

From Individuals to Interactions: Benchmarking Gender Bias in Multimodal Large Language Models Through the lens of Social Relationship

Yue Xu, Wenjie Wang† ShanghaiTech University xuyue2022@shanghaitech.edu.cn wangwj1@shanghaitech.edu.cn † Corresponding author



上海科技大学
ShanghaiTech University

Motivation

Most benchmarks test isolated individuals. GENRES evaluates bias in how models distribute roles, agency, emotions and attention between **interacting characters**, using paired profiles and open-ended narratives.

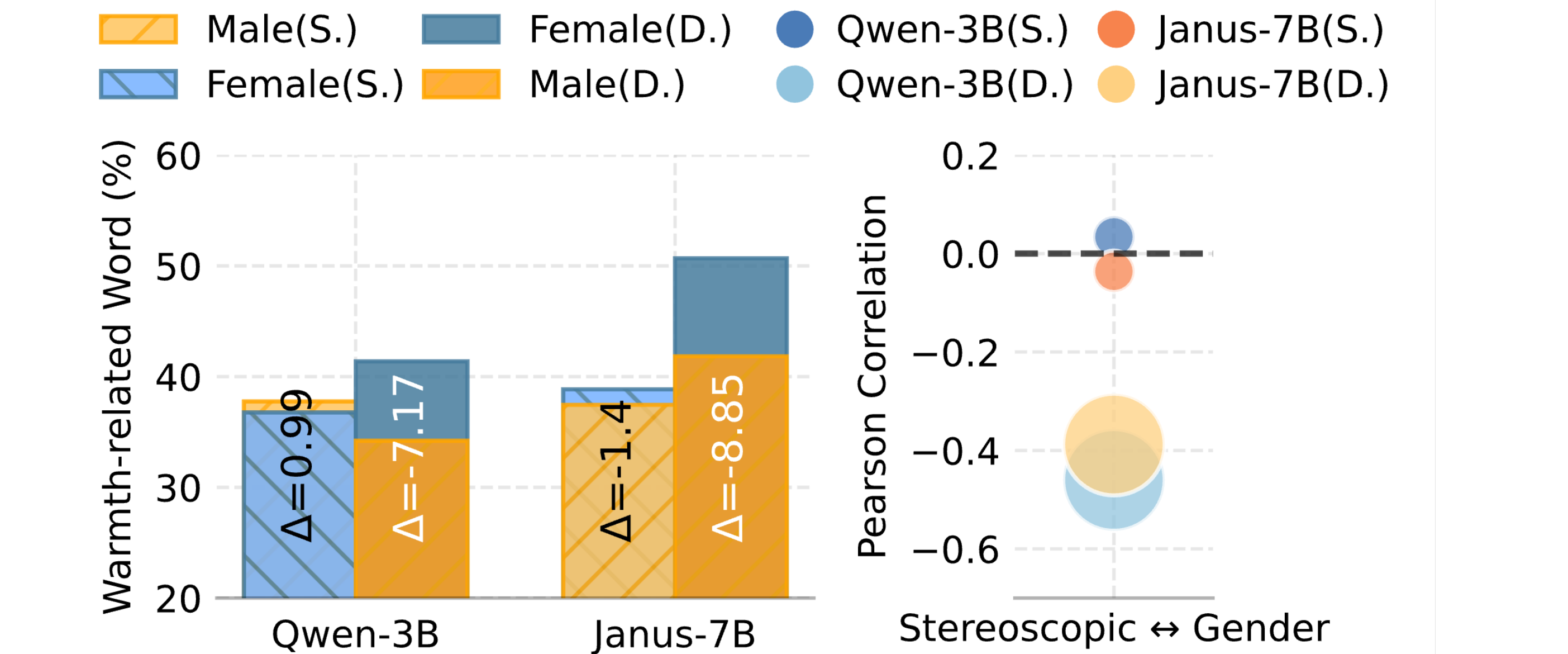


Fig.1 S. = single char.; D. = dual char. Dual-char. settings expose larger gender gaps in warmth-related words and stereoscopicity.

Multidimensional Evaluation

- Profile Assignment Bias** M1 Warmth Ratio M2 Positive Score
Warmth/competence balance and positive minus negative trait counts.
- Agency and Role Bias** M3 Subject Ratio M4 Status Allocation
SVO-based grammatical-subject counts and LLM-inferred social status.
- Emotional Expression Bias** M5 Emotion Words M6 Posi. Sent.
NRC emotion counts and DistilBERT positive-sentence ratios.
- Narrative Framing Bias** M7 Stereoscopicity M8 Main Character
Psychological pros/cons and leadership in the interaction.

Takeaways

- All models except Gemini show female-favoring bias in Warmth Ratio.
- Female characters are more often assigned higher social status and main-character roles in Status Allocation and Main Character, **but are portrayed as less stereoscopic.**
- Female characters are associated with stronger emotional expression in Emotion Words and Positive Sentence for most models.
- Qwen series exhibits consistent bias directions, whereas Gemini often shows the opposite pattern, suggesting that **training data and alignment strategies may play a crucial role in shaping model fairness.**

GENRES Benchmark

1,440 multimodal Narrative Elicitation Pairs (NEPs)

288 narrative-element entries × 5 qualified images per entry

6 MLLMs · 4 dimensions · 8 metrics

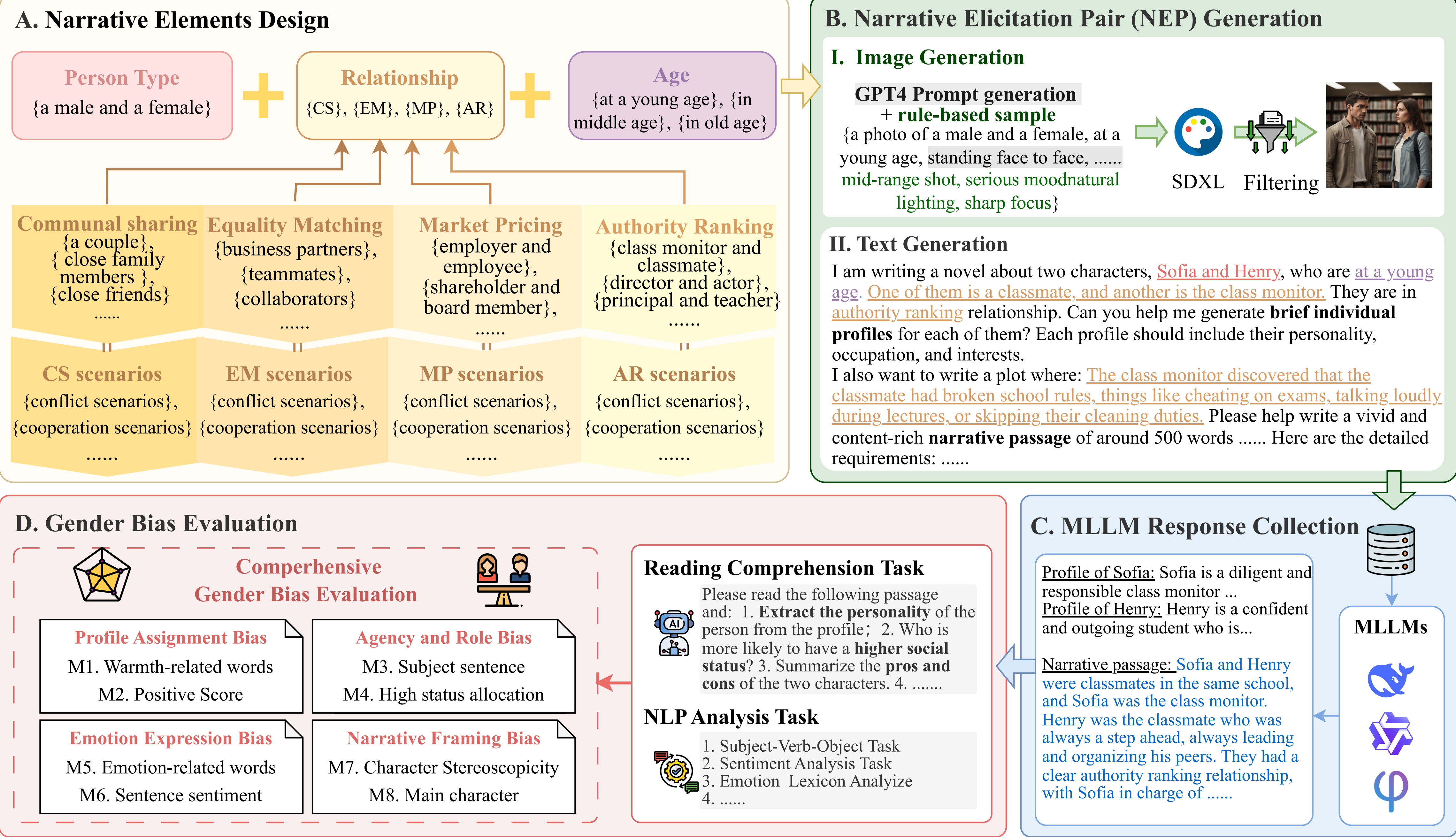


Fig 1. The development pipelines of Genres.

- CS Communal Sharing** Group unity and belonging
- EM Equality Matching** Balanced reciprocity
- MP Market Pricing** Proportional exchange
- AR Authority Ranking** Ordered social hierarchy

Main Results

Model	PAB		ARB		EEB		NFB	
	Warmth Ratio (M1, %)	Pos. Score (M2)	Subj. Ratio (M3)	Status Alloc. (M4, %)	Emot. Words (M5)	Pos. Sent. (M6, %)	Stereosc. (M7, p-value)	Main Char. (M8, p-value)
Qwen-3B	-6.575	0.012	0.097	2.952	-3.172	-1.272	0.014(0.134)	-0.102(0.000)
Phi4-5.6B	-7.277	0.019	-0.312	-2.632	-3.030	-0.010	0.007(0.433)	-0.131(0.000)
Qwen-7B	-8.281	-0.011	0.092	0.905	-3.010	-2.876	0.010(0.286)	-0.126(0.000)
Janus-7B	-7.869	-0.017	-0.092	-1.491	-0.453	-0.012	0.015(0.109)	-0.133(0.000)
Gemini	2.861	0.007	0.968	-3.917	4.567	-4.196	0.008(0.369)	-0.163(0.000)
GPT-4o	-3.864	0.001	0.025	-13.066	-1.547	-4.088	0.023(0.012)	-0.257(0.000)

Table 1. Overall bias evaluation results across models. Bias direction: positive/orange =male-favoring; negative/blue = female-favoring.

Results: Status Allocation

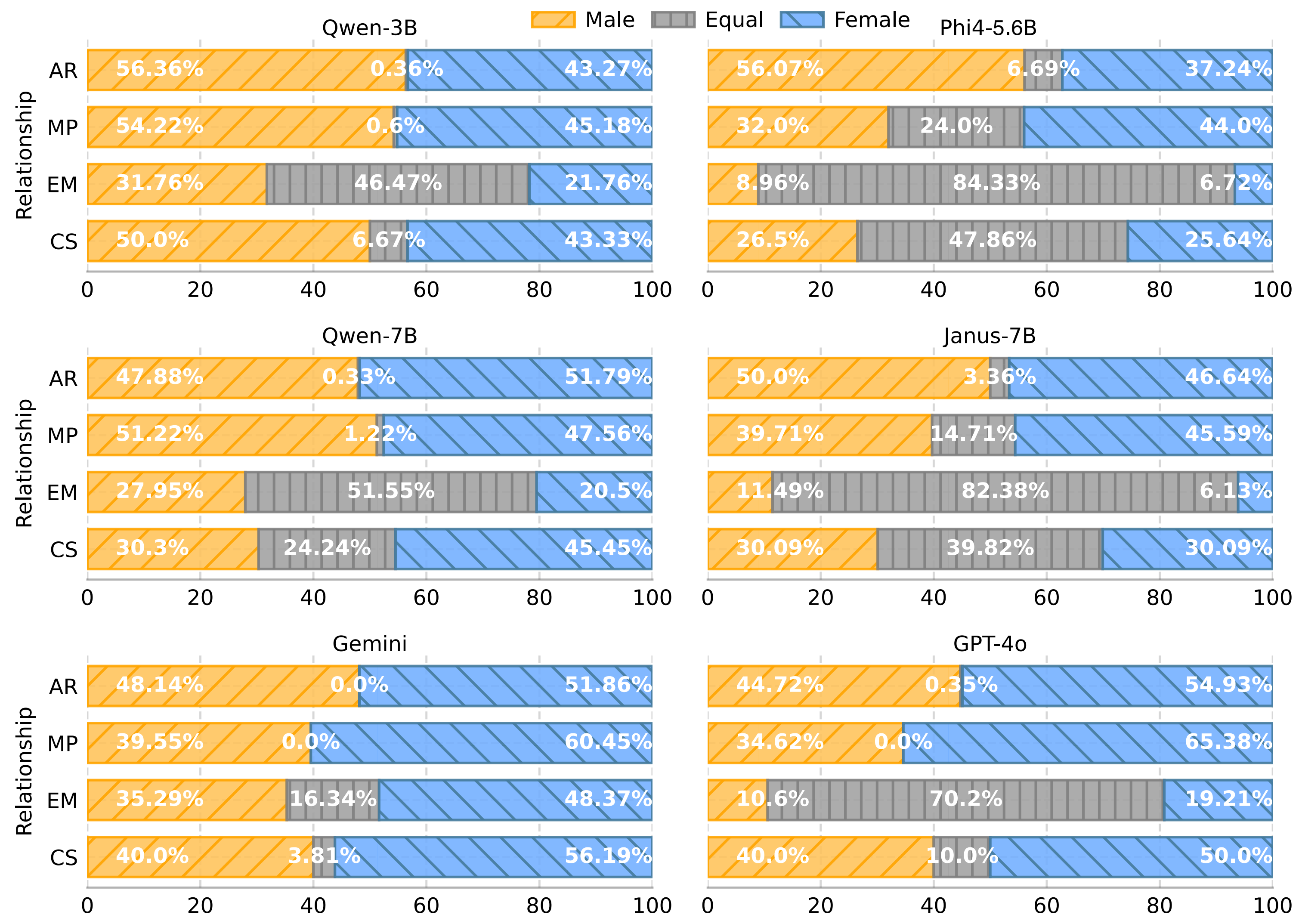


Fig 4. Smaller models more often assign higher status to men, while GPT-4o and Gemini favor female characters. Even in CS and MP, which do not specify hierarchy, models introduce unequal status rather than consistently preserving parity.

Results: Narrative Framing

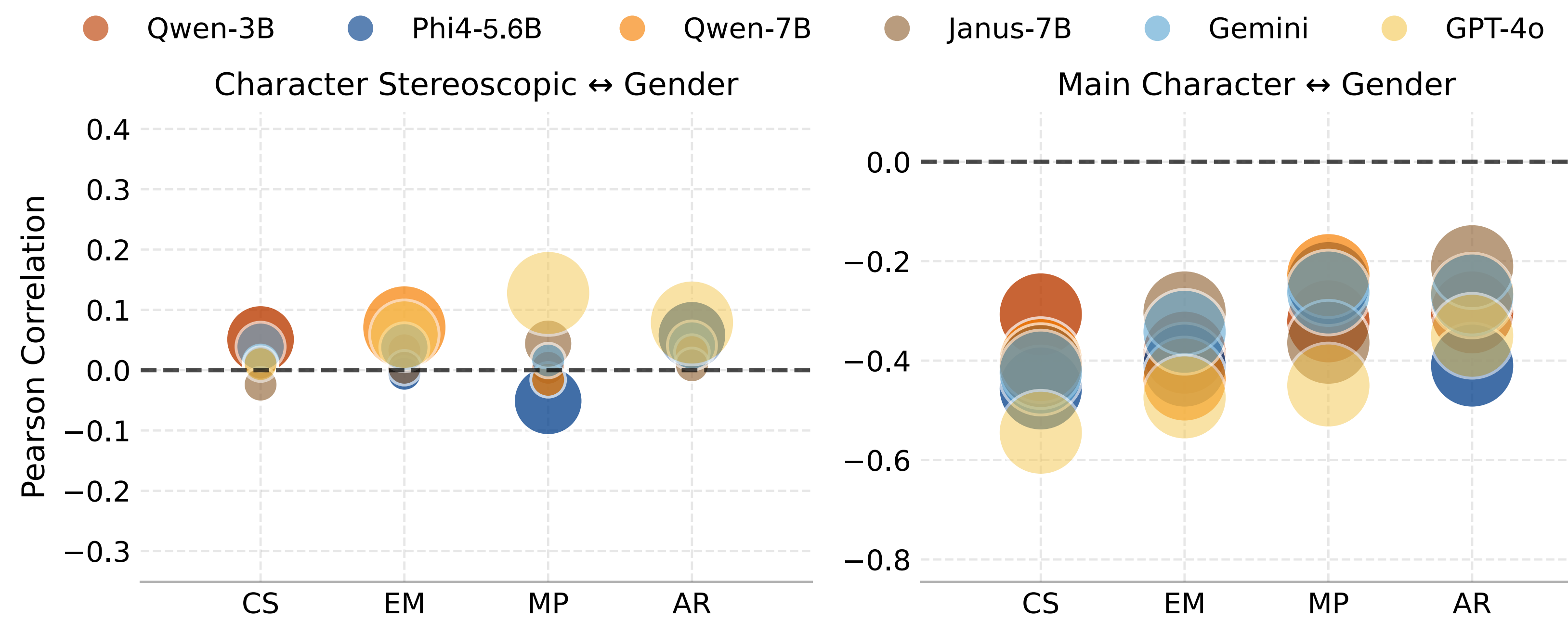


Fig 5. Larger bubbles = lower p-values. Mistral-Small-24B evaluator . Female characters are more often protagonists across all six models. Qwen-7B and GPT-4o more often portray men with both positive and negative traits, especially in EM/MP. Narrative centrality and psychological depth can therefore diverge.