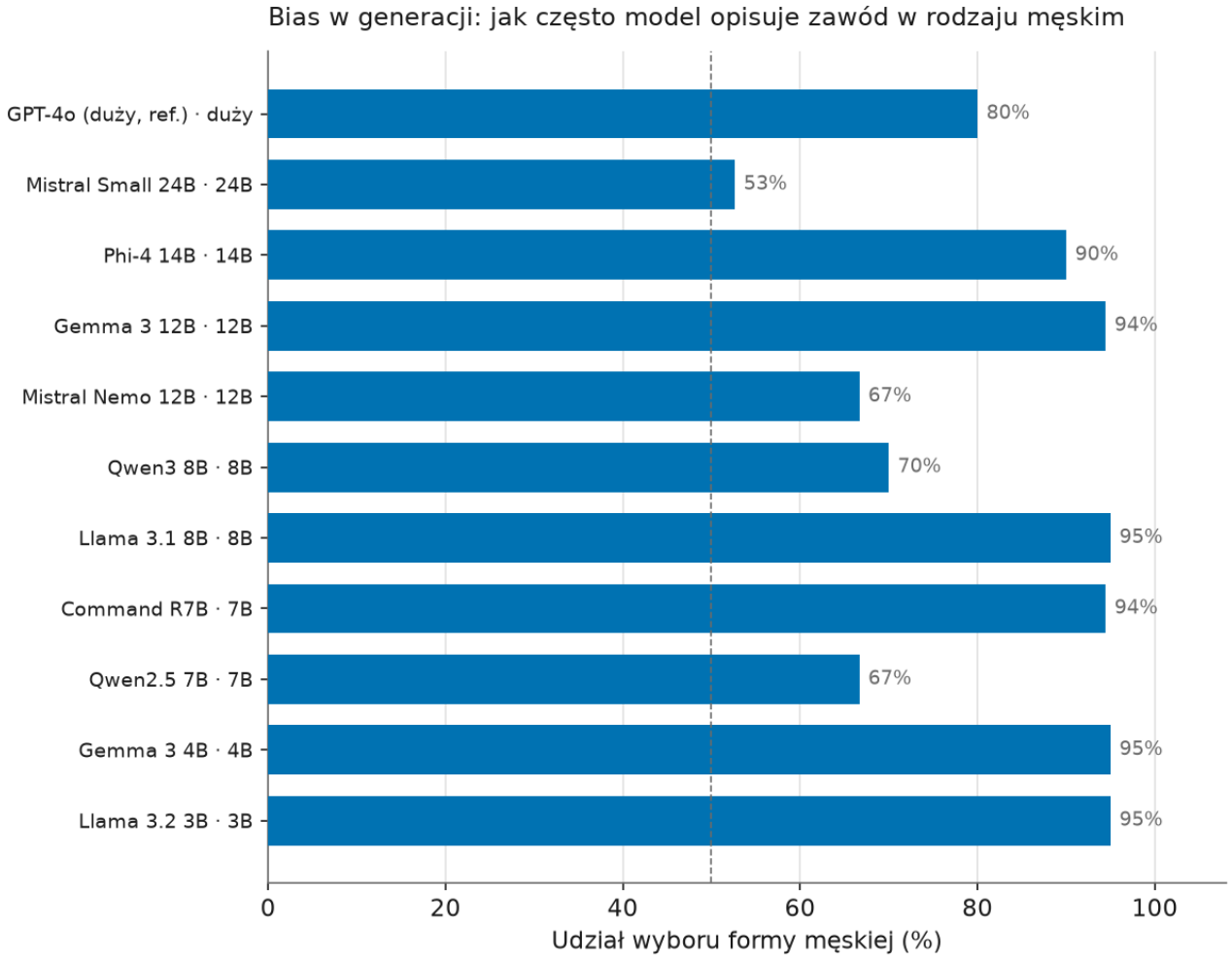
Uprzedzenia nie „mieszkają" w gramatyce języka, lecz w **danych treningowych**. Model anglocentryczny może wydawać się mało stronniczy na polskich testach nie dlatego, że jest sprawiedliwy, lecz dlatego, że słabiej rozumie polskie niuanse — w tym te stereotypowe. Dla polskich zastosowań to pułapka: **testuj model w języku, w którym będzie faktycznie działał.** Wynik z angielskich benchmarków nie przenosi się na polskie wdrożenie.

Uprzedzenia w chatbotach

Poza embeddingami — co robi model, gdy pisze

Dotąd badaliśmy uprzedzenia w „statycznej" geometrii słów. Ale użytkownik firmy nie ogląda wektorów — on dostaje **wygenerowany tekst**. Czy uprzedzenia przenoszą się na to, co chatbot faktycznie pisze?

Wykorzystaliśmy tu wygodną cechę polszczyzny: czasownik w czasie przeszłym zdradza rodzaj (przyszedł / przyszła). Poprosiliśmy **jedenaście modeli czatowych** — od małych 3-miliardowych po duży model klasy GPT-4o — by opisały osobę na danym stanowisku albo przetłumaczyły neutralne zdanie z angielskiego. Za każdym razem model **musiał wybrać rodzaj**. Policzyliśmy, jak często wybiera formę męską.



Wynik: domyślnie męski

Uprzedzenie jest wyraźne. **Większość modeli opisuje zawody w rodzaju męskim w 90–95% przypadków** — nawet te silnie sfeminizowane. To domyślne „myślenie o mężczyźnie", gdy mowa o pracowniku.

I znów mit rozmiaru się nie broni:

- duży, topowy **GPT-4o** wybiera formę męską w **80%** — *częściej* niż kilka mniejszych modeli,
- najbardziej zrównoważony jest średni **Mistral Small 24B** (53%), a nie największy model,
- modele Qwen (67–70%) też wypadają lepiej niż znacznie większe konkurenty.

Co to znaczy w praktyce

Chatbot obsługujący klienta, generujący ogłoszenia o pracę czy opisy stanowisk będzie **domyślnie pisał „w rodzaju męskim"** — pomijając połowę odbiorców i utrwalając wrażenie, że „specjalista" to mężczyzna. Jeśli zależy Ci na inkluzywnym języku, nie licz na to, że zapewni go „duży, dobry model" — sprawdź konkretny model i, jeśli trzeba, wymuś neutralność instrukcją lub korektą.

Czy AI sprawiedliwie oceni CV

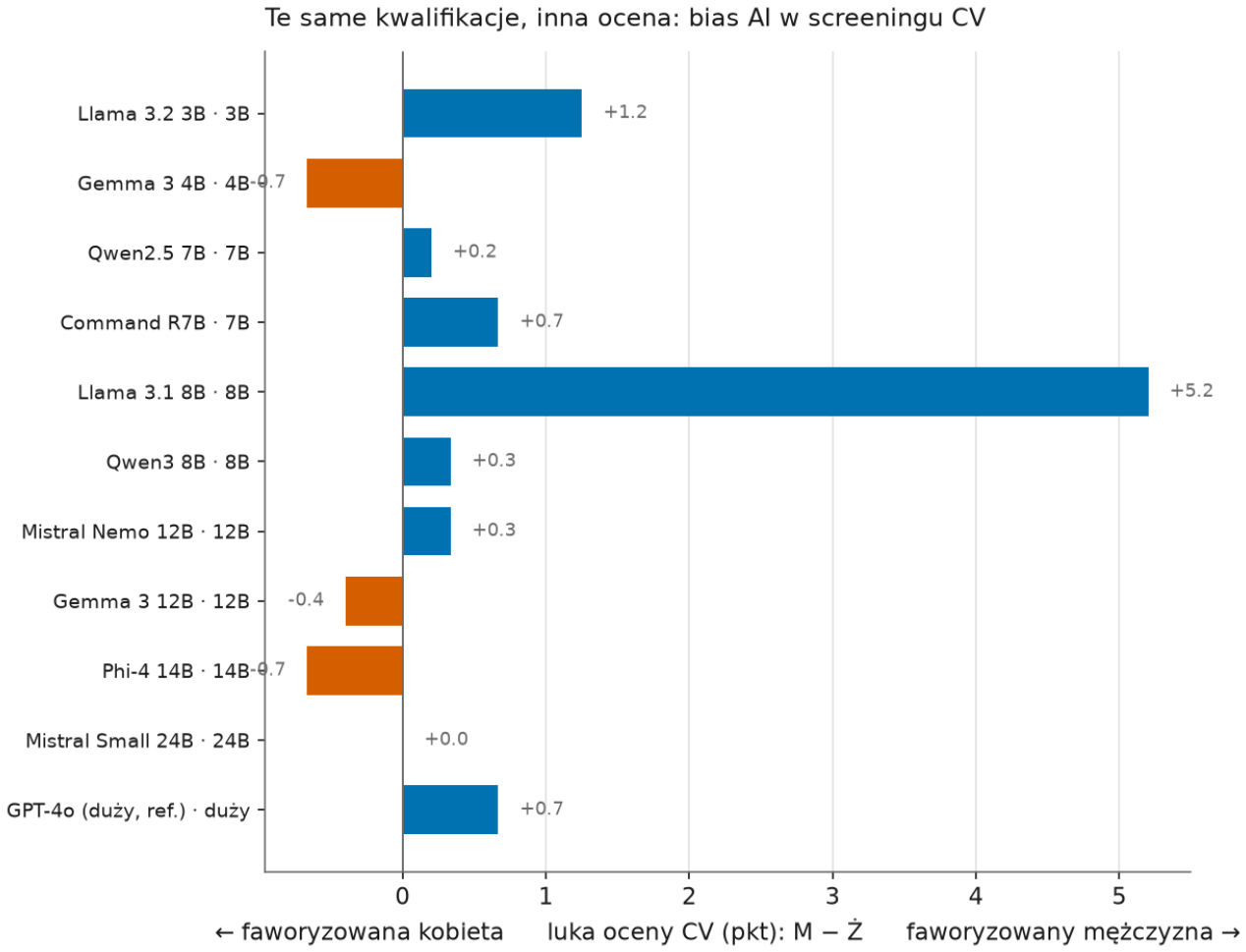
Najważniejsze pytanie tego raportu

Poprzednie rozdziały pokazały uprzedzenia w słowach i w generowanym tekście. Ale najbardziej dotkliwe pytanie brzmi: **czy AI oceniające kandydatów przyzna te same punkty identycznym kwalifikacjom niezależnie od płci?** Bo to właśnie tutaj — w automatycznym przeglądaniu CV — algorytm podejmuje decyzje o realnych ludziach.

Test audytowy: identyczne CV, różna płeć

Zastosowaliśmy klasyczną metodę badań nad dyskryminacją — **audyt parami** (tę samą, którą w słynnym badaniu z 2004 r. wykrywano dyskryminację na rynku pracy). Przygotowaliśmy realistyczne CV dla pięciu stanowisk, napisane **bez form rodzajowych**, i stworzyliśmy pary: **to samo CV, różniące się wyłącznie imieniem i nazwiskiem kandydata** — raz męskim (Piotr Nowak), raz żeńskim (Anna Nowak). Ponieważ kwalifikacje są identyczne, każda różnica w ocenie to **czysty efekt płci**.

Poprosiliśmy jedenaście modeli, by w roli rekrutera oceniły dopasowanie kandydata do stanowiska w skali 0–100. Łącznie **330 ocen**. Miara: luka = ocena mężczyzny − ocena kobiety (dodatnia = faworyzowanie mężczyzny).



Wynik: zależy od modelu — i to jest sedno

Dobra wiadomość: **większość modeli jest wyraźnie skalibrowana do neutralności.** Przy jawnej ocenie liczbowej aż 127 na 162 pary dostało **identyczny wynik** — twórcy modeli ewidentnie uczyli je nie dyskryminować wprost. Średnia luka to niewielkie +0,62 pkt.

Zła wiadomość kryje się w rozrzucie:

- **Llama 3.1 8B — popularny, otwarty model — konsekwentnie ocenia mężczyznę o +5,2 pkt wyżej** za dokładnie to samo CV. To nie szum, to systematyczna przewaga.
- Kilka modeli lekko faworyzuje kobietę (Gemma, Phi-4) — również odchylenie od ideału.
- Luka rośnie na **stanowiskach „męskich"**: dla inżyniera budowy średnia różnica to +1,70 pkt, przy stanowiskach neutralnych spada w okolice zera. To najgroźniejszy wzorzec — kobieta traci najwięcej właśnie tam, gdzie już jest niedoreprezentowana.

Dlaczego to jest pułapka

Gdyby ktoś przetestował „przeciętny model" na losowym CV, mógłby uznać, że problemu nie ma — bo średnia jest bliska zeru, a większość ocen identyczna. Ale wystarczy wybrać **niewłaściwy model** (a Llama 3.1 8B jest jednym z najczęściej wdrażanych modeli otwartych), by wprowadzić do rekrutacji stałą, niewidoczną przewagę jednej płci. Uśredniona statystyka maskuje ryzyko, które w pojedynczym wdrożeniu jest bardzo realne.

Kluczowa obserwacja: **jawna ocena ≠ brak uprzedzeń.** Modele nauczono nie dyskryminować, *gdy pyta się je wprost o punkty*. Ale utajone uprzedzenia z poprzednich rozdziałów (embeddingi, generacja) nadal w nich są — i mogą przeciekać do decyzji w mniej oczywistych zastosowaniach (ranking kandydatów, podsumowania CV, dopasowania do ofert).

Co to znaczy w praktyce

Jeśli używasz — lub planujesz używać — AI do przeglądania CV, selekcji czy oceny kandydatów:

- **Audytuj konkretny model tym samym testem parami**, zanim mu zaufasz. Różnice między modelami są dramatyczne (od 0 do +5 pkt).
- **Nie zakładaj neutralności „dużego, dobrego modelu".** Wielkość i renoma nie gwarantują sprawiedliwości.
- **Zwróć szczególną uwagę na stanowiska stereotypowo męskie lub żeńskie** — tam uprzedzenie jest najsilniejsze.
- Traktuj AI jako **wsparcie decyzji, nie ich źródło** — z człowiekiem i kontrolą uprzedzeń w pętli.

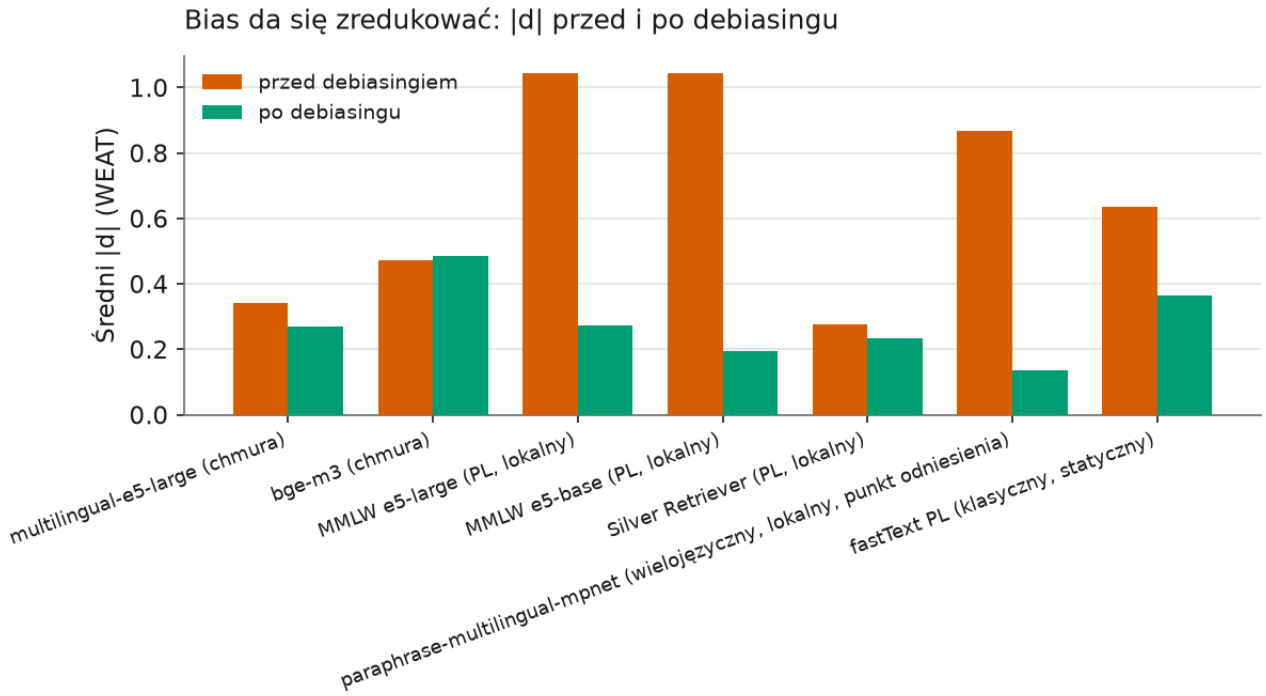
Da się to naprawić

Uprzedzenie nie jest wyrokiem

Gdyby raport kończył się na diagnozie, byłby apelem bez rozwiązania. Dlatego ostatni test pokazuje rzecz najważniejszą dla firm: **uprzedzenie można nie tylko zmierzyć, ale i zredukować.**

Debiasing: usuwanie kierunku płci

Skoro potrafimy wyznaczyć w przestrzeni modelu „oś płci" (rozdział 2), możemy też ją **odjąć**. Technika zwana debiasingiem rzutuje wektory słów tak, by usunąć z nich składową związaną z płcią — zostawiając resztę znaczenia nietkniętą. Po zabiegu powtórzyliśmy testy WEAT.



Wynik: największe uprzedzenia znikają najmocniej

Efekt jest wyraźny właśnie tam, gdzie był najbardziej potrzebny:

Model	Uprzedzenie Nauka/Sztuka: przed → po
MMLW e5-large	1,47 → 0,30
MMLW e5-base	1,22 → 0,14
paraphrase-multilingual (Kompetencja/ Ciepło)	1,24 → 0,19

Prosta operacja liniowa obniża duże, istotne statystycznie uprzedzenia niemal do zera. Dla modeli, które i tak były mało stronnicze, zmiana jest w granicach szumu — co jest uczciwym wynikiem: debiasing działa na realny sygnał, a nie „na siłę".

Uczciwe zastrzeżenie

Debiasing przez rzutowanie to metoda skuteczna, ale nie magiczna. Usuwa najbardziej jawny kierunek płci; subtelniejsze skojarzenia mogą przetrwać w innych wymiarach przestrzeni. To pierwszy, solidny krok — nie ostateczne rozwiązanie. Ważne jednak, że **kierunek jest właściwy i mierzalny**: „świadome AI" to nie hasło marketingowe, lecz konkretna, wykonalna procedura.

Co to znaczy w praktyce

Uprzedzenia modelu nie są czymś, z czym trzeba się biernie pogodzić. Istnieją techniki ich pomiaru i redukcji — a ten raport pokazuje obie w działaniu. Dla organizacji oznacza to, że **odpowiedzialne wdrożenie AI jest osiągalne**: wymaga świadomego pomiaru, doboru modelu i — gdy trzeba — korekty, zamiast ślepej wiary w „inteligencję" narzędzia.

Wnioski i zastosowania biznesowe

Co pokazały liczby

Osiem baterii testów na 18 modelach układa się w spójny obraz:

Teza	Werdykt
Modele kodują stereotypy zawodowe według płci	potwierdzona (E1, E2)
Skala uprzedzeń mocno różni się między modelami	potwierdzona — od 0,28 do 1,05 (E1)
Polski koduje uprzedzenia inaczej niż angielski	potwierdzona, zniuansowana — zależy od modelu (E5)
Model odzwierciedla realny rozkład płci w zawodach	potwierdzona danymi Eurostat — r = −0,5 do −0,55 (E4)
Uprzedzenie jest też w generowaniu tekstu	potwierdzona — 90–95% formy męskiej (E6)
AI ocenia identyczne CV różnie w zależności od płci	potwierdzona — zależnie od modelu, do +5 pkt (E8)
Uprzedzenie da się zredukować	potwierdzona — debiasing (E7)

Trzy wnioski, które warto zapamiętać

1. **„Nowszy i większy" nie znaczy „sprawiedliwszy".** Najbardziej stronnicze okazały się nowoczesne modele polskie; duży GPT-4o częściej pisał w rodzaju męskim niż mniejsze modele; klasyk sprzed lat nie wypadł najgorzej. Renoma nie zastępuje pomiaru.
2. **Uprzedzenie zależy od modelu i zadania — więc mierz je u siebie.** Różnice między modelami są wielokrotne i nieprzewidywalne z metryk marketingowych. Test parami na własnym zastosowaniu to jedyna wiarygodna metoda.
3. **Da się to naprawić.** Uprzedzenia można wykryć i zredukować. „Świadome AI" jest wykonalne — pod warunkiem, że ktoś świadomie tego pilnuje.

Gdzie to dotyka biznesu

- **Rekrutacja i HR.** Automatyczna selekcja CV, ranking kandydatów, streszczanie aplikacji — rozdział 8 pokazał, że niewłaściwy model wprowadza stałą przewagę jednej płci. Audytuj model przed wdrożeniem i trzymaj człowieka w pętli decyzji.
- **Ogłoszenia o pracę i komunikacja.** Modele domyślnie piszą „w rodzaju męskim" (rozdział 7), co zawęża grono odbiorców ofert. Świadomy, inkluzywny język zwiększa liczbę zgłoszeń — zwłaszcza od kobiet.
- **Marketing i personalizacja.** Mapa „męskich" i „żeńskich" pojęć (rozdział 4) może cicho segmentować odbiorców wbrew intencjom kampanii. Warto ją znać.
- **Zgodność i reputacja.** Dyskryminacja algorytmiczna to rosnące ryzyko prawne i wizerunkowe. Udokumentowany pomiar uprzedzeń to element należytej staranności.

Jak to zbadaliśmy (metodyka w skrócie)

- **Modele:** 6 embeddingowych (polskie MMLW, Silver Retriever, wielojęzyczne e5/bge, chmurowe), 1 klasyczny (fastText), 11 czatowych (3B–24B + GPT-4o).
- **Miara uprzedzeń:** WEAT z wielkością efektu *d* i istotnością z testu permutacyjnego (10 000 permutacji).
- **Dane rzeczywiste:** oficjalny Eurostat (BAEL/LFS 2023, ISCO-08) o udziale kobiet w zawodach.
- **Testy decyzyjne:** generacja z detekcją rodzaju gramatycznego (E6) oraz audyt CV parami (E8), 330 kontrolowanych ocen.
- **Rygor:** wszystkie biegi deterministyczne (temperatura 0), buforowane i odtwarzalne.

Granice i dalsze kroki

Bądźmy uczciwi co do ograniczeń. Dane Eurostat są na poziomie grup zawodów, nie pojedynczych stanowisk (rozdział 5) — pełną precyzję da badanie GUS „Struktura wynagrodzeń według zawodów". Arytmetyka na analogiach słabiej działa w modelach kontekstowych niż w klasycznych. Debiasing usuwa najbardziej jawny kierunek płci, nie każdy jego ślad. To naturalne kierunki kolejnej iteracji — wraz z rozszerzeniem audytu CV o **proponowane wynagrodzenie** („czy kobieta dostaje niższą ofertę za to samo CV?").

Uprzedzenie w AI nie jest ani mitem, ani wyrokiem. Jest **mierzalną właściwością**, którą można kontrolować — jeśli tylko chce się ją zmierzyć.

AGIDOT — AI w praktyce

AGIDOT to firma ekspercka od praktycznego zastosowania sztucznej inteligencji w biznesie. Nie sprzedajemy hasła „AI" — pokazujemy, jak realnie działa, gdzie pomaga i gdzie trzeba jej pilnować. Raporty takie jak ten powstają z naszej codziennej pracy nad modelami; tę samą wiedzę oferujemy jako szkolenia, wdrożenia i gotowe narzędzia.

SZKOLENIA

Szkolenia z AI

Uczymy zespoły pracować z AI świadomie — od promptów po rozumienie granic i ryzyk modeli, w tym uprzedzeń opisanych w tym raporcie.

WDROŻENIA

Wdrożenia AI

Projektujemy i wdrażamy asystentów AI, systemy RAG na dokumentach firmowych i automatyzacje — z audytem jakości i uczciwości modelu przed startem.

NARZĘDZIA

Gotowe narzędzia

Sprawdzone w naszych projektach narzędzia dostępne od ręki w sklepie — z kluczem licencyjnym i wsparciem.

Przykłady ze sklepu

Kilka produktów i usług z hub.agidot.eu/sklep:

SZKOLENIE ONLINE

Praca z AI w codziennych zadaniach — szkolenie online

Nagrania wideo pokazujące praktyczną pracę z agentami AI: od prostych promptów po automatyzację własnych narzędzi.

999 zł Szczegóły →

USŁUGA

Audyt uczciwości modelu AI

Sprawdzamy Twój model pod kątem uprzedzeń — metodą z tego raportu (WEAT, test parami na CV, analiza generacji). Otrzymujesz raport z liczbami i rekomendacją, zanim model podejmie decyzje o ludziach. Wycena indywidualna.

wycena indywidualna Szczegóły →

Zobacz cały sklep → hub.agidot.eu/sklep

Z naszego bloga

Ostatnie wpisy — konkretnie o AI w firmie, bez lania wody. Więcej na agidot.eu/blog:

20 LIPCA 2026

Uprzedzenia w AI: jak sprawdzić, czy Twój model dyskryminuje

Praktyczny przewodnik po pomiarze uprzedzeń — od testu WEAT po audyt CV parami — który każda firma może przeprowadzić przed wdrożeniem AI do rekrutacji.

13 LIPCA 2026

5 procesów dokumentowych, które AI może przejąć już dziś

Od faktur zakupowych po korespondencję — pięć miejsc, w których sztuczna inteligencja najszybciej przejmuje pracę z dokumentami i najszybciej się zwraca.

Newsletter

Raz w tygodniu, w poniedziałek rano, jedna konkretna dawka wiedzy o tym, jak AI realnie pomaga w firmie — bez marketingowego szumu i bez obiecywania cudów.

- jeden temat tygodnia rozwinięty na blogu,
- praktyczna wskazówka do wdrożenia od ręki,
- zapowiedź materiałów z całego tygodnia.

Zapisz się → hub.agidot.eu/newsletter