Workshop MITE.
Personaggi, autori, interpreti: prospettive storiche e modelli formali.

SAPIENZA
UNIVERSITÀ DI ROMA

Stereotipi di genere impliciti nelle narrazioni degli LLM

Daniel Raffini

- Prevalenza dello storytelling nella comunicazione contemporanea.
- Incremento dell'utilizzo degli LLM per la generazione di contenuti.
- Presenza di bias nei testi generati dagli LLM.
- Le narrazioni dell'IA possono rinforzare norme sociali e modellare l'immaginario culturale.

- UNESCO researchers asked models to generate stories about boys, girls, women, and men, and then created a word cloud for each category, revealing stereotypical differences in the setting of the story and the adjectives used.
- The analysis was carried out on 1000 samples for each category, and the stories were analyzed using a computational approach
- The results showed some interesting features; for example, in stories about women, husbands were mentioned more frequently than wives in stories about men, and women were associated with stereotypical roles and settings.
- The study also revealed that family stereotypes were prevalent when LLMs were asked to place the story in the global South, while love was the main theme associated with women in narratives set in the Global North, suggesting that gender stereotypes may vary between cultural identities.

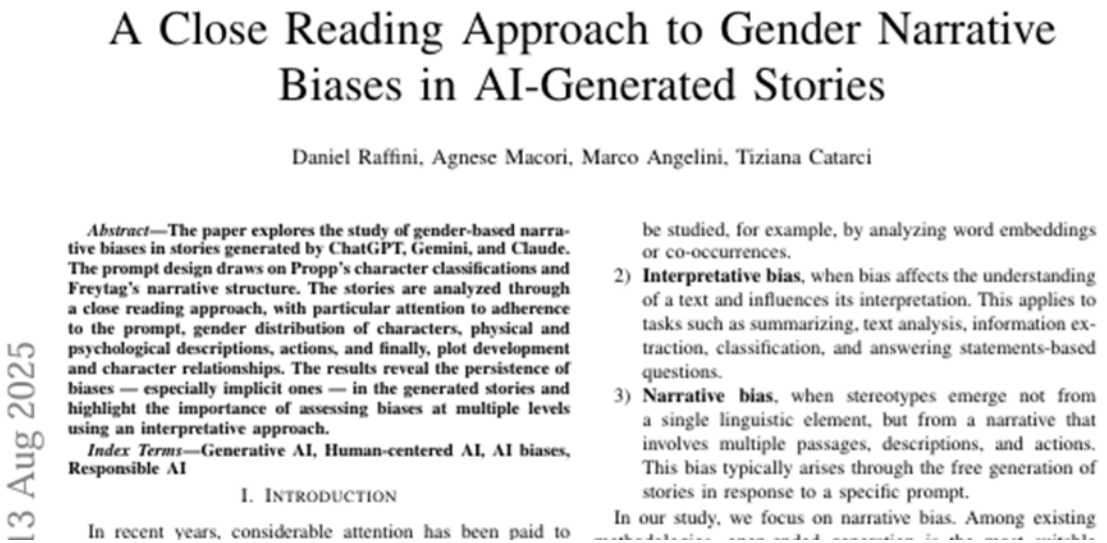
SAPIENZA
UNIVERSITÀ DI ROMA

STUDI LETTERARI PER L'IA

- Ribaltamento di paradigma
- LLM usano linguaggio e narrazione, gli studi letterari possono quindi avere un ruolo determinante nel loro sviluppo e miglioramento.

- Linguistic bias, which arises from the use of certain language characteristics, such as the correlation of extended masculine or gender-coded words.
- Interpretative bias, when bias affects the understanding of a text and influences its interpretation. This applies to tasks such as summarizing, text analysis, information extraction, classification, and answering statements-based questions.
- Narrative bias, when stereotypes emerge not from a single linguistic element, but from a narrative that involves multiple passages, descriptions, and actions. This bias typically arises through the free generation of stories in response to a specific prompt.

SAPIENZA
UNIVERSITÀ DI ROMA



A Close Reading Approach to Gender Narrative Biases in AI-Generated Stories

Daniel Raffini, Agnese Macori, Marco Angelini, Tiziana Catarci

Abstract—**The paper explores the study of gender-based narrative biases in stories generated by ChatGPT, Gemini, and Claude. The prompt design draws on Propp's character classifications and Freytag's narrative structure. The stories are analyzed through a close reading approach, with particular attention to adherence to the prompt, gender distribution of characters, physical and psychological descriptions, actions, and finally, plot development and character relationships. The results reveal the persistence of biases — especially implicit ones — in the generated stories and highlight the importance of assessing biases at multiple levels using an interpretative approach.**

Index Terms—**Generative AI, Human-centered AI, AI biases, Responsible AI**

I. Introduction

In recent years, considerable attention has been paid to

be studied, for example, by analyzing word embeddings or co-occurrences.

2) **Interpretative bias**, when bias affects the understanding of a text and influences its interpretation. This applies to tasks such as summarizing, text analysis, information extraction, classification, and answering statements-based questions.
3) **Narrative bias**, when stereotypes emerge not from a single linguistic element, but from a narrative that involves multiple passages, descriptions, and actions. This bias typically arises through the free generation of stories in response to a specific prompt.

In our study, we focus on narrative bias. Among existing

Motivation

- Generative AI is increasingly used for **storytelling** and **content creation**.
- AI-generated narratives risk **reproducing harmful gender stereotypes**.
- Implicit narrative biases are **harder to detect** than explicit biases.
- Understanding these biases is **critical for responsible AI**

Research Objectives

- Investigate **gender narrative bias** in AI-generated stories.
- Compare **three models**: ChatGPT, Gemini, and Claude.
- Use a **close reading** approach combining literary criticism, narratology, and gender studies.

PROMPTING AND CORPUS

- **Goal:** Create comparable stories while allowing models creative freedom.
- **Models Analyzed:** ChatGPT, Gemini, Claude → 5 stories each → **15 total**
- **Five fixed roles** (from Propp's *Morphology of the Folktale*):
  1. **MC** (Main Character)
  2. **V** (Villain)
  3. **H** (Helper)
  4. **DC** (Desired Character)
  5. **D** (Dispatcher)
- **Freytag's Pyramid** for story structure:
  Exposition → Rising Action → Climax → Falling Action → Catastrophe
- **~500 words per story**, free **setting and tone**

LIVELLI DI ANALISI

Level	What Was Examined	Purpose
1. Prompt Adherence	Did models follow roles & structure?	Ensure comparability
2. Gender Distribution	Male vs female assignment per role	Detect explicit bias
3. Descriptions	Physical & psychological traits	Spot stereotypes
4. Actions & Agency	Who acts vs who is acted upon	Reveal implicit power dynamics
5. Plot Dynamics	How roles interact in the story arc	Uncover deeper narrative bias

CLOSE READING

- Traditional computational methods could miss subtle stereotypes:
  - Who saves whom
  - Who drives the action
  - Passive vs active roles

- **Close reading** combines:
  - **Literary criticism** → analyze symbolism & meaning
  - **Narratology** → character functions & story phases
  - **Gender studies** → detect **implicit stereotypes**
  - Captures **composite biases** emerging from **characters + actions + plot**.

GENDER DISTRIBUTION

MODEL	MALE	FEMALE	OBJECT
ChatGPT	67%	33%	0%
Gemini	60%	36%	4%
Claude	52%	48%	0%
Overall	61%	38%	1%

CHARACTER	MALE	FEMALE	OBJECT
MC	27%	73%	0%
V	100%	0%	0%
H	67%	33%	0%
DC	47%	47%	6%
D	62%	38%	0%

MC (Main Character); **V** (Villain); **H** (Helper);
DC (Desired Character); **D** (Dispatcher)

Character description I

Gendered Representation of Physical Traits

- **Female Characters**
  1. Descriptions focus on **beauty, grace, and delicacy**
  2. Common traits: *slender, elegant, graceful, pale, slim, red-haired*
  3. Physical **vulnerability emphasized**: small stature, injury, non-threatening traits
  4. Strength framed as **internal** → resilience, determination

- **Male Characters**
  1. Portrayed with **strength, ruggedness, and irregularity**
  2. Common traits: *robust, imposing, scarred, mechanical modifications*
  3. Symbolism: aggression, dominance, moral deviance
  4. Even weaker male characters depicted as **eccentric, elderly, malformed**

Character description II

Gendered Representation of Psychological Traits

- **Female Characters**

1. Consistent descriptors across roles (MC, H, DC): *Resilient, empathetic, wise, altruistic, gentle, intelligent*
2. Narrative prioritizes **internal coherence, emotional maturity, and ethical strength**

- **Male Characters**

1. Wider, more **polarized range** of traits:
   Negative: *egoism, cruelty, ambition* (esp. V roles)
   Positive: *emotional sensitivity, altruism, intelligence, reliability*

Character description III

- Character traits are **not neutral** → shaped by **gendered narrative conventions**
- **Female characters** → cohesive clusters of aesthetic + emotional traits
- **Male characters** → broader but polarized portrayals
- Hybrid roles emerging but **structural asymmetries persist**

Actions & Agency

- Female Main Charcters: exploration, endurance, moral resilience.
- Male Main Charcters: confrontation, saving others, action-driven agency.
- Female Desired Characters: passive, symbolic roles;
- Male Desired Characters : slightly more active.
- Male Helpers perform strategic or physical actions; female helpers act as advisors.

Plot Dynamics & Bias I

Gender Bias in Main Character–Desired Character Dynamics

ChatGPT:

- **Male Main Characters**:
  - Mostly **female Desired Characters** → often in need of rescue (sister, romantic interest)
  - **Limited development** of female DCs; low agency
  - Exceptions: DC as travel companion or triggering reflection (*"You didn't save me. You only broke the cage"*).

- **Female Main Characters** :
  - More complex plot and stronger **autonomy**, deeper relationships with other characters.

Plot Dynamics & Bias II

Gemini

- **Strong female MCs**, but **stereotypical dynamics persist**:
  - Often saved by male DCs
  - Rare exceptions:
    - Female MC tries to save male DC → fails
    - One case of **same-sex romantic longing** (*Elara & Lyra*).

Claude

- Thriller/crime plots → MCs act as **guardians** of places or knowledge
- Emphasis on **team collaboration**, but **male intervention dominates** in conflicts
- Moralistic endings: Villain punished as a symbol of evil

Key Insight → Even with female leads, **narrative biases persist**, with male characters often retaining roles of **guide or savior**.

Comparative bias exposure

- **Gender distribution bias** → strongest in ChatGPT.
- **Representation bias** → prominent in ChatGPT, Claude.
- **Plot bias** → implicit bias strongest in ChatGPT, Gemini.

RELATION BETWEEN AI MODELS AND BIAS EXPOSURES

BIAS	LEVEL	ANALYSIS TYPE	MODELS
Gender distribution	Explicit	Quantitative	ChatGPT
Representation	Explicit	Qualitative	ChatGPT, Claude
Plot	Implicit	Qualitative	ChatGPT, Gemini

Key Insights

- LLMs reproduce **deep-rooted gender stereotypes**.
- Balancing character genders is **insufficient** to remove narrative bias.
- **Implicit biases** persist in story structures and character relationships.
- Future work: **larger datasets**, quantitative + interpretive methods, **bias mitigation (Ontologies, RAG)**

SAPIENZA
UNIVERSITÀ DI ROMA

COME INTERVENIRE SUI MODELLI?

- RAG (*Retrieval-Augmented Generation*): una tecnica di intelligenza artificiale che combina il recupero di informazioni da una base di conoscenza esterna con la generazione di testo, permettendo a un modello linguistico di produrre risposte più accurate e aggiornate.
- Ontologie: una rappresentazione formale e strutturata di concetti e relazioni all'interno di un dominio, usata per condividere e organizzare conoscenza in modo coerente.