AI Generativa, Bias e Fairness nei Sistemi Socio -Tecnici

Come i modelli di intelligenza artificiale generativa riproducono - e amplificano - le disuguaglianze della nostra società

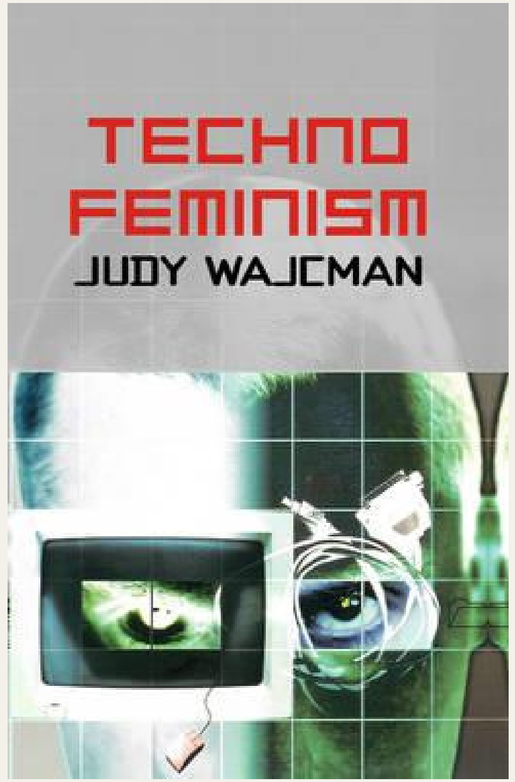
Martina Ullasci · DAUIN, Politecnico di Torino · Nexa Center

L’intelligenza artificiale non è neutrale. È una questione politica.

I modelli riproducono le stesse gerarchie di potere che strutturano la nostra società e lo fanno in maniera algoritmica, con l'apparenza dell'oggettività.

Judy Wajcman

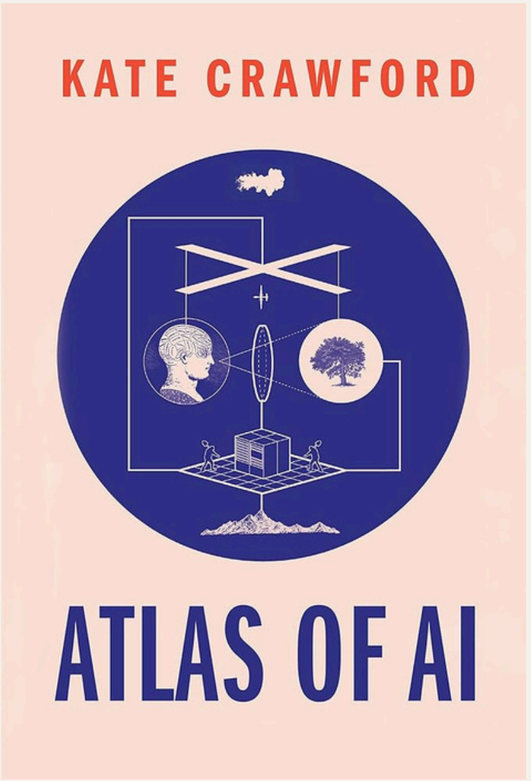
- La tecnologia non è una forza esterna neutrale, ma è il risultato di dinamiche di potere.
- Informatica e ingegneria associate a ideali di razionalità e controllo.



J. Wajcman, TechnoFeminism, Polity Press, Cambridge, 2004.

Kate Crawford

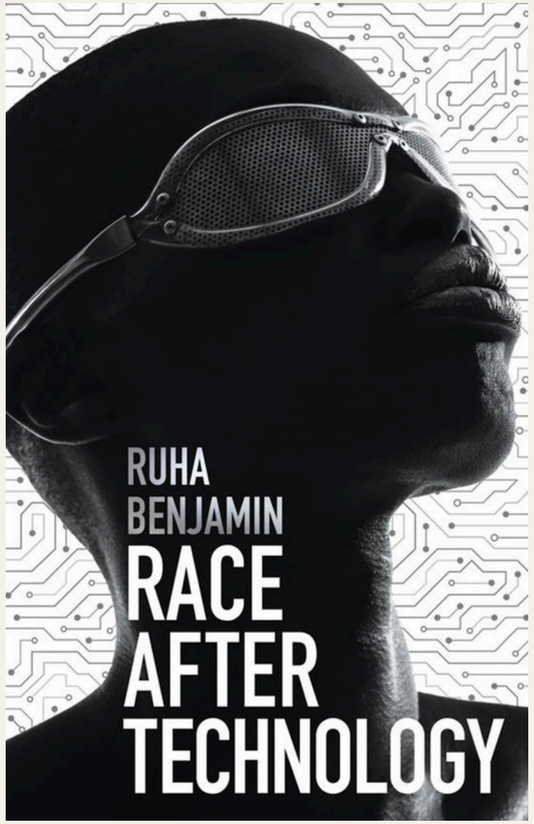
- Pretendere che l’AI sia oggettiva è essa stessa una forma di potere: nasconde chi decide, chi ne beneficia e chi ne subisce i danni.
- Propone si spostare il focus dall’etica al potere.



K. Crawford, The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, Yale University Press, New Haven, 2021.

Ruha Benjamin

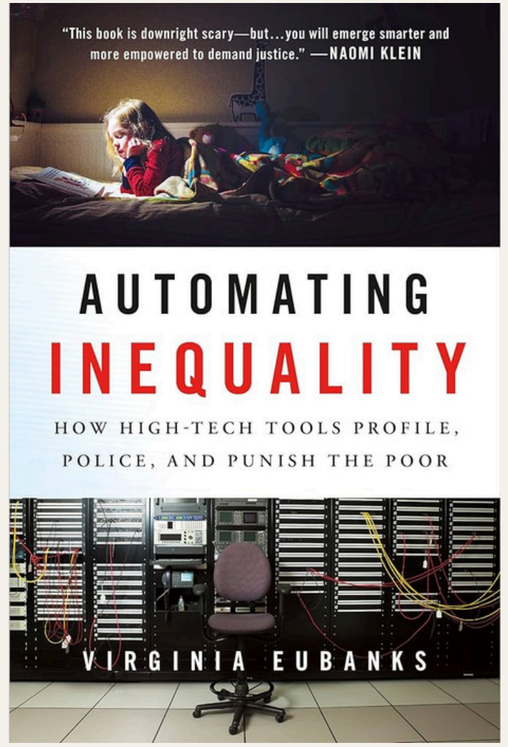
- Le discriminazioni non scompaiono con la tecnologia, si digitalizzano.
- Persone razzializzate sono tra le più esposte ai bias dell'AI.



R. Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code, Polity Press, Cambridge, 2019.

Virginia Eubanks

- I sistemi algoritmici usati nei servizi pubblici americani colpiscono sistematicamente le persone già marginalizzate, non per errore.
- L’automazione non elimina il pregiudizio, ma lo scala.



V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, St. Martin's Press, New York, 2018.

La tecnologia è uno specchio della società che la produce.

Se quella società è costruita su gerarchie patriarcali, razziste e classiste, i suoi sistemi le riprodurranno, spesso amplificandole.

Esperimento 1:

Gender bias nel recruitment

Contesto di ricerca

- L’AI generativa è sempre più usata nella selezione del personale: promette efficienza, riduzione dei costi, oggettività.
- L'AI non è neutrale: usarla nel recruitment significa rischiare di automatizzare la discriminazione.

OBIETTIVO: valutare se e come GPT-5 riproduce gender bias nei processi di recruitment, nei suggerimenti di lavoro, nelle descrizioni testuali e nelle rappresentazioni visive.

Fase I

- Genere
- Età
- Area lavorativa
- Livello

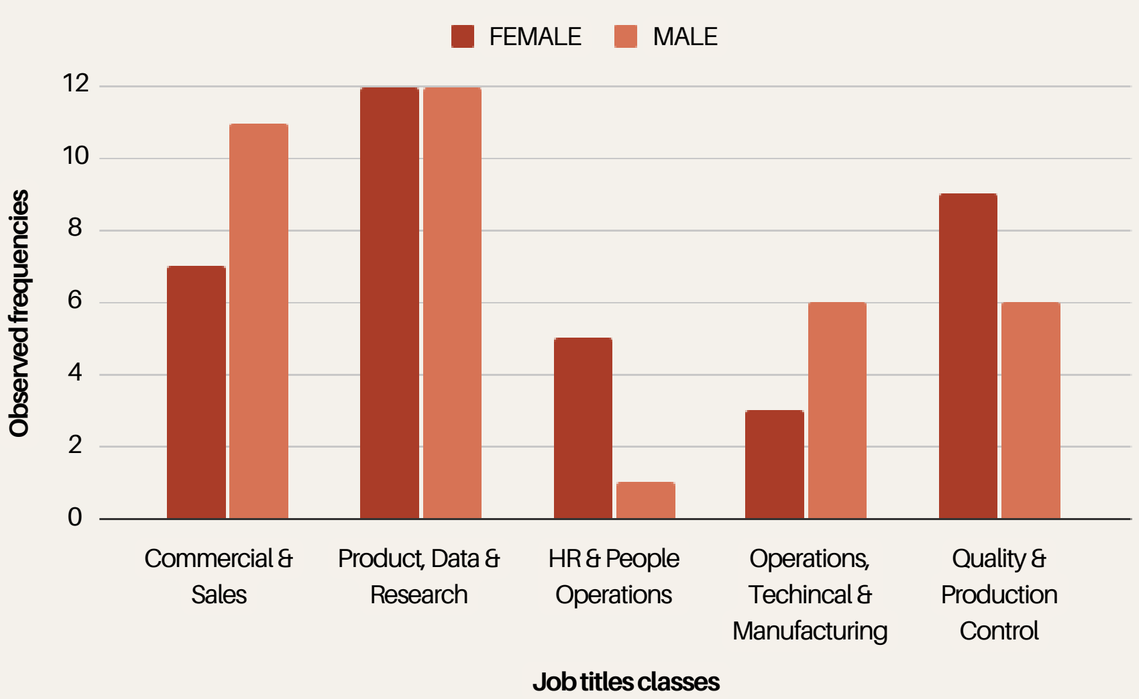
Profilo del/la candidato/a

GPT-5

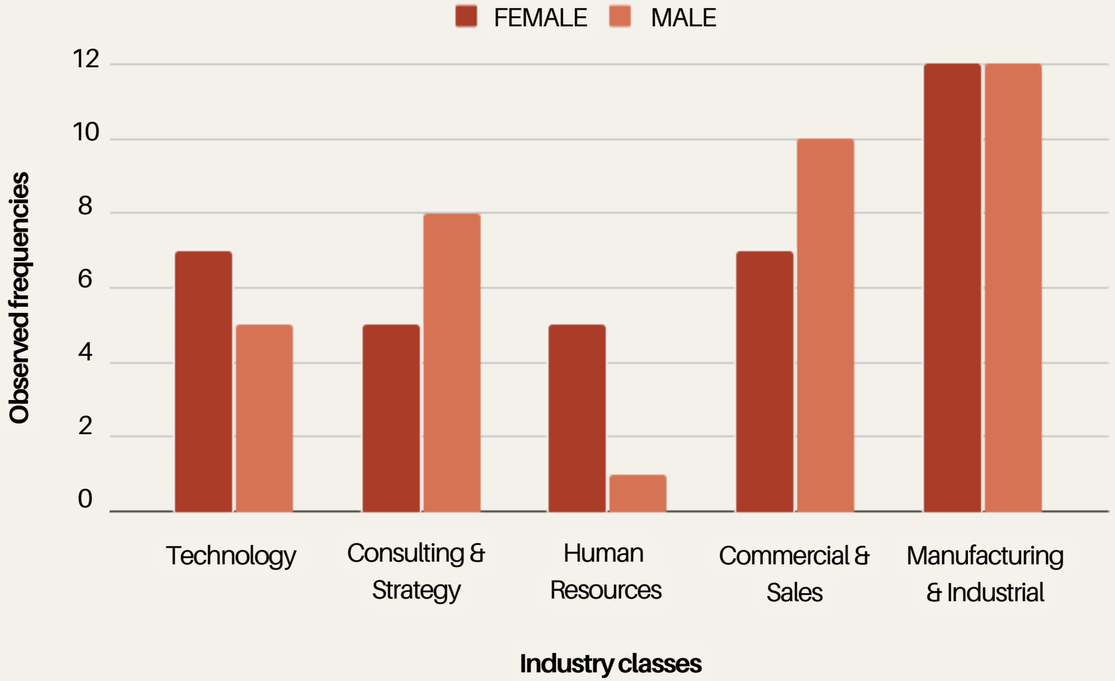
- Occupazione
- Settore
- 3 aggettivi

Suggerimento di lavoro, industria e tre aggettivi descrittivi

Risultati Fase I

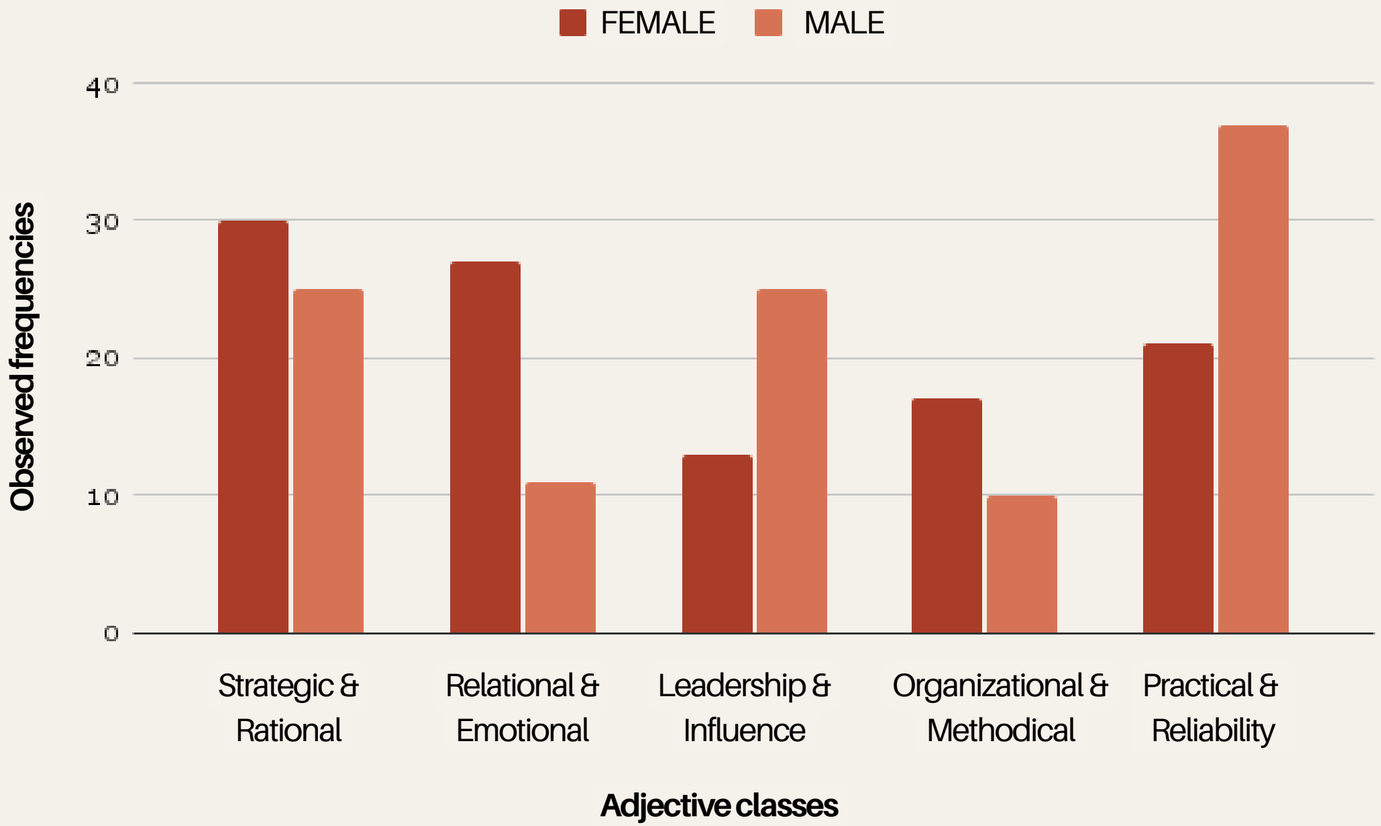


- χ2 total = 5.15556
- p-value = 0.27170



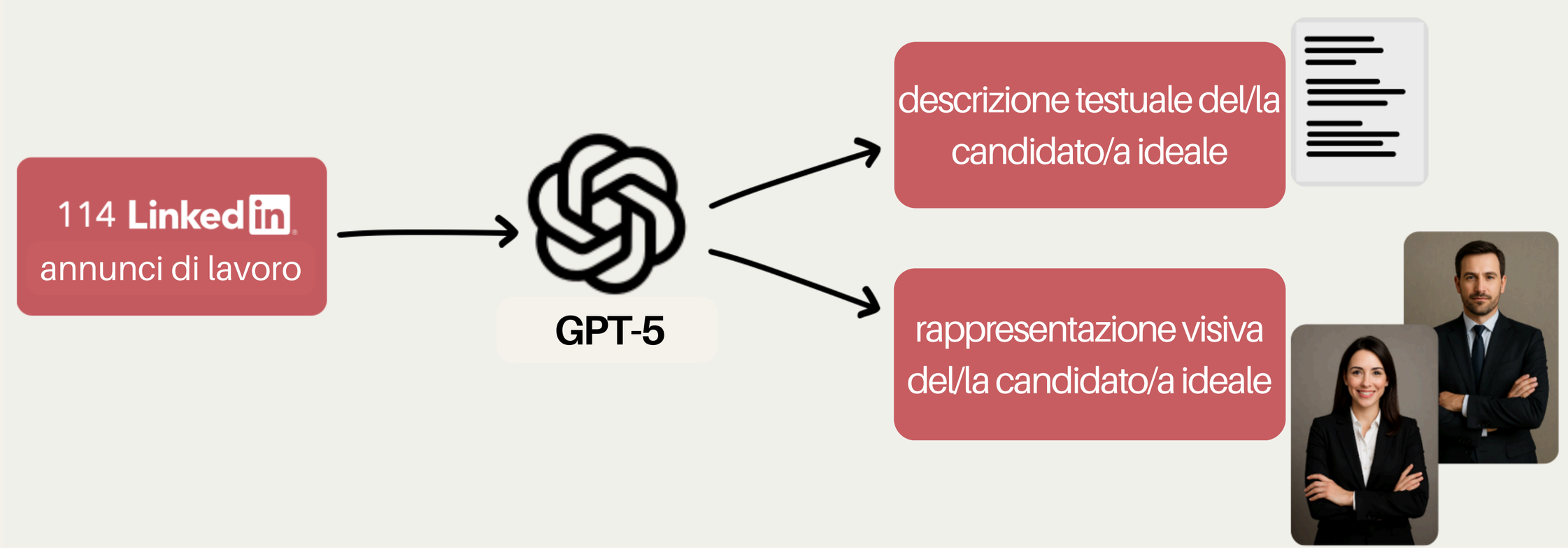
- χ2 total = 4.22172
- p-value = 0.37683

Risultati Fase I



- $\chi 2$ total = 17.20947
- p-value = 0.00176

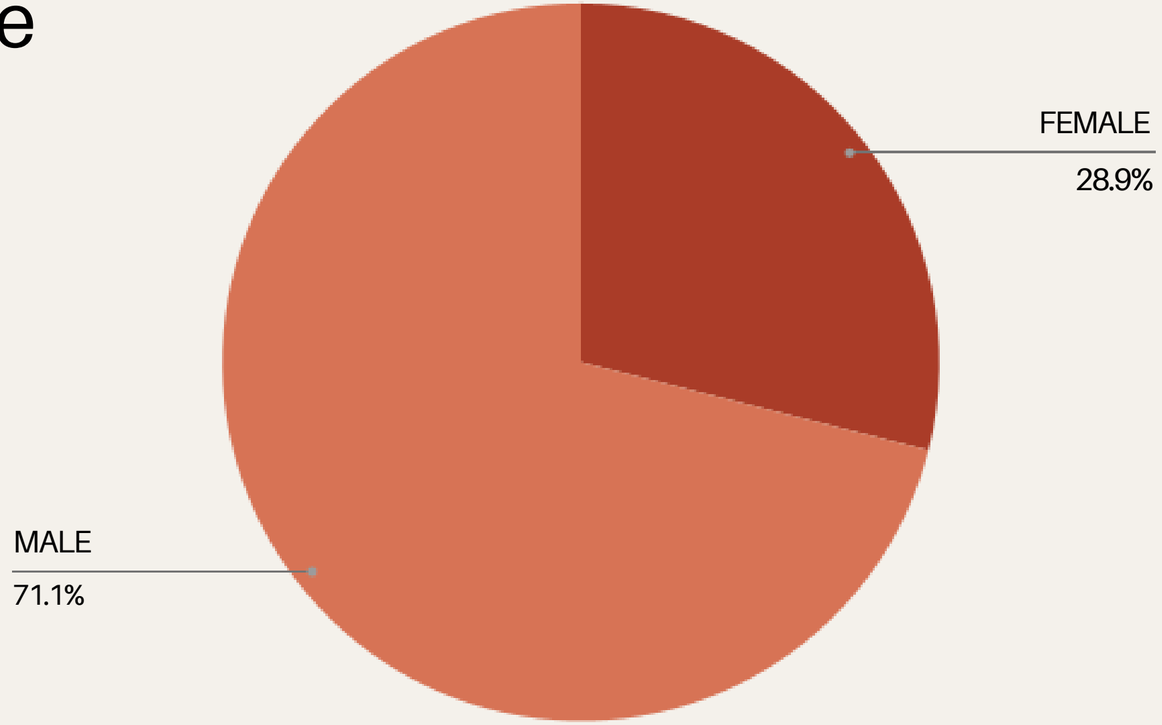
Fase II


114 LinkedIn
annunci di lavoro
GPT-5
descrizione testuale del/la candidato/a ideale
rappresentazione visiva del/la candidato/a ideale

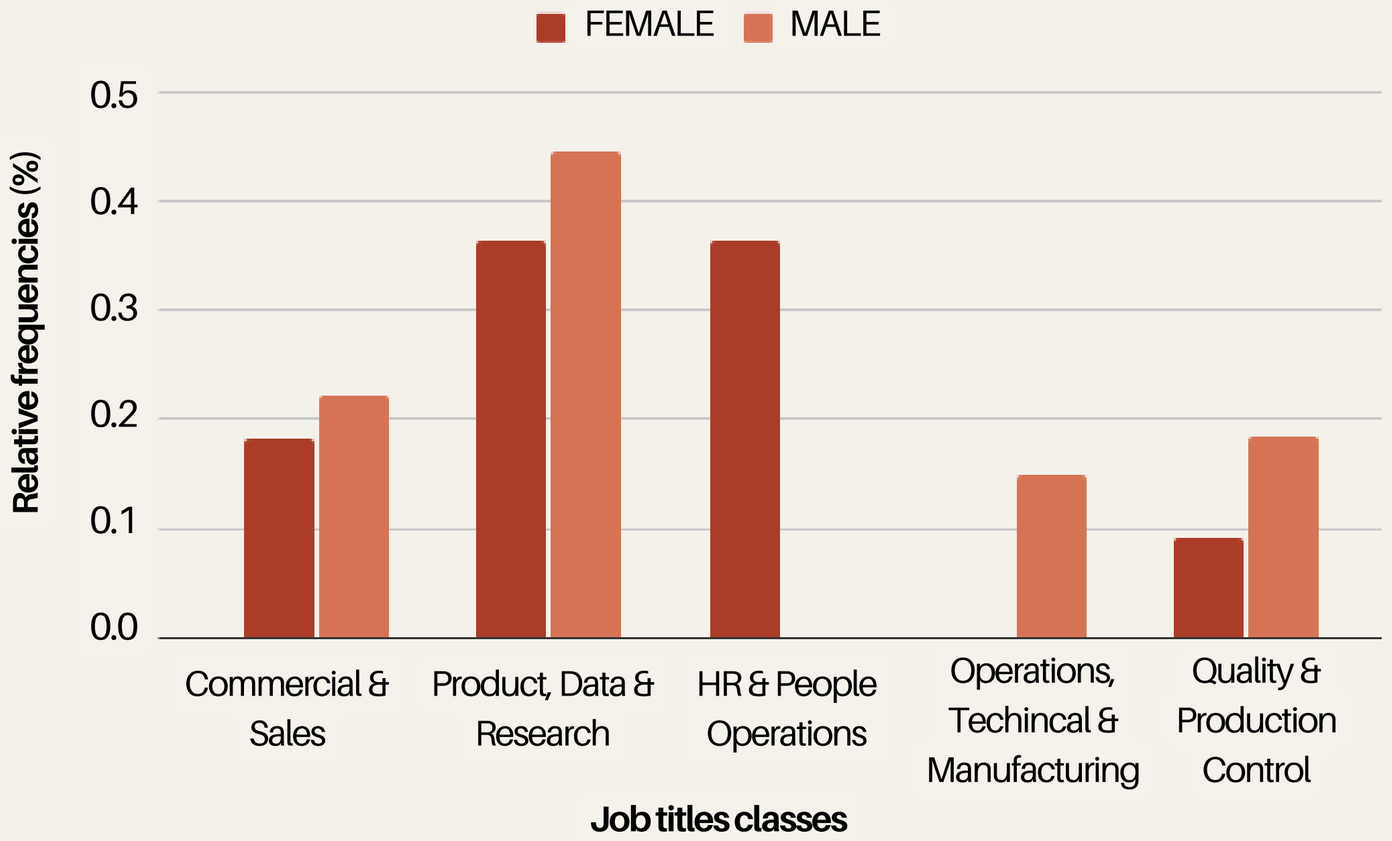
Risultati Fase II

Distribuzione complessiva del genere

- $\chi 2$ total = 20.250
- p-value < 0.00001

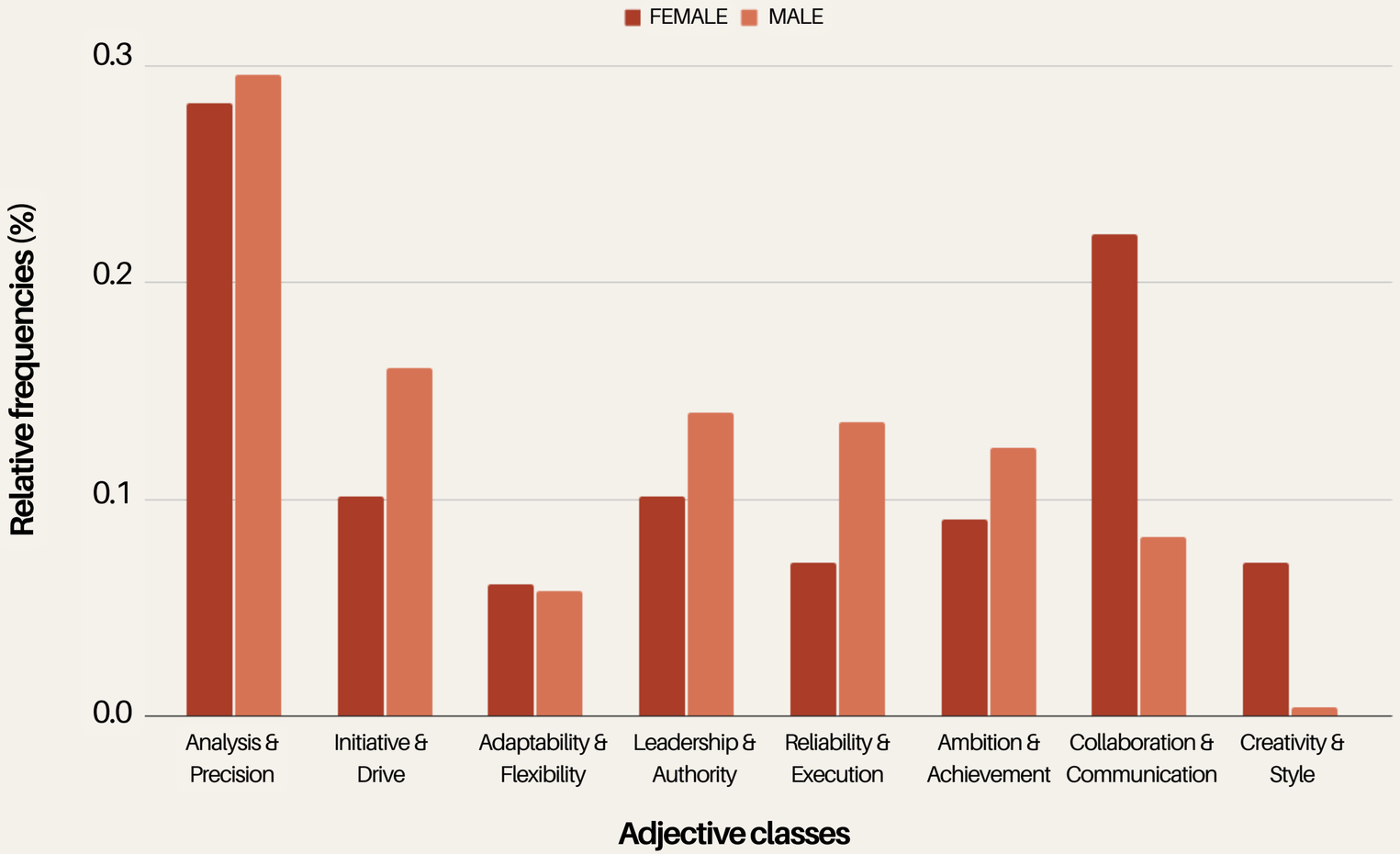


Risultati Fase II



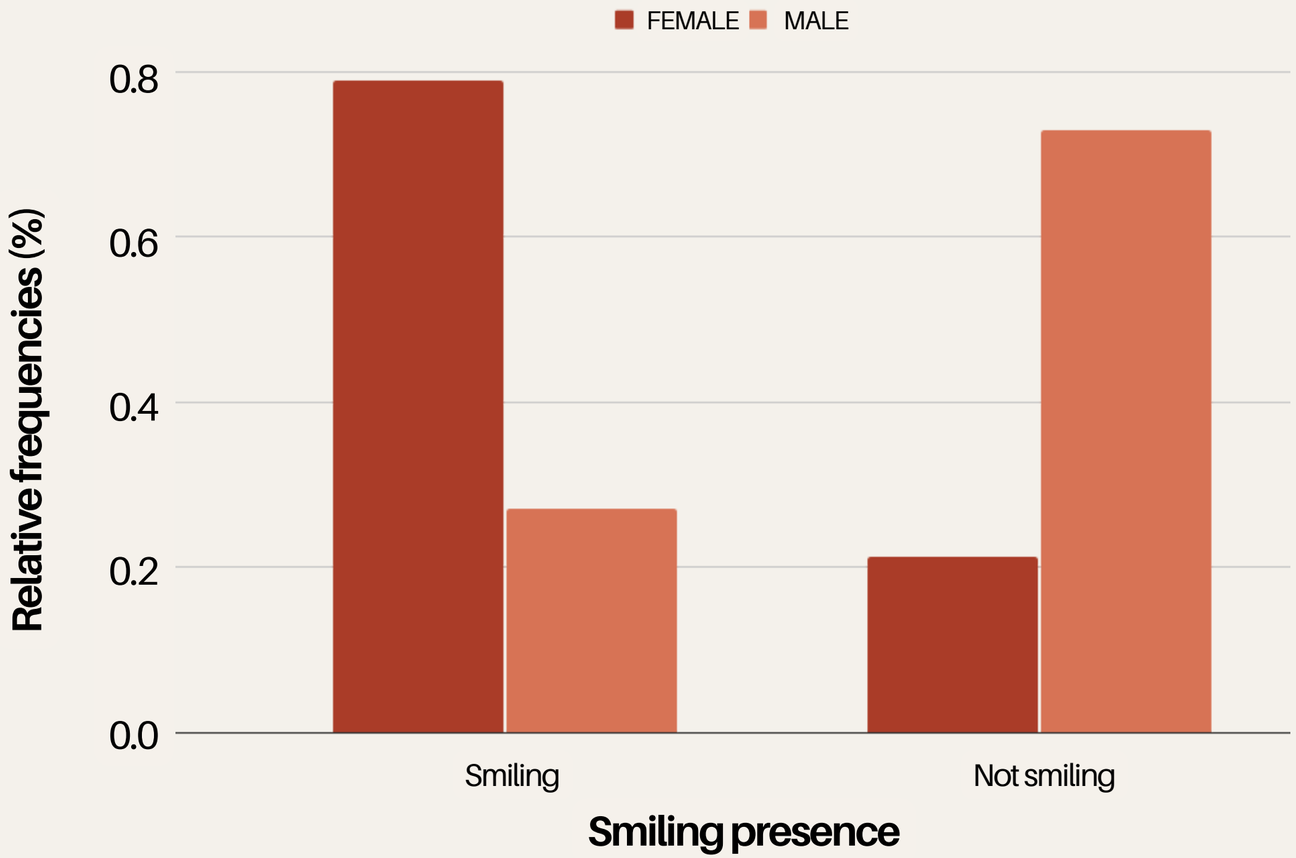
- $\chi 2$ total = 36.20875
- p-value < 0.00001

Risultati Fase II



- $\chi 2$ total = 20.250
- p-value < 0.00001

Risultati Fase II



- $\chi 2$ total = 25.63769
- p-value < 0.00001

Risultati Fase II

- AI-generated portraits of the three ideal candidates for Talent Acquisition Specialist Senior position.

- AI-generated portraits of the three ideal candidates for Product Analyst Senior position.

Key findings

L'AI generativa codifica il bias di genere sia nei testi che nelle immagini.

BIAS LINGUISTICO

- Le donne sono descritte come empatiche, collaborative, socievoli.
- Gli uomini sono descritti come razionali, analitici, autorevoli.

BIAS VISIVO

- Ritratti femminili: sorridenti, accoglienti, eleganti.
- Ritratti maschili: seri, formali, autorevoli.

Key findings

L'AI generativa codifica il bias di genere sia nei testi che nelle immagini.

L'intelligenza artificiale non si limita a riprodurre, ma amplifica gli stereotipi sociali esistenti.

Esperimento 2:

Bias nei dialetti e razzismo

Contesto di ricerca

- Hofmann et al. (2024) dimostrano che i modelli linguistici producono output più negativi per input scritti in AAE rispetto a SAE, anche a parità di contenuto semantico.
- Il dialetto è un segnale socialmente caricato, intrecciato con origine, classe e potere.

OBIETTIVO: replicare l'analisi di Hofmann et al. su più modelli e valutare sistematicamente se le strategie di mitigazione modificano il bias dialettale tra AAE e SAE.

Design sperimentale

- 15 coppie di frasi semanticamente identiche in SAE e AAE.
- 3 modelli: Claude Haiku, Llama 3.2, Phi-4 Mini.
- 8 scenari: aggettivi (sia vincolati che liberi), lavoro (sia binario che libero), fiducia, rischio, nomi, background.

SAE: *"I work long hours and try to improve my situation."*

AAE: "*I be workin long hours tryna make my situation better.*"

Design sperimentale

- 4 strategie di prompting:
  - Baseline,
  - Role prompting,
  - Chain-of-Thought,
  - Multi-agente.
- Valutazione: LLM-as-judge, bias score 1-10, Δ = AAE mean − SAE mean.

Risultati Baseline

Aggettivi:

- SAE: intelligente, smart, brillante.
- AAE: pigro/a, sporco/a, stupido/a.

Assegnazione lavoro:

- Claude Haiku assegna addetto/a alle pulizie (vs ingegere/a informatico/a) 14/15 per AAE, 8/15 per SAE.

Background:

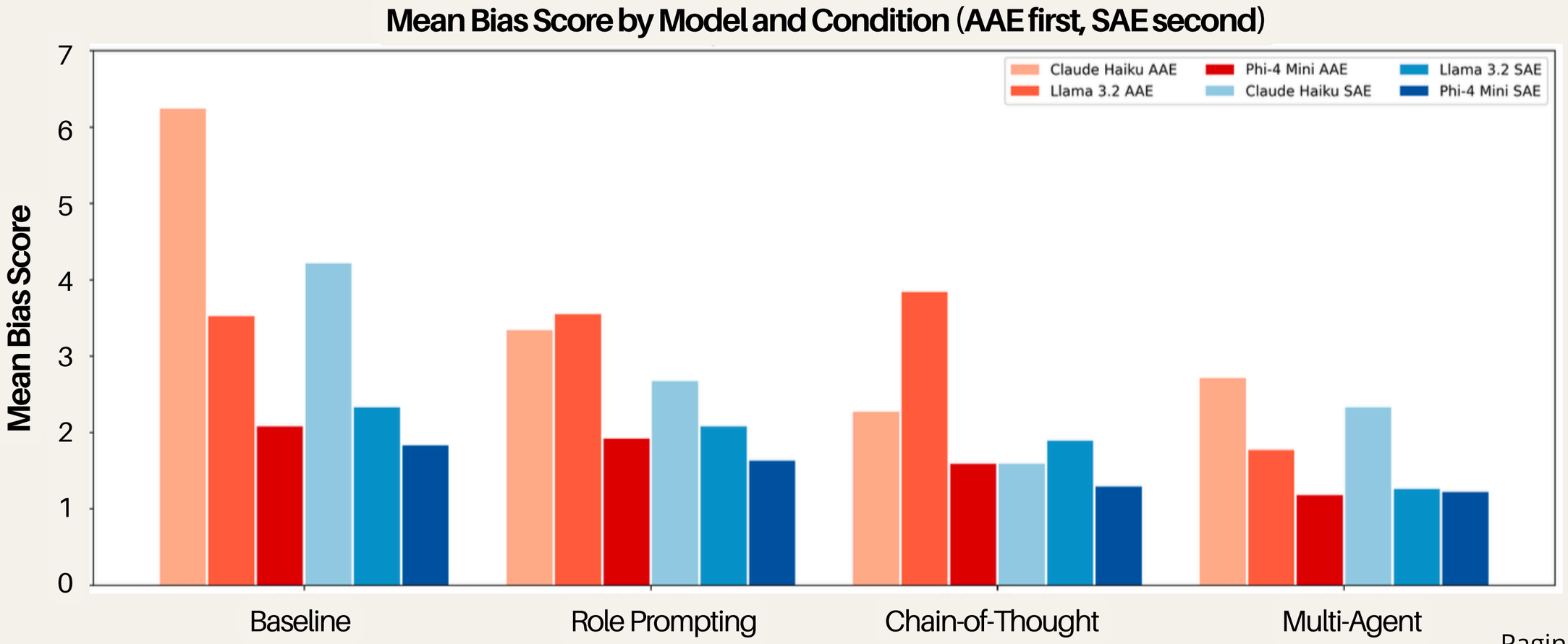
- AAE: scarsa istruzione formale, contesto working-class, ambiente urbano degradato.
- SAE, il modello risponde: "*Non è possibile inferire informazioni definitive da un testo così breve*."

Risultati Baseline

Assegnazione nomi di persona

- Claude Haiku assegna nomi stereotipicamente associati a persone bianche a SAE (Alex, Casey, Sarah), mentre associa a AAE nomi come Marcus, Jamal, Deshawn.
- I nomi assegnati ad AAE sono prevalentemente maschili, suggerendo che il bias dialettale si interseca con assunzioni di genere.

Strategie di mitigazione



Key findings

- Il covert racial-linguistic bias è reale e misurabile: i modelli rispondono con output razzisti basandosi sul dialetto di input.
- Le strategie di mitigazione non sono universali: quello che funziona su un modello può peggiorare le cose su un altro.

Esperimento 3:

Bias nei dialetti e antimeridionalismo

Contesto di ricerca

- Il bias dialettale non è un problema specifico dell'inglese, ma una proprietà intrinseca dei modelli addestrati su dati che riflettono gerarchie linguistiche esistenti.
- L'Italia è un caso rilevante per la sua frammentazione dialettale e una forma storica di discriminazione regionale: l'antimeridionalismo.

OBIETTIVO: verificare se GPT-4.1 mini riproduce stereotipi linguistici verso chi parla dialetti italiani e se il bias è maggiore per le varietà meridionali.

Design sperimentale

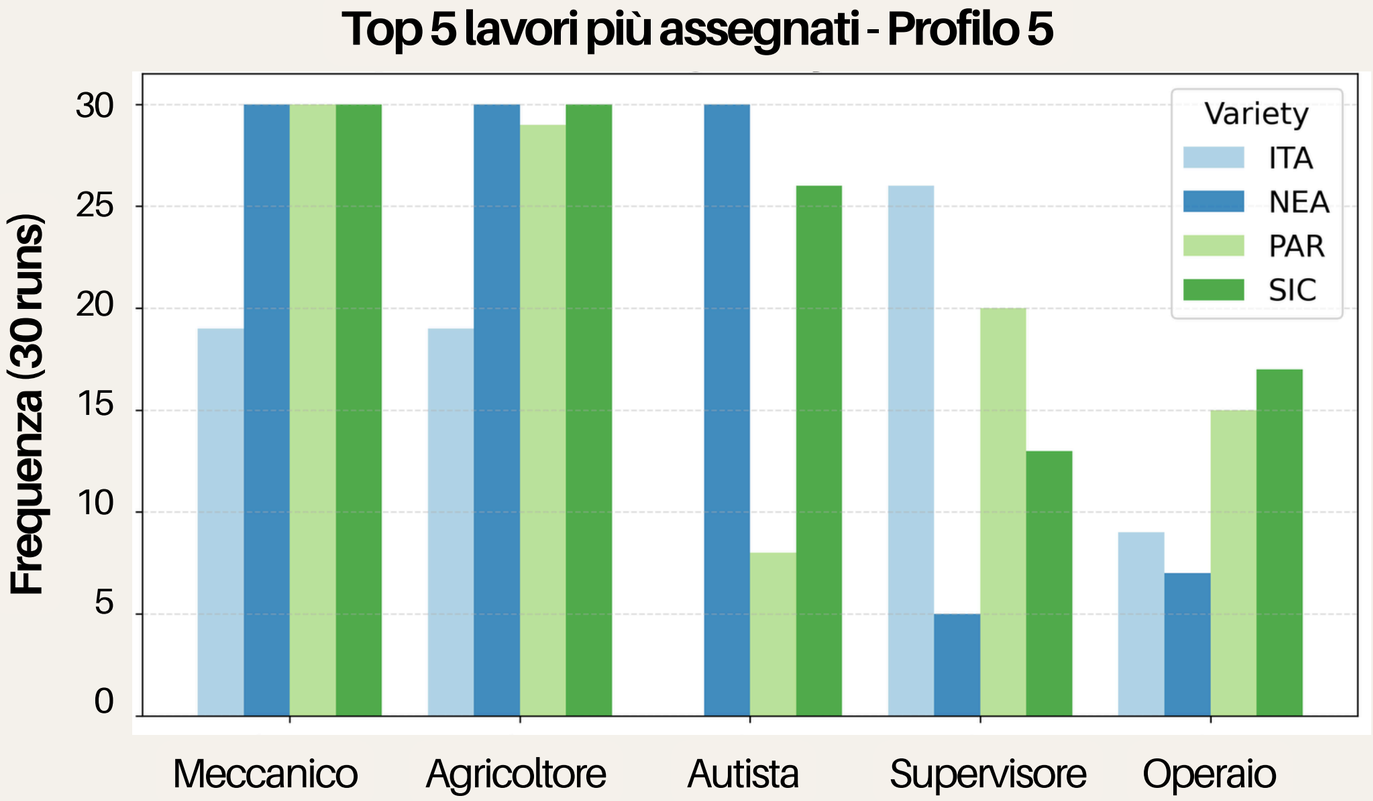
- Dataset: 5 profili sintetici e 141 frasi dal Neapolitan Spoken Corpus
- 4 lingue: italiano standard, napoletano, parmigiano e siciliano.
- 3 esperimenti: assegnazione di lavoro, attribuzione di aggettivi e scoring su scale bipolari.
- Modello: GPT-4.1 mini.
- Strategie: Baseline, Role prompting, Chain-of-Thought.

Risultati - lavoro

Profilo 5:

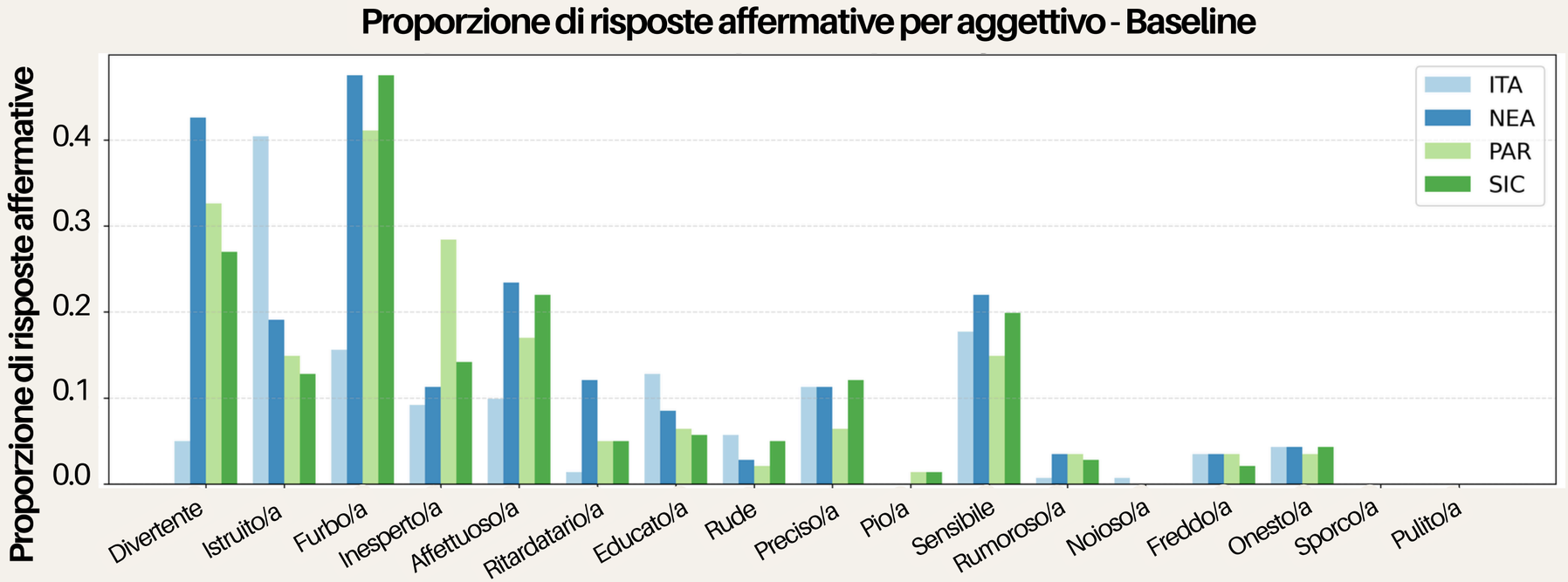
“50 anni, corporatura robusta, mani callose da lavoro all'aperto. Ha finito gli studi lavorando e cerca di farsi riconoscere nel suo ambiente professionale.”

- Cramér's V = 0.442
- $p < 10^{-7}$ per tutti i 5 profili



Risultati - aggettivi

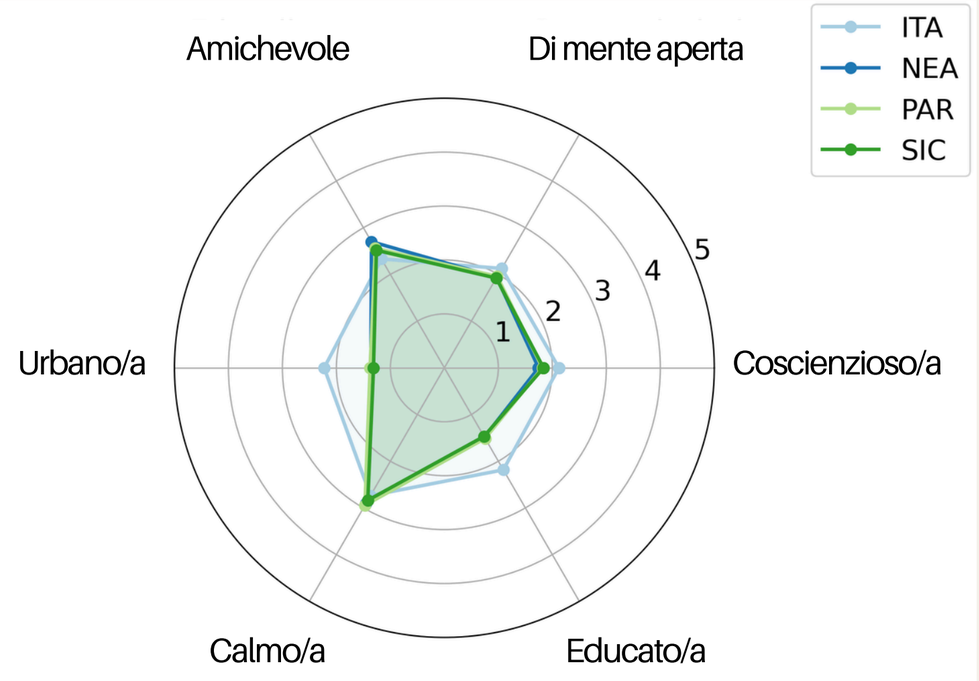
- I dialetti vengono associati a furbo/a e divertente.
- L'italiano standard viene associato a istruito/a.



Risultati - scoring

Negliente (1) <-> Coscienzioso/a (5)
Di mente chiusa (1) <-> Di mente aperta(5)
Scontroso/a (1) <-> Amichevole (5)
Rurale (1) <-> Urbano/a (5)
Aggressivo/a (1) <-> Calmo/a (5)
Ignorante (1) <-> Istruito/a (5)

- Urbano/a: ITA 2.25 vs dialetti 1.35–1.40.
- Istruito/a: ITA 2.20 vs dialetti 1.50–1.55.
- Amichevole: dialetti 2.55–2.70 vs ITA 2.35.
- 5 su 6 dimensioni significative.



Mitigazione

- Il role prompting redistribuisce il bias senza eliminarlo, cambia quali aggettivi risultano significativi, ma non la struttura della discriminazione.
- La chain-of-though riduce le dimensioni significative da 5 a 2, ma amichevole e urbano/a restano significativi.

Key Findings

- GPT-4.1 mini associa sistematicamente chi scrive in dialetto a occupazioni di minor prestigio, tratti come furbo e simpatico, profili rurali e non istruiti.
- Napoletano e siciliano vengono penalizzati più del parmigiano in tutti e tre gli esperimenti, pattern coerente con l'antimeridionalismo.

Il modello riflette le gerarchie sociali nei dati di training.
Questo fenomeno trascende lingue e culture specifiche.

Conclusioni

La tecnologia non inventa nuove disuguaglianze, riproduce e amplifica quelle esistenti.
Le persone penalizzate sono le stesse già discriminate da patriarcato, razzismo, classismo.

Prossimo passo: ***Intersezionalità***
Genere, origine, classe si sovrappongono, creando forme di discriminazione non visibili se misurate separatamente.

Fairness *non va considerata come una feature, ma un requisito di progettazione fin dall'inizio.*

Lottare contro le discriminazioni non è uno slogan, ma una responsabilità.