Issues in Information Systems
Volume 27, Issue 3, pp. 250-264, 2026

DOI: https://doi.org/10.48009/3_iis_2026_118

When AI tells the story: Systematic stereotyping in AI-generated narratives

Miaoyi Zeng, *University of Wisconsin - Whitewater, zengm@uww.edu*
Meng Tian, *Northern State University, meng.tian@northern.edu*

Abstract

As Artificial Intelligence (AI) tools are increasingly deployed across digital platforms to generate content about people, questions arise about whether AI-generated narratives exhibit systematic representational biases. This study examines whether AI-generated narratives show structured framing differences across social role contexts and gender, and whether such patterns differ across three leading commercial AI tools. Using a controlled prompt experiment, we generated 469 narratives from ChatGPT (GPT-4o), Claude (Opus 4.6), and Gemini (2.5 Pro), varying social role context (professional vs. domestic) and character gender (man vs. woman). Each narrative was scored using the warmth and competence dimensions derived from the Stereotype Content Model. The results reveal three notable patterns. First, the AI models told a different story depending on the context in which each character was placed: professional narratives were consistently high in competence and low in warmth, while domestic narratives received the reverse. Second, gender shaped this pattern specifically in professional contexts, where female characters received notably warmer portrayals than male characters despite similarly high levels of competence. Third, although all three models exhibited similar framing patterns, they differed in intensity. ChatGPT generated the warmest narratives overall and assigned higher competence portrayals to domestic characters than either Claude or Gemini, whereas Gemini generated the coldest portrayals overall. These findings suggest that AI-generated narratives encode structured stereotype-consistent framing, that gender and social role context jointly shape those portrayals, and that AI tool choice carries representational consequences that are rarely considered in practice.

Keywords: AI-generated narratives, gender stereotyping, social role context, large language models, stereotype content model

Introduction

Artificial Intelligence (AI) tools such as ChatGPT, Claude, and Gemini are increasingly being deployed as storytellers across digital platforms. They are used to draft news summaries, write biographical sketches, generate educational materials, and produce content about people at a scale no human workforce could match. According to Henriksson (2023), half of the newsrooms worldwide now use AI tools in content production, and that number is growing dramatically. Most audiences encounter AI-generated content without knowing an AI produced it, and even professional journalists cannot reliably identify the origin of an article with any statistically significant accuracy (Kotenidis & Veglis, 2021). As these systems take on a growing share of the informational content that people encounter, a fundamental question arises: when AI tells a story about a person, does it tell it neutrally?

While this question may seem unimportant, the stakes are anything but trivial. Decades of media research have shown that consistent exposure to stereotyped portrayals of individuals shapes how audiences perceive various social groups over time (Gerbner et al., 1986). When specific roles are consistently framed in the same ways across the content people consume, those patterns accumulate into social expectations about

who belongs (and who does not) and who is capable of what. For example, in the U.S. decades of movies depicting workers in the computing fields have created the expectation of a young, white male, who is a socially awkward loner (Cheryan et al., 2013). If the stories generated by AI follow a stereotyped logic, the consequences for how audiences perceive social groups may be largely invisible but have a substantial societal impact.

Prior research has established that AI systems can encode social biases into their outputs (Bender et al., 2021; Bolukbasi et al., 2016; Sargent et al., 2024). However, most prior work has focused primarily on whether biased or stereotype-consistent portrayals are present, rather than examining whether AI-generated narratives follow structured and predictable stereotype-consistent patterns. Specifically, when AI generates stories about individuals across different social role contexts, does it portray them differently in systematic ways? And does gender shape those patterns, such that the same role context produces different portrayals depending on whether the character is a man or a woman? These questions remain largely unexplored.

A further question concerns which AI is doing the storytelling. ChatGPT, Claude, and Gemini are distinct products developed by different companies with different design priorities. Each of them has been shaped differently by its developers, meaning the same prompt could produce meaningfully different narratives depending on which model generates it. Even if all three AI tools produce stereotyped narratives, they may not do so in the same way or to the same degree. Thus, AI tool choice, a decision users and organizations often make without considering its representational implications, may matter more than is commonly recognized.

This study seeks to answer these questions. To do this we generated a corpus of narratives from three leading AI tools (ChatGPT [GPT-4o], Claude [Opus 4.6], and Gemini [2.5 Pro]) across two social roles (professional and domestic) and two genders (male and female), and analyze the resulting stories for structure and consistency in how characters are portrayed. The results indicate that AI-generated narratives exhibit systematic stereotype-consistent framing patterns across social role contexts. Specifically, professional and domestic characters are described in consistently and markedly different ways, reflecting distinct patterns of trait attribution. These differences are further shaped by gender, which systematically influences how warmth is assigned across these roles. The AI tools, however, do not tell identical stories. For warmth (friendliness, sincerity, communal orientation), all three AI tools apply the same domain logic but differ in the level of warmth. Gemini produced the coldest narratives, whereas ChatGPT produced the warmest. For competence (intelligence, skill, agency), all three tools generated narratives portraying professional characters as highly competent. Where they parted ways was in how they portrayed domestic characters. ChatGPT rated domestic characters "relatively" competent, while Claude and Gemini rated domestic characters as lower in terms of competence.

By examining extended narratives rather than isolated words or sentences, this study demonstrates that stereotype-consistent framing in AI-generated content can emerge at the narrative level. By analyzing both gender and social role context simultaneously, the study further shows that representational patterns vary in more nuanced ways than a simple professional-versus-domestic distinction alone. Finally, by directly comparing three leading AI tools, the findings suggest that AI tool selection carries meaningful representational consequences. The remainder of this paper is organized as follows. The next section reviews relevant literature and develops the theoretical background. The Methodology section describes the data collection and analysis procedures. The Results section presents findings for each research question, followed by a Discussion of theoretical and practical implications, Limitations, and a Conclusion.

Theoretical Background and Research Questions

Bias and Stereotypes in AI-Generated Narratives

Existing research documents pervasive biases in AI-generated content, showing that AI can produce unequal representations and outcomes across social groups (e.g., Bansal et al., 2023; Ferrara, 2024; Wan et

al., 2023). For example, Sheng et al. (2019) showed that AI tools produce more negative outputs for prompts referencing women and minority groups; and Kirk et al. (2021) found that AI systems exhibited intersectional occupational biases. These findings established the foundation for AI bias research, though they are based on model generations and architectures that predate current commercial systems.

However, two limitations are consistently found in this literature. The first limitation is the unit of analysis, as the majority of the research has examined bias at the level of individual word(s), short text responses, or tightly controlled sentence prompts (Bolukbasi et al., 2016; Sheng et al., 2019). This unit of analysis identifies the traits an AI tool associates with a given social or demographic group, but it does not capture whether these associations coalesce into a broader narrative that systematically shapes how a character is portrayed.

The second limitation concerns the nature of the question being asked. Most prior studies focus primarily on whether AI bias is present, rather than examining whether the bias produced has a structured, predictable pattern. Knowing that an AI tool produces biased output is different from understanding whether the output contains a pattern of bias across contexts (Blodgett et al., 2020).

This study addresses both limitations in extant research by examining AI-generated narratives as a broader unit of analysis, and moving beyond acknowledging the existence of bias, to study whether AI-generated narratives follow systematic stereotype-consistent patterns across social role contexts, gender, and leading AI tools.

The Stereotype Content Model as a Measurement Tool

Although AI bias can take many forms, the present study focuses specifically on stereotype-consistent portrayals in AI-generated narratives. Stereotypes refer to generalized associations attached to particular social groups or roles and often shape how individuals are represented in communication and media contexts. To examine whether stereotyping in AI-generated narratives follows structured and recognizable stereotype patterns, we need a framework that can capture the structure of social perception in text. The Stereotype Content Model (SCM; Fiske et al., 2002) provides exactly that.

The SCM proposes that perceptions of social groups are organized along two fundamental dimensions: warmth (friendliness, sincerity, communal orientation) and competence (intelligence, skill, agency). These dimensions have been validated across decades of research and vary systematically across social roles and gender. Professional settings tend to elicit competence-dominant portrayals (i.e., the way a person or group is characterized through descriptions, traits, and behaviors), while domestic contexts tend to elicit warmth-dominant portrayals (Cuddy et al., 2008; Eagly & Karau, 2002; Mastro & Greenberg, 2000). In this study, domestic context refers to private and family-oriented social settings rather than formal workplace or professional environments, consistent with distinctions commonly discussed in social role research (Eagly & Karau, 2002). Gender patterns have also been consistently documented across cultures and media formats, with women more frequently associated with warmth and communal orientation and men more often associated with competence and agency (Fiske et al., 2002; Koenig & Eagly, 2014).

Using the warmth and competence dimensions of the SCM as an analytical lens, we evaluate each narrative based on the extent to which it emphasizes warmth-related and competence-related traits. Narratives emphasizing capability, achievement, decisiveness, and independence are interpreted as reflecting stronger competence framing, whereas narratives emphasizing care, emotional support, relationships, and communal orientation are interpreted as reflecting stronger warmth framing.

Whether modern AI systems reproduce, reverse, or modify these established social role and gender patterns remains an open empirical question, particularly given that current commercial models have undergone explicit alignment and safety training that may influence the direction of representational differences. This study asks whether these patterns appear systematically in AI-generated narratives and whether they differ across AI tools.

Inter-Model/Tool Variation

AI tools are often treated as a single technology, but ChatGPT, Claude, and Gemini are distinct systems developed by different companies with different design priorities. Each AI tool has been deliberately shaped by its developers through post-training adjustments, processes intended to make the tool more helpful, safe, or consistent with certain standards of output (Ouyang et al., 2022). Although many aspects of these alignment processes remain proprietary, such design choices can meaningfully influence AI-generated outputs.

These differences raise the possibility that representational patterns in AI-generated narratives may vary systematically across AI tools. If the three AI tools examined in this study were aligned similarly, stereotype-consistent portrayals would be expected to appear relatively consistently across systems. However, differences in training data, alignment strategies, and platform objectives may lead the same prompt to generate meaningfully different narratives.

Prior research on AI bias and stereotyping frequently examines a single AI tool, implicitly treating its findings as broadly representative of AI systems more generally. In contrast, this study treats AI tool choice as an important variable rather than merely a methodological detail. Specifically, the study examines whether the structure and intensity of stereotype-consistent portrayals differ across ChatGPT, Claude, and Gemini. In doing so, the study evaluates whether AI tool choice carries meaningful consequences for the representational content of AI-generated narratives.

Using warmth and competence as the analytical dimensions, this study asks:

Research Question 1: How do AI-generated narratives exhibit systematic framing differences across (a) social role contexts and (b) gender?

Research Question 2: How do ChatGPT (GPT-4o), Claude (Opus 4.6), and Gemini (2.5 Pro) differ in the level and pattern of stereotyping in their generated narratives?

Methodology

Data Collection

We designed a controlled prompt experiment to elicit AI-generated narratives under systematically varied conditions. The study employed a 2 (gender: man, woman) × 2 (domain: professional, domestic) × 3 (AI tool/model: Claude Opus 4.6, ChatGPT GPT-4o, Gemini 2.5 Pro) design, yielding 12 unique prompt cells. All narratives featured American individuals as story characters, with a single male name (James Smith) and female name (Mary Smith) used across all conditions to hold cultural context and name-level variation constant, isolating domain and gender as the primary variables of interest. Each cell was administered 40 times as independent replicates, producing a target corpus of 480 texts.

All prompts used a standardized template: "Write a 150-word journalistic narrative about [NAME], an American [GENDER], [DOMAIN PHRASE]. Use third-person narration and a journalistic register. Write in English." The professional domain phrase was "working in his/her professional capacity"; the domestic domain phrase was "in a domestic or family context." The professional and domestic domain phrases were selected to operationalize the social role contexts central to SCM research, consistent with established practice in gender stereotype studies using context-based experimental manipulations (Fiske, 2018; McArthur & Resko, 1975).

Texts were collected via each AI tool's API with rate-limiting and automatic retry logic. All API calls used each provider's default temperature setting (1.0 across all three platforms) and no additional system prompt beyond the generation instruction, to minimize non-task variation. No memory, personalization, or account-specific settings were active; data collection was conducted programmatically via each provider's API rather

than through consumer-facing interfaces, ensuring that account history and user-level customizations did not influence outputs. Data were collected between April 18, 2026 and April 19, 2026 using the following model versions: Claude: claude-opus-4-6 (Anthropic); GPT: gpt-4o (OpenAI); and Gemini: gemini-2.5-pro (Google). Because commercial AI models are updated frequently, these version identifiers are reported to facilitate reproducibility. After excluding texts that returned generation errors (e.g., API overload or empty outputs), the final analytic corpus comprised 469 narratives (Claude: n = 150; ChatGPT: n = 160; Gemini: n = 159).

Each narrative was scored on two dimensions from the Stereotype Content Model (SCM): warmth and competence. The warmth dimension reflects the extent to which a narrative emphasizes communal, relational, caring, and emotionally supportive characteristics, whereas the competence dimension reflects the extent to which a narrative emphasizes capability, agency, achievement, independence, and effectiveness. Each dimension was evaluated on a 1-5 Likert-type scale, with higher values indicating a stronger emphasis on the corresponding dimension. Scoring was conducted using automated API calls to Claude Opus 4.6 (Anthropic), with each narrative submitted programmatically via the Anthropic Messages API. A structured system prompt instructed the AI tool to act as a trained content analyst and return only two integer scores, one for warmth and one for competence. No rationale or additional output was requested. To reduce the influence of explicit demographic cues during scoring, identifying information (including character names, nationality references, and domain labels) was removed from each narrative prior to submission. The tool therefore received only the narrative text itself and was instructed to evaluate the extent to which warmth and competence-related characteristics were reflected in the narrative content. Scores were collected with checkpoint-based resumability to allow interrupted runs to continue without re-scoring completed texts.

The use of AI-assisted coding is consistent with emerging research suggesting that large language models can perform psychometric text analysis with levels of reliability comparable to traditional NLP-based approaches and human evaluation when appropriately validated (Mousavi et al., 2026). Recent studies have similarly reported substantial convergence between AI based evaluations and expert human judgments across a variety of text evaluation tasks (Chiang & Lee, 2023).

To further assess the validity of the AI-based coding procedure, a validation subsample of 48 narratives was selected through stratified random sampling across the 12 experimental cells, with 4 narratives randomly drawn from each cell. Five native English speakers who were not authors of this study served as human raters, trained on the rating rubric prior to rating. Inter-rater reliability of the averaged human ratings, estimated using intraclass correlation coefficients (ICC), was 0.896 for warmth and 0.635 for competence. ICC between Claude-generated and human-averaged scores was 0.988 for warmth and 0.967 for competence, indicating strong agreement between automated and human evaluations.

Data Analysis

Analyses were conducted in SPSS (v. 29) using a two-stage approach corresponding to the two research questions. To address concerns about Type I error inflation arising from the use of two correlated outcome measures, a Bonferroni-adjusted significance threshold of $\alpha = 0.025$ per test was applied. Because Levene's tests indicated heterogeneity of error variances across cells for both dependent variables (all $ps < 0.001$), results were interpreted with appropriate caution; however, ANOVA is considered robust to this violation when cell sizes are approximately equal (ns ≈ 117-118 per gender × domain cell), which was satisfied in the data.

To assess whether AI-generated narratives exhibit systematic stereotyping (RQ1), a 2 (Gender: Man, Woman) × 2 (Domain: Professional, Domestic) between-subjects univariate analysis of variance (ANOVA) was conducted separately for each dependent variable. This design allowed for the examination of main effects for gender and domain as well as their interaction, providing a direct test of the SCM-predicted warmth-competence trade-off. Partial eta-squared (η^2p) was reported as the effect size index, with values of 0.01, 0.06, and 0.14 interpreted as small, medium, and large, respectively (Cohen, 1988).

To examine whether the three AI tools differed in the level and pattern of stereotyping (RQ2), a 3 (Model/Tool: Claude, ChatGPT, Gemini) × 2 (Gender) × 2 (Domain) factorial ANOVA was estimated separately for each dependent variable, expanding the RQ1 design with Model as an additional between-subjects factor. The primary effects of theoretical interest were the Model main effect, the Model × Domain interaction, and the Model × Gender interaction. Significant main effects involving Model (which has three levels) were followed up using Bonferroni-adjusted pairwise comparisons to identify specific between-model differences. Where significant two-way interactions involving Model were observed, simple effects analysis was conducted by decomposing the interaction at each level of the relevant factor, with Bonferroni correction applied to pairwise comparisons within each section.

Results

Descriptive Statistics

Means and standard deviations for warmth and competence, by gender, domain, and model, are presented in Tables 1 and 2, respectively. Across all conditions, warmth ratings ranged from a low of M = 1.75 (SD = 0.84) for Gemini-generated, male professional narratives to a high of M = 4.95 (SD = 0.22) for ChatGPT-generated, female domestic narratives. Competence ratings were near the scale ceiling for professional narratives (range: M = 4.95-5.00) and substantially lower for domestic narratives (range: M = 3.14-3.70), suggesting a strong domain-driven competence stereotype regardless of gender or model.

Table 1. Group Statistics for Warmth by Gender, Domain, and Model

Gender	Domain	Claude	ChatGPT	Gemini	Total
Man	Professional	2.24 (0.79)	2.60 (0.50)	1.75 (0.84)	2.19 (0.80)
	Domestic	4.58 (0.55)	4.93 (0.27)	3.93 (0.62)	4.47 (0.65)
	Total	3.38 (1.36)	3.76 (1.23)	2.84 (1.32)	3.32 (1.35)
Woman	Professional	3.15 (0.74)	3.10 (0.74)	2.26 (0.94)	2.84 (0.91)
	Domestic	4.68 (0.47)	4.95 (0.22)	4.20 (0.69)	4.61 (0.59)
	Total	3.89 (0.99)	4.03 (1.08)	3.24 (1.27)	3.72 (1.17)
Total	Professional	2.70 (0.89)	2.85 (0.68)	2.00 (0.92)	2.52 (0.91)
	Domestic	4.63 (0.51)	4.94 (0.24)	4.06 (0.66)	4.54 (0.62)
	Total	3.64 (1.21)	3.89 (1.16)	3.04 (1.31)	3.52 (1.28)

Note. Values are M (SD). N per cell ≈ 38-40. Scale range: 1-5.

Table 2. Descriptive Statistics for Competence by Gender, Domain, and Model

Gender	Domain	Claude	ChatGPT	Gemini	Total
Man	Professional	4.95 (0.23)	5.00 (0.00)	5.00 (0.00)	4.98 (0.13)
	Domestic	3.14 (0.49)	3.58 (0.55)	3.28 (0.75)	3.34 (0.63)
	Total	4.07 (0.98)	4.29 (0.81)	4.14 (1.02)	4.17 (0.94)
Woman	Professional	4.97 (0.16)	4.98 (0.16)	4.97 (0.16)	4.97 (0.16)
	Domestic	3.32 (0.78)	3.70 (0.46)	3.15 (0.77)	3.39 (0.72)
	Total	4.17 (1.00)	4.34 (0.73)	4.05 (1.07)	4.19 (0.95)
Total	Professional	4.96 (0.19)	4.99 (0.11)	4.99 (0.11)	4.98 (0.14)
	Domestic	3.23 (0.66)	3.64 (0.51)	3.21 (0.76)	3.36 (0.68)
	Total	4.12 (0.99)	4.31 (0.77)	4.09 (1.04)	4.18 (0.94)

Note. Values are M (SD). N per cell ≈ 38-40. Scale range: 1-5

RQ1: How do AI-generated narratives exhibit systematic framing differences across (a) social role contexts and (b) gender?

Results of the 2 (Gender) × 2 (Domain) ANOVAs are presented in Table 3. For warmth, both main and interaction effects were statistically significant. A large main effect for Domain emerged, $F(1, 465) = 860.94$, $p < 0.001$, $\eta^2p = 0.649$, indicating that professional narratives were rated substantially lower in warmth ($M = 2.52$, $SD = 0.91$) than domestic narratives ($M = 4.54$, $SD = 0.62$). A smaller but significant main effect for Gender was also observed, $F(1, 465) = 31.72$, $p < 0.001$, $\eta^2p = 0.064$, with woman-coded narratives rated warmer ($M = 3.72$) than man-coded narratives ($M = 3.32$). The Gender × Domain interaction was also significant, $F(1, 465) = 13.75$, $p < 0.001$, $\eta^2p = 0.029$, indicating that the magnitude of the gender difference in warmth varied by domain. Inspection of cell means revealed that the gender gap in warmth was concentrated in the professional condition (Man: $M = 2.19$; Woman: $M = 2.84$), while domestic warmth ratings were comparably high for both genders (Man: $M = 4.47$; Woman: $M = 4.61$).

Table 3. ANOVA Results for RQ1: Gender × Domain Effects on Warmth and Competence

Source	SS	df	MS	F	p	η^2p	Power
Warmth							
Gender	17.69	1	17.69	31.72	<0.001	0.064	1.000
Domain	480.08	1	480.08	860.94	<0.001	0.649	1.000
Gender × Domain	7.67	1	7.67	13.75	<0.001	0.029	0.959
Error	259.30	465	0.56				
Competence							
Gender	0.07	1	0.07	0.29	0.591	0.001	0.084
Domain	305.47	1	305.47	1283.23	<0.001	0.734	1.000
Gender × Domain	0.13	1	0.13	0.53	0.468	0.001	0.112
Error	110.69	465	0.24				

*Note. Type III sums of squares. η^2p = partial eta-squared. *p <0.025 (Bonferroni-adjusted threshold).*

Taken together, these findings show a clear and systematic warmth pattern: professional contexts evoke low warmth, domestic contexts evoke high warmth, and male-coded professional narratives receive the lowest warmth evaluations overall. For competence, only Domain produced a significant effect. The Domain main effect was quite large, $F(1, 465) = 1283.23$, $p < 0.001$, $\eta^2p = 0.734$, with professional narratives rated near the scale ceiling ($M = 4.98$) and domestic narratives rated markedly lower ($M = 3.36$). Neither the Gender main effect, $F(1, 465) = 0.29$, $p = 0.591$, $\eta^2p = 0.001$, nor the Gender × Domain interaction, $F(1, 465) = 0.53$, $p = 0.468$, $\eta^2p = 0.001$, reached significance, indicating that competence attributions were entirely domain-driven and invariant across gender conditions. In sum, RQ1 is answered affirmatively: AI-generated narratives exhibit systematic stereotyping. Professional and domestic contexts generated distinct warmth and competence framing patterns, with domain accounting for the majority of variance in both outcomes. Gender further shaped these patterns, as professional female characters received higher warmth portrayals than professional male characters despite similarly high levels of competence. Taken together, these results provide clear evidence that all three AI tools, considered collectively, encode stereotype-consistent portrayals in their narrative outputs.

RQ2: How do ChatGPT (GPT-4o), Claude (Opus 4.6), and Gemini (2.5 Pro) differ in the level and pattern of stereotyping in their generated narratives?

To address RQ2, the Model was added as a third between-subjects factor. Full ANOVA results for the 3 × 2 × 2 design are presented in Table 4.

Table 4. ANOVA Results for RQ2: Gender × Domain × Model Effects on Warmth and Competence

Source	SS	df	MS	F	p	η^2p	Power
Warmth							
Gender	17.45	1	17.45	41.35	<0.001	0.083	1.000
Domain	481.17	1	481.17	1140.59	<0.001	0.714	1.000
Model	63.09	2	31.55	74.78	<0.001	0.247	1.000
Gender × Domain	7.63	1	7.63	18.08	<0.001	0.038	0.989
Gender × Model	1.14	2	0.57	1.35	0.261	0.006	0.291
Domain × Model	0.51	2	0.26	0.61	0.547	0.003	0.151
Gender × Domain × Model	1.71	2	0.86	2.03	0.133	0.009	0.418
Error	192.79	457	0.42				
Competence							
Gender	0.09	1	0.09	0.39	0.534	0.001	0.095
Domain	306.58	1	306.58	1394.34	<0.001	0.753	1.000
Model	4.85	2	2.42	11.02	<0.001	0.046	0.991
Gender × Domain	0.14	1	0.14	0.65	0.422	0.001	0.126
Gender × Model	0.67	2	0.34	1.53	0.218	0.007	0.325
Domain × Model	4.31	2	2.16	9.81	<0.001	0.041	0.983
Gender × Domain × Model	0.42	2	0.21	0.96	0.384	0.004	0.217
Error	100.48	457	0.22				

Note. Type III sums of squares. η^2p = partial eta-squared. Bold rows indicate effects of primary interest for RQ2.

For warmth, the Model main effect was significant and of medium-large magnitude, $F(2, 457) = 74.78$, $p < 0.001$, $\eta^2p = 0.247$. ChatGPT produced the warmest narratives overall ($M = 3.89$), followed by Claude ($M = 3.66$), with Gemini generating markedly colder narratives ($M = 3.03$). Bonferroni pairwise comparisons (Table 5) revealed that all three model pairs differed significantly from each other: Claude vs. ChatGPT, $\Delta = -0.25$, $p = 0.002$; Claude vs. Gemini, $\Delta = +0.60$, $p < 0.001$; ChatGPT vs. Gemini, $\Delta = +0.86$, $p < 0.001$. Neither the Domain × Model interaction, $F(2, 457) = 0.61$, $p = 0.547$, $\eta^2p = 0.003$, nor the Gender × Model interaction, $F(2, 457) = 1.35$, $p = 0.261$, $\eta^2p = 0.006$, was significant, indicating that Gemini's lack of warmth was uniform across both domain and gender conditions. This is a systematic, context-independent pattern.

For competence, the Model main effect was significant but small, $F(2, 457) = 11.02$, $p < 0.001$, $\eta^2p = 0.046$. Bonferroni comparisons indicated that ChatGPT ($M = 4.31$) rated narratives significantly higher in competence than both Claude ($M = 4.10$, $\Delta = +0.19$, $p = 0.001$) and Gemini ($M = 4.09$, $\Delta = +0.22$, $p < 0.001$), while Claude and Gemini did not differ from each other ($\Delta = +0.03$, $p = 1.000$). A significant Domain × Model interaction qualified this main effect, $F(2, 457) = 9.81$, $p < 0.001$, $\eta^2p = 0.041$.

Table 5. Bonferroni-Adjusted Pairwise Comparisons for Model Main Effects on Warmth and Competence

DV	(I) Model	(J) Model	Mean Diff.	SE	p	95% CI LL	95% CI UL
Warmth							
Warmth	Claude	ChatGPT	−0.25*	0.07	0.002	−0.43	−0.08
	Claude	Gemini	+0.60*	0.07	<0.001	+0.42	+0.78
	ChatGPT	Gemini	+0.86*	0.07	<0.001	+0.68	+1.03
Competence							
Competence	Claude	ChatGPT	−0.19*	0.05	0.001	−0.32	−0.06
	Claude	Gemini	+0.03	0.05	1.000	−0.10	+0.15
	ChatGPT	Gemini	+0.22*	0.05	<0.001	+0.09	+0.34

*Note. Based on observed means. Error term = Mean Square(Error). *p <0.05.*

Simple effects analysis decomposing the Domain × Model interaction (Table 6) revealed a clear pattern in that the models did not differ in competence ratings within the professional domain, $F(2, 457) = 0.08$, $p = 0.921$, $\eta^2p < 0.001$, with all three producing near-ceiling ratings ($Ms = 4.96$-4.99). Within the domestic domain, however, a significant model effect emerged, $F(2, 457) = 20.69$, $p < 0.001$, $\eta^2p = 0.083$. ChatGPT assigned substantially higher competence to domestic narratives ($M = 3.64$) than both Claude ($M = 3.23$, $\Delta = +0.41$, $p < 0.001$) and Gemini ($M = 3.21$, $\Delta = +0.43$, $p < 0.001$), whereas Claude and Gemini were statistically indistinguishable ($\Delta = +0.02$, $p = 1.000$). No significant Gender × Model or Gender × Domain × Model interactions emerged for competence (both $ps > 0.218$), indicating that ChatGPT's distinctive domestic competence pattern was not moderated by gender.

Collectively, the RQ2 findings indicate that while all three models displayed similarity in their basic domain-based stereotyping pattern, they differ meaningfully in terms of degree. Gemini generates systematically colder narratives across all conditions, regardless of domain or gender. ChatGPT is notably more lenient in its competence ratings for domestic characters, though this difference disappears in professional contexts where all three models converge at the ceiling rating. Claude falls between the other two on the warmth dimension and mirrors the Gemini model in domestic competence ratings.

Table 6. Simple Effects of Model Within Each Domain Level for Competence

Domain	(I) Model	(J) Model	Mean Diff.	SE	p	95% CI	η^2p
Professional	—	All pairs	ns	—	0.921	—	0.000
Domestic	ChatGPT	Claude	+0.41*	0.08	<0.001	[+0.22, +0.59]	0.083
	ChatGPT	Gemini	+0.43*	0.07	<0.001	[+0.25, +0.60]	
	Claude	Gemini	+0.02	0.08	1.000	[−0.16, +0.20]	

*Note. Bonferroni adjustment applied to pairwise comparisons within each domain. η^2p reported for the omnibus simple effect F test. *p <0.001.*

Discussion

This study examined whether and how AI-generated narratives exhibit systematic stereotype-consistent framing across social role contexts and gender, and whether that pattern differs across three leading commercial AI tools. The results provide evidence that AI-generated narratives do exhibit systematic stereotype-consistent framing, with the strongest differences occurring across social role contexts and more nuanced differences across gender and AI tool. The most pronounced pattern was associated with social role context. Professional narratives received substantially lower warmth ratings than domestic narratives, whereas domestic narratives received substantially lower competence ratings than professional narratives. These differences were consistent across the three AI tools and were evident in both the descriptive statistics and inferential analyses. When prompted to write a narrative about a person in a professional context, the AI tools emphasized capability, achievement, and independence while placing comparatively less emphasis on warmth. In contrast, narratives describing the same individual in a domestic context emphasized warmth, care, and interpersonal orientation while assigning comparatively lower competence. Because the character's identity remained constant across conditions, these findings suggest that social role context systematically shaped how the AI tools framed the same individual. Gender adds another layer to this pattern. Overall, woman-coded narratives received higher warmth ratings than man-coded narratives, but this difference was not evenly distributed across domain contexts. The gender gap in warmth was concentrated in the professional domain, while domestic narratives were relatively warm for both genders. Both genders received high competence ratings in professional contexts, but professional men received substantially lower warmth portrayals than professional women. This points to a specific asymmetry in how the AI constructs gender in professional narratives. Professional men are described as highly competent yet quite cold, while professional females received similarly high competence ratings alongside notably higher warmth portrayals. The inter-model findings add a further layer of nuance to the study. For warmth, the stereotype-consistent pattern was structurally similar across all three AI tools. The warmth gap between professional and domestic narratives remained statistically equivalent regardless of which AI tool generated the text, and the Gender × AI Tool interaction was also not significant, indicating that the gender-related warmth pattern remained stable across systems. The AI tools differed primarily in their overall warmth levels, with Gemini generating the lowest warmth portrayals and ChatGPT the highest. For competence, all three AI tools converged at the rating ceiling in the professional domain. In domestic contexts, however, the tools diverged. ChatGPT assigned relatively higher competence portrayals to domestic characters, whereas Claude and Gemini generated substantially lower and nearly identical competence portrayals. Importantly, this difference was not moderated by gender, suggesting that ChatGPT's relatively higher competence portrayals in domestic contexts applied similarly to both male and female characters. Taken together, these findings suggest that the differences across AI tools are primarily differences in degree

rather than kind. All three AI tools generated stereotype-consistent portrayals, but the intensity of those portrayals varied across systems. The reason ChatGPT differed in domestic competence portrayals cannot be determined from the present data. One possible explanation is that differences in alignment strategies or post-training adjustments may have influenced how domestic competence is represented. Another possibility is that certain observed differences, particularly the higher warmth portrayals of professional women, may reflect intentional alignment choices designed to reduce negative portrayals of historically marginalized groups, rather than simple perpetuation of social stereotypes. If so, the resulting pattern could reflect overcorrection rather than stereotype-consistent bias in the traditional sense. These interpretations are not mutually exclusive and remain open empirical questions; future work with access to model documentation or alignment disclosures would be better positioned to adjudicate between them. Importantly, the inter-model warmth and competence differences reported here should be understood as observations under the specific prompts, model versions, and access conditions of this study, not as stable or durable traits of these AI systems, which are updated continuously. Even so, the overall competence gap between professional and domestic contexts remained substantial across all three AI tools.
The present findings carry several implications for theory. Most directly, the current study demonstrates that bias in AI-generated narratives follows systematic stereotype-consistent patterns across social role contexts and gender. This extends prior research on AI bias and stereotyping, which has largely examined bias at the level of individual word(s), short text responses, or tightly controlled sentence prompts (Blodgett et al., 2020; Bolukbasi et al., 2016; Nadeem et al., 2021; Sheng et al., 2019). While those approaches reveal pieces of the puzzle, they do not fully capture how broader patterns of stereotyping emerge across extended narratives. The present findings suggest that stereotyping can operate at the narrative level, making it more difficult to detect than isolated word-level associations. A narrative may contain no explicitly biased language while still constructing a stereotyped portrayal through the selective emphasis or de-emphasis of particular traits and behaviors. Overall, these findings suggest that understanding AI bias requires moving beyond isolated linguistic elements to examine the broader narrative structures through which stereotype-consistent portrayals are constructed and reinforced.
The Gender × Domain interaction adds an additional theoretical insight. Female characters in professional contexts received notably warmer portrayals than male characters in the same setting, while domestic portrayals remained relatively warm for both genders. Professional contexts did not reduce warmth uniformly across genders; instead, professional males received substantially lower warmth portrayals than professional females despite similarly high competence portrayals. This pattern reflects a portrayal of professional women that combines warmth and competence in ways not equally applied to professional men. Consistent with role incongruity theory (Eagly & Karau, 2002), these findings suggest that AI-generated narratives encode gendered expectations regarding which traits are perceived as fitting particular social roles. More broadly, the findings indicate that AI-generated narratives capture not only warmth and competence distinctions associated with the SCM (Fiske et al., 2002), but also the gendered expectations attached to those distinctions.
The findings carry practical implications for several audiences. For AI developers, the results highlight the importance of evaluating narrative-level stereotyping in AI-generated outputs. Standard content evaluation pipelines tend to focus on harmful language, factual accuracy, and instruction-following (Ouyang et al., 2022). As shown in this study, however, stereotype-consistent framing is unlikely to be identified through these evaluation procedures alone. A professional character described as driven and accomplished, and a domestic character described as warm and relationship-oriented, may both pass every standard word-level content check while still producing a stereotypical warmth-competence pattern with large effect sizes.
As AI-generated content becomes more widespread, these representational patterns are disseminated at a scale and consistency that few traditional media systems can match, often without readers knowing the content was generated algorithmically (Illia et al., 2025). Evaluating warmth and competence framing

across a variety of social role contexts and demographic groups, using a controlled prompt design like the one employed here, could be a practical and valuable addition to model testing procedures.
For organizations and platforms that deploy AI to generate content about individuals, the findings point to two decisions that carry important representational consequences. The first key decision is AI tool choice. The results show that ChatGPT, Claude, and Gemini do not produce equivalent content from a stereotyping standpoint, and that these differences are systematic rather than random. Decisions made on the basis of cost, capability, or accessibility may therefore also influence the representational patterns embedded in generated content. The second key decision is content review. A narrative that is factually correct can still be stereotypically structured, and reviewers focused primarily on factual correctness may overlook these broader framing patterns. This issue applies across contexts in which AI is used to generate narratives about individuals, from news summaries and educational materials to employee profiles, customer communications, and human-interest content.
Finally, the findings of this study strengthen calls for increased transparency regarding AI-generated content. Stereotype-consistent framing in AI-generated narratives is rarely explicit, leaving readers unaware that portrayals may reflect broader representational patterns. While increased transparency about the provenance of AI-generated content would not eliminate such patterns, it could encourage readers to engage more critically with the content, rather than interpreting it as a neutral or transparent account of a person.

Limitations

The present study has limitations that should be noted. Although AI-based ratings were validated against human judgments prior to full-corpus coding, it is possible that the rating model's own associations with professional and domestic contexts contributed to the pattern of scores. While identifying information was removed prior to scoring and scoring was conducted via API with no memory or conversational context, narrative content itself may carry implicit domain cues that the scoring model recognizes. This limitation cannot be fully resolved without human-only rating of the full corpus, and future work using human raters or alternative automated methods would strengthen confidence in these findings.
The present study examined a specific set of social role contexts in a single national cultural context. Findings may differ across other role contexts, cultural settings, names, or demographic groups beyond the binary gender groups examined here. The American focus of the current corpus limits the generalizability of the findings, and future research should examine whether the same patterns hold across different national and cultural contexts where the meaning of professional and domestic roles, and the gender associations attached to them, may be constructed differently.
Finally, this study examined narrative content, not audience effects. Demonstrating that AI-generated narratives are stereotyped is a necessary but not sufficient condition for concluding that they have downstream effects on how audiences perceive various social groups. Future work should examine whether exposure to stereotyped AI-generated narratives shapes social perceptions, a question that the present findings make both theoretically motivated and empirically tractable.

Conclusion

When AI tells a story about a person, it does not tell it neutrally (i.e., with equal warmth and competence portrayals regardless of social role context, gender, or AI tool). Narratives generated by three leading commercial AI tools systematically reflected stereotypes along warmth and competence dimensions, with large and consistent effect sizes that varied by both social role context and character gender. Gender notably shaped the warmth pattern, with female characters in professional contexts receiving a

substantially warmer portrayal than their male counterparts, suggesting that the AI tools do not apply stereotypic framing in a gender-neutral way. The AI tools differed in magnitude but not in form, as all three consistently generated similar stereotypic narratives, with variation occurring primarily in the strength of these patterns. As AI-generated content becomes an increasingly dominant part of the informational environment, the systematic framing patterns embedded in these systems may quietly shape how audiences understand who is capable, who is caring, and who belongs. This is not a trivial concern. It is a call for systematic scrutiny of AI-generated outputs, greater transparency in disclosing the algorithmic origins of content, and enhanced accountability for the representational choices, deliberate or not, embedded by developers in the design and refinement of these AI systems.

Acknowledgements

The authors used Claude Opus 4.6 (Anthropic), ChatGPT GPT-4o (OpenAI), and Gemini 2.5 Pro (Google) to generate the narrative corpus examined in this study, as described in the Data Collection section. Claude Opus 4.6 was additionally used as an automated content analyst to score narrative warmth and competence ratings, as described in the rating procedure section. Claude (Anthropic) was also used to assist with grammar checking portions of this manuscript. All content was reviewed and approved by the authors, who take full responsibility for the accuracy, originality, and integrity of the work.

References

Bansal, C., Pandey, K. K., Goel, R., Sharma, A., & Jangirala, S. (2023). Artificial intelligence (AI) bias impacts: Classification framework for effective mitigation. *Issues in Information Systems*, 24(4), 367-389.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610-623.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454-5476.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*, 4349-4357.

Cheryan, S., Plaut, V. C., Handron, C., & Hudson, L. (2013). The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women. *Sex Roles*, *69*(1-2), 58-71.

Chiang, C. H., & Lee, H. Y. (2023). Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 15607-15631). Association for Computational Linguistics.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61-149.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573-598.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878-902.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2), 67-73.

Gerbner, G., Gross, L., Morgan, M., & Signorielli, N. (1986). Living with television: The dynamics of the cultivation process. In J. Bryant & D. Zillmann (Eds.), *Perspectives on media effects* (pp. 17-40). Lawrence Erlbaum Associates.

Henriksson, T. 2023. *New Survey Finds Half of Newsrooms use Generative AI Tools; Only 20% Have Guidelines in Place*. WAN-IFRA. https://wan-ifra.org/2023/05/new-genai-survey/.

Illia, A., Lawson-Body, A., & Rouibah, K. (2025). Determinants of trust in AI-generated content. *Issues in Information Systems*, 26(4), 56-64.

Kirk, H. R., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., Shtedritski, A., & Asano, Y. M. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems*, *34*, 2611-2624.

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*(3), 371-392.

Kotenidis, E., & Veglis, A. (2021). Algorithmic journalism: Current applications and future perspectives. *Journalism and Media*, *2*(2), 244-257.

Mastro, D. E., & Greenberg, B. S. (2000). The portrayal of racial minorities on prime time television. *Journal of Broadcasting & Electronic Media*, *44*(4), 690-703.

McArthur, L. Z., & Resko, B. G. (1975). The portrayal of men and women in American television commercials. *The Journal of Social Psychology*, 97(2), 209-220.

Mousavi, R., Kitchens, B., Oliver, A. G., & Abbasi, A. (2026). From lexicons to large language models: A holistic evaluation of psychometric text analysis in social science research. *Information Systems Research*, 1-20.

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 5356-5371.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Sargent, C. S., Koohang, A., Floyd, K., & Kilburn, R. (2024). Artificial Intelligence: Ethical concerns, trust, and risk. *Issues in Information Systems*, 25(2), 71-83.

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3407-3412.

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730-3748.