

---

Università degli Studi di Padova – Dipartimento di Ingegneria dell'Informazione  
Corso di Laurea in Ingegneria dell'Informazione

***Relazione per la prova finale***  
***«Un dataset italiano per l'analisi degli stereotipi di genere nei testi mediante Large Language Models»***

Relatore: Prof. Antonio Rodà  
Correlatore: Prof.ssa Silvana Badaloni

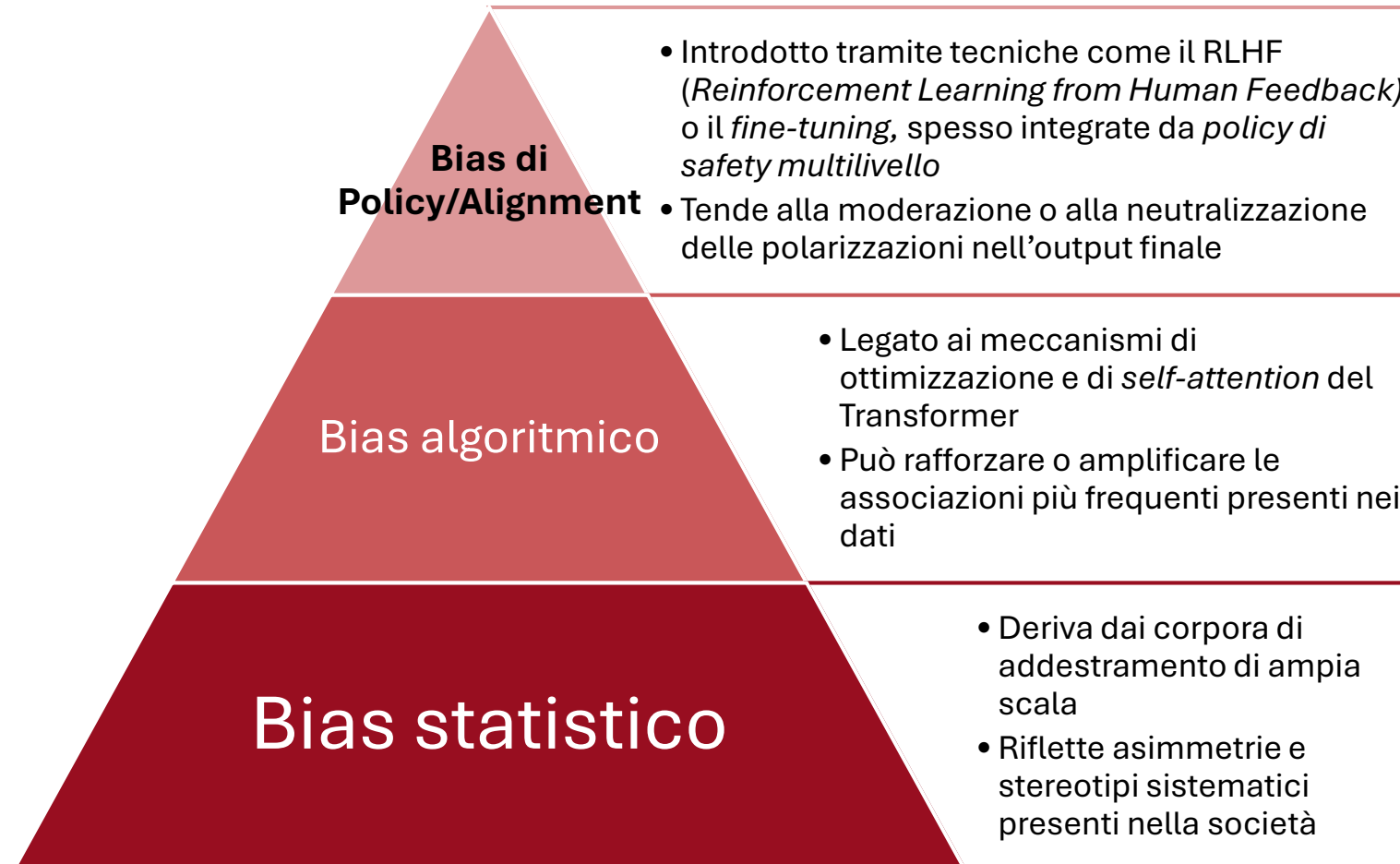
Laureanda: Elisa Famulari  
Matricola n.: 2010940

Padova, 13/03/2026





## Inquadramento Teorico: Genere e Linguaggio

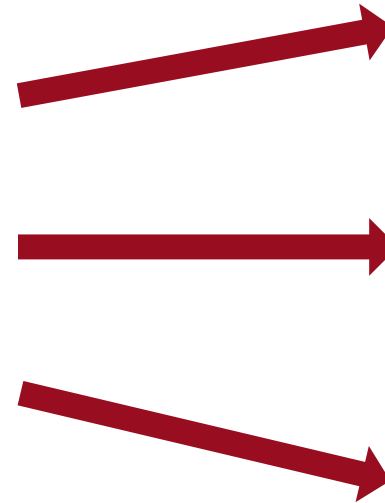



## La stratificazione del Bias negli LLM





## CONFRONTO SISTEMATICO UOMO-LLM

-  Umani
-  Gpt-2.1-mini
-  Gemini-2.5-flash
-  DeepSeek-chat



 Convergenze e divergenze

 Tendenze strutturali

 Intensità e dispersione delle valutazioni

L'obiettivo primario della ricerca non risiede nella mera espansione del dataset, quanto nel suo impiego come strumento per un confronto sistematico tra la percezione umana e la logica dei Large Language Models.

## Struttura del corpus

- Origine: Ampliamento del corpus di Vismara (2024)
- Numerosità: 180 estratti

Categoria	N. Sezioni
Femminile (F)	60
Maschile (M)	60
Neutro (N)	60

## Scala di valutazione

Scala discreta bipolare a 5 livelli:



**-2:** Completamente Femminile

**-1:** Più Femminile che Maschile

**0:** Neutro

**+1:** Più Maschile che Femminile

**+2:** Completamente Maschile



## DESCRIZIONE DEL LAVORO

**Domanda chiave:**  
**«A QUALE GENERE PENSA CHE IL TESTO  
SIA RIVOLTO?»**

### Valutazione umana

Somministrazione del questionario  
con PsyToolkit

Campione finale: 75 partecipanti

Campione bilanciato:  
56% F - 41,3% M

Validazione: esclusione delle  
sezioni con <5 valutazioni, 111  
sezioni finali

### Valutazione dei modelli

Sistemi analizzati: GPT-4.1-mini,  
Gemini-2.5-flash e DeepSeek-chat

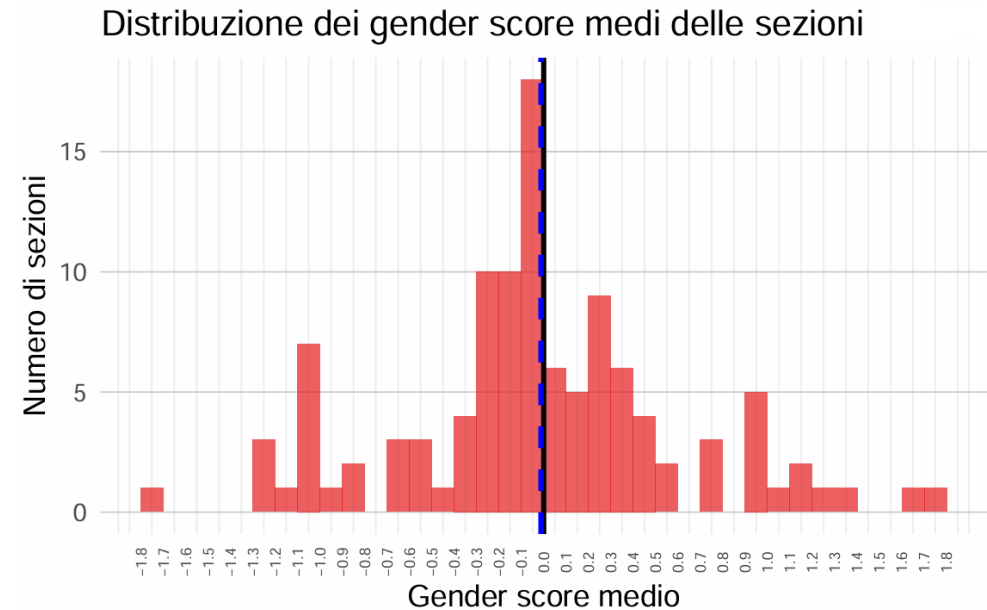
Pipeline Tecnica: Implementazione  
in Python per chiamate API  
autenticare e strutturate

Parametri: 5 run per modello a  
temperature variabili  
(T=0, T=0.3, T=0.5)

Prompt: identico alla domanda  
umana per garantire piena  
comparabilità

## Risultati: La percezione umana

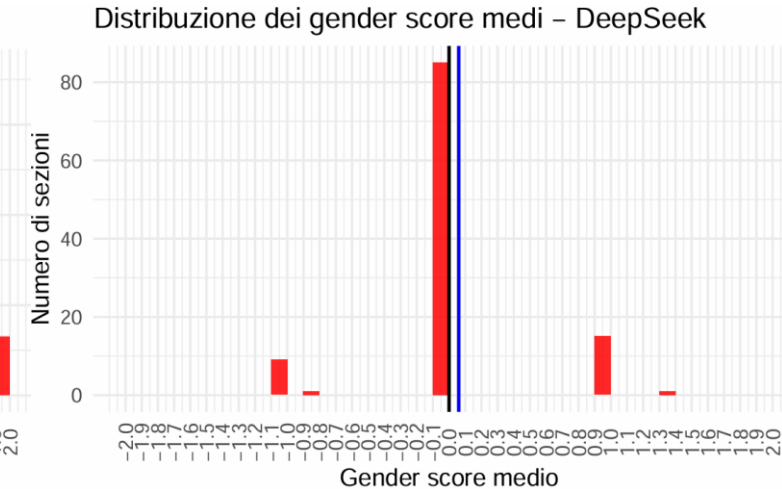
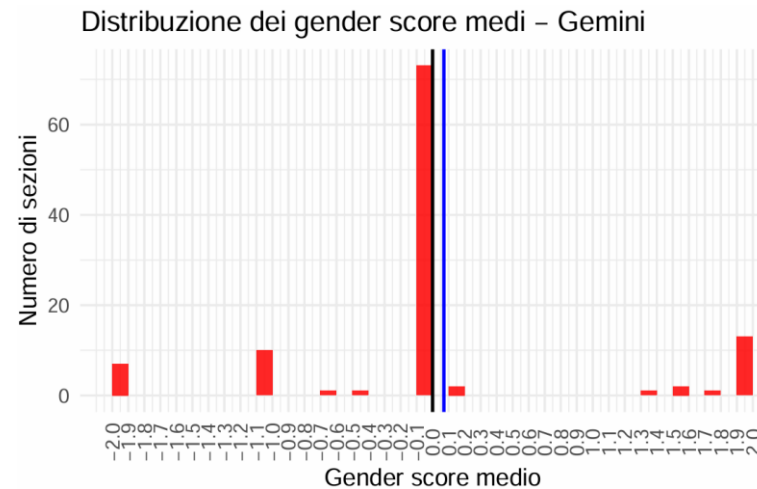
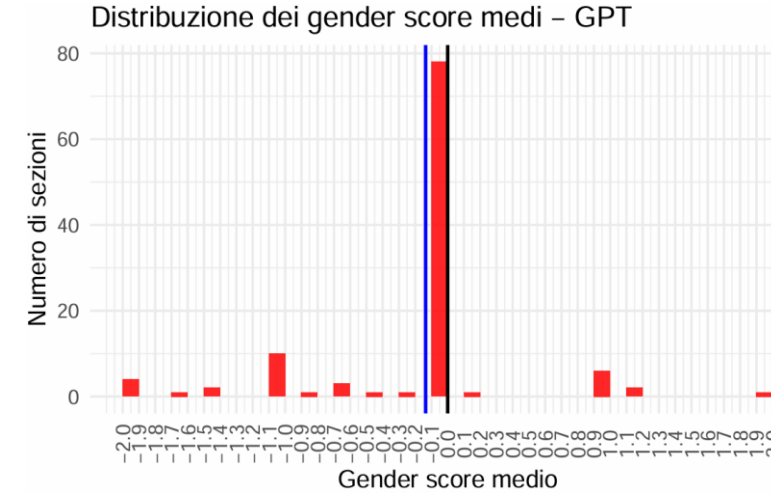
Categoria	N sezioni	Score medio	Deviazione std
Globale	111	-0.012	0.639
F	41	-0.395	0.512
M	39	0.446	<b>0.628</b>
N	31	-0.081	0.409



Distribuzione dei gender score medi delle sezioni valutate. La linea nera indica lo zero della scala, la blu lo score medio della distribuzione globale.

## Risultati: Le prestazioni dei modelli

Modello	Score medio	Deviazione std
GPT-4.1-mini	-0.135	0.642
Gemini-2.5-flash	0.07	0.955
Deepseek-chat	0.059	0.488



## Confronto

Coefficienti di correlazione  
lineari di Pearson a confronto:

Confronto	r
GPT-Gemini	0.12
GPT-DeepSeek	0.10
Gemini-DeepSeek	0.77

Confronto	r
Umano-GPT	0.49
Umano-Gemini	0.59
Umano-DeepSeek	0.61

## Sincronia delle divergenze

*N3\_1: Il primo dato che balza all'occhio è quello delle vittime di sesso femminile. Dal 1 gennaio 2023 al 31 luglio 2024 sono state 175. Conosciamo bene purtroppo i fatti di cronaca dietro queste storie, così come sappiamo che il 31,5% delle donne in Italia ha subito nel corso della propria vita una qualche forma di violenza fisica o sessuale.*

*M3\_3: Invece di risultare un po' piatto (come può accadere a volte con la monocromia), il mix di colori è un ottimo modo per rendere un outfit interessante, aggiungendo l'elemento cruciale e apparentemente spontaneo del "l'ho appena indossato".*

## Discussione



## Limiti :

Campione umano prevalentemente giovane

Assenza di accesso ad i dati di addestramento e alle strategie di alignment

Utilizzo di modelli 'mini'

## Sviluppi futuri:

Estensione del campione umano

Ampliamento del dataset

Confronto con nuovi modelli linguistici

La ricerca evidenzia che l'IA non inventa il bias, ma lo eredita e lo replica con precisione:  
dove l'uomo esita, anche la macchina esita.