Neutrality Bites: Gender Representation in AI-Generated Animal Stories

IMANI FINKLEY, University of Washington, USA
YUANXI LI, University of Washington, USA
MELANIE WALSH, University of Washington, USA

Gender bias in AI-generated stories is a well-documented problem. While much attention has been paid to reducing or mitigating this bias, it is not always clear whether interventions produce genuinely fairer results. To investigate this issue, we examine how large language models (LLMs) handle gender assignment in a narrative context that is popular, highly ambiguous, and also known to closely reproduce human stereotypes: stories about talking animals. We prompt six leading LLMs to complete an English-language story about seven different anthropomorphic animal characters whose gender is unstated. We additionally iterate with four different narrative settings and a range of model temperatures. Across the 23.8K stories, we find that models frequently avoid gendering the animal character in the story (19% on average) or use gender-neutral language like "it" or "its" (38.2% on average). However, when gender is assigned, there is a significant masculine bias. Feminine animal characters are virtually absent, present in just 2.2% of stories vs. 40.6% that feature masculine characters. Our findings point to a broader argument: neutrality bites. In other words, models that prioritize neutrality to address social bias may actually contribute to the erasure of marginalized perspectives and identities. We suggest that alternative strategies beyond neutrality need to be pursued, such as ones that more equally distribute social possibilities across imagined subjects.

CCS Concepts: • **Social and professional topics → Gender**; • **Information systems → Content analysis and feature selection**; • **Applied computing → Arts and humanities**; • **Computing methodologies → Natural language generation**.

Additional Key Words and Phrases: fiction, gender neutrality, gender bias, large language models

ACM Reference Format:
Imani Finkley, Yuanxi Li, and Melanie Walsh. 2026. Neutrality Bites: Gender Representation in AI-Generated Animal Stories. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3805689.3812287

1 Introduction

Gender bias in AI-generated outputs has been a widespread problem [5, 27, 41]. These systems frequently overrepresent cisgender men and masculine perspectives while rendering women, non-binary people, and femininity either less visible, or reductively and offensively portrayed [7, 24]. This is concerning not only for algorithmic decision-making contexts like hiring, but also for storytelling and fiction, where AI models are increasingly used. For example, Google's Gemini "Storybook" tool [17], which creates 10-page illustrated narratives based on prompts, is now being used by parents to read stories to their children [35].

Authors' Contact Information: Imani Finkley, ifinkley@uw.edu, University of Washington, Seattle, Washington, USA; Yuanxi Li, yuanxili@uw.edu, University of Washington, Seattle, Washington, USA; Melanie Walsh, melwalsh@uw.edu, University of Washington, Seattle, Washington, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812287

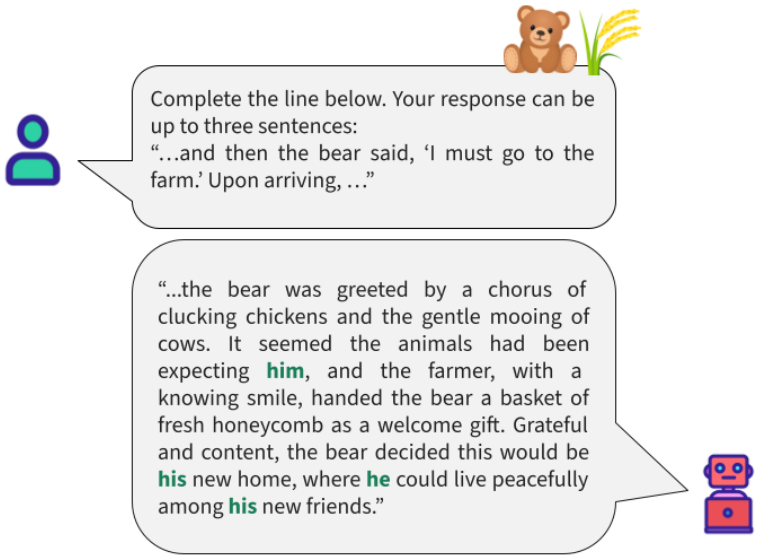

Fig. 1. Example prompt and response from GPT-4o with the following parameters: bear (subject), farm (setting), and 1.0 (temperature). We label this story Masculine due to the pronouns (bold green) used to refer to the bear. Generated December 2025.

Reducing or mitigating gender bias in large language models (LLMs) is now an active area of research in both academia and industry [15, 46, 48]. But it is not always clear whether proposed or adopted interventions produce genuinely fairer representations, especially for fiction. For instance, if a user prompts an LLM to generate a story about a character whose gender is ambiguous, and the model avoids gendering the subject, is that real fairness? Or is it a band-aid fix that covers up a larger bias problem, leaving it to seep out elsewhere?

To investigate this issue, we examine how LLMs handle gender assignment in a narrative context that is popular, highly ambiguous, and also known to closely reproduce human stereotypes: stories about talking animals (Figure 1). We turn to anthropomorphic fiction for several reasons.

First, animal stories are popular and influential. Human authors have been writing stories about anthropomorphic animals for thousands of years, back to *Aesop's Fables* (620-564 BCE) and beyond. Today, they feature prominently in children's books and media—like *Pete the Cat* (2008-present) or *Bluey* (2018-present)—which are important for human development and shape early conceptions of gender and gender roles. Additionally, in the last several years, users have turned to AI models to create children's books [19, 23, 25], such as *Alice and Sparkle*, one of the first known children's books to include AI-generated text and images [38].

Second, anthropomorphic narratives represent a unique projection and reflection of human culture. Animal characters that resemble humans offer useful imaginative affordances for creators. Some authors reportedly turn to animal characters to conjure "universal" subjects who "transcend" gender, race, or other identity categories and physical characteristics [28, 47]. Yet, counterintuitively, research shows that gender bias is actually more pronounced in stories about animal characters than in stories about human characters [10, 28]. In other words, paradoxically, human writers project human stereotypes more strongly in animal stories, making them a striking test case for LLMs.

Third, we hypothesize that model developers have not created guardrails or aligned models in response to this specific context, as they almost certainly have for stories about humans in different occupations, like doctors or flight attendants, a genre that has received more scholarly and popular attention [13, 24, 27, 43].

In our experiment, we iteratively prompt six leading LLMs to complete an English-language story about seven different human-like animal characters whose gender is unstated—bear, dog, rabbit, cat, mouse, pig, and bird. We additionally iterate with four different narrative settings and a range of model temperatures, which are known to influence the randomness or creativity of the output. For each story, we identify the pronouns used to describe the animal character, which we take as a proxy for gender assignment.

Across the 23.8K LLM-generated stories, we seek to answer the following questions:

- **RQ1:** How often do LLMs assign **masculine** (he/him), **feminine** (she/her) **neutral** (it/its, they/them), or **no pronouns** (animal name) to the fictional animal character, and how does this compare across models?
- **RQ2:** How do different factors—model, model temperature, animal character, setting—impact gender assignment, if at all?
- **RQ3:** How do LLMs' pronoun assignments compare to human responses from prior work?

Overall, we find that models frequently avoid gendering the animal character in the story (**19%** on average) or use gender-neutral language like "it" or "its" (**38.2%** on average). However, when gender *is* assigned, there is a significant masculine bias. Feminine animal characters are virtually absent, present in just **2.2%** of stories vs. **40.6%** featuring masculine characters. They mostly appear in stories about cats (**53.2%**), which are stereotypically associated with femininity. Even the model with the highest feminine representation, Claude Sonnet 4.5, only features feminine characters in **3.8%** of the stories vs. **34.3%** masculine stories. Compared to a human study that used an almost identical story prompt [45], LLMs were **six times *less* likely** to imagine a feminine character and **1.2 times more likely** to use neutral or animal name references.

Notably, gender assignment is not consistent across LLMs, including those from the same developer. Surprisingly, the most recent OpenAI model, GPT-5.1, displayed the strongest masculine bias, with **65.2%** of the stories featuring masculine characters. Though differences in pretraining data or architecture cannot be ruled out, these observed disparities suggest that post-training interventions likely play a role in shaping gendered behavior in these models. Model temperature and narrative setting did not have major impacts on gender assignment.

Our findings point to an overarching claim: ***neutrality bites***. We argue, in other words, that models that prioritize neutrality in order to address social bias may unintentionally contribute to the erasure of marginalized perspectives and identities. We suggest that alternative strategies beyond neutrality need to be pursued, such as ones that more equally distribute gender possibilities across imagined subjects. Code and materials to support this work can be found at https://github.com/imanif/animal-stories.

2 Related Work

2.1 Gender Bias in Children's Literature

Anthropomorphic animals, like a bear who lives in a house or a tiger that talks, feature prominently in children's literature, a widely consumed genre that plays a formative role in establishing social norms and frameworks for kids [6]. A large body of research has pointed to significant gender bias in English-language children's literature, such as the overrepresentation of masculine characters, especially as protagonists [26] or in active storylines [2], and the reinforcement of a gender binary [9, 18, 20, 31]. While McCabe et al. [28] found that the representation of feminine human characters steadily improved throughout the 20th century, they found that representation of feminine animal characters did not, remaining stagnant. Walsh et al. [45] found that gender bias persisted among a selection of the most popular and widely-read English-language children's books featuring anthropomorphic animals, even into the 21st century. This concern also extends beyond books themselves to the

broader information environments; for example, children's products are often retrieved and suggested through stereotypically gendered associations [37].

2.2 Gender Bias in LLM-Generated Fiction

Gender bias is a well-demonstrated and widely discussed problem in LLMs [5, 22, 41]. Notably, narrative texts—especially short stories—are frequently used as test cases for audits of gender and social bias, even in research not primarily concerned with fiction or narrative form [16, 27, 42]. As with human-authored literature, stories serve as a useful window into broader cultural imaginaries: they make visible assumptions, norms, and stereotypes, and they can expose implicit biases that may remain less visible in more constrained or task-oriented settings, like the review of job applications.

Many studies that explicitly investigate LLM-generated stories or use them as test cases have found significant gender bias. For example, Lucy and Bamman [27] found thematic biases in GPT-3-generated stories, showing that feminine characters were primarily described by their appearance while masculine characters were associated with high power. Spillner [43] similarly found that, in ChatGPT stories, women were overrepresented in stereotypically women-dominated fields—though also found, in a reversal, that women were overrepresented in some stereotypically *male*-dominated fields, perhaps reflecting alignment efforts to avoid bias. A study that used LLMs to produce children's bedtime stories also found stereotypical gender representations, as well as limited diversity in the socioeconomic backgrounds and ethnicity of characters in narratives given different geographical settings [39].

Perhaps most closely related to our own work, Gillespie [16] found that leading models from Microsoft and OpenAI often "avoided" gendering the subject in an ambiguous story, speculating that "this is almost certainly the result of interventions meant to address the uneasy cultural politics around pronouns and gender identity."

2.3 Gender Debiasing & Finetuning

Attempting to "debias" LLMs has been a growing trend in response to this evidence, and it has taken a variety of forms, in recognition of the array of harms to various gender minority groups [8, 22]. Proposed interventions include finetuning models to correctly employ neopronouns—such as, in English, xe/xem/xyr or ze/zir/zirs, or, in Dutch, hen/hen/hun or die/die/diens—to reflect a category of pronoun used by some non-binary people [44]. There is also work on so-called "gendertuning," which uses bias word perturbation and fine-tuning classification to revise texts with non-gendered pronouns or pronouns that disrupt common stereotypes (e.g. "women in the kitchen" becomes "men in the kitchen") [15]. Adopting generated-language reforms remains a challenge. For example, when tasked with explaining their text revisions, LLMs have been found to treat gender-inclusivity as more relevant for women or trans people, rather than as a default approach [46].

Recent work further suggests that apparent gender neutrality does not necessarily translate into equitable representation. Mickel et al. [29] show that increased representation of women in model outputs does not eliminate differences in how women and men are described, while Abolghasemi et al. [1] show that seemingly balanced results can still encode biased exposure patterns. Together, these studies suggest that fairness must be evaluated not only in terms of who appears in outputs, but also in terms of how gendered subjects are described and distributed. A broader perspective from information retrieval is also useful here. Seyedsalehi [40] show that gender bias can emerge at multiple stages of an information system, including queries, datasets, ranking, and evaluation, which helps frame gender neutrality not simply as a matter of wording, but as a broader systems problem.

3 Methods

3.1 Story Generation

To create our animal story dataset, we prompt six diverse, state-of-the-art LLMs to complete a specific English-language story template. We test two generations of OpenAI models—GPT-4o [33] and GPT-5.1 [34]—and leading models from Anthropic and Google—Claude Sonnet 4.5 [3] and Gemini 2.5 Flash [11]—as well as two widely used open or hybrid models—Mistral Medium 3.1 [30] and OLMo3 7B Instruct [32]. We prompt each model with an in-media-res story, introducing a talking animal character who speaks in the first person and is not clearly gendered in any way:

> *Complete the line below. Your response can be up to three sentences:*
> *"…and then the **{animal}** said, 'I must go to the **{setting}**.' Upon arriving, …"*

We adopt this template from a previous study conducted by the senior author with journalists from *The Pudding*, in which they gave the same prompt to human survey participants [45]. This allows us to compare model behavior to human responses from prior work. We systematically vary the story with seven different animals and four different settings, in order to account for biases that may be introduced with any particular animal or location. The animal characters were selected based on the same prior work [45], which included a survey of 300 popular English-language children's books on *Goodreads* that feature at least one anthropomorphized animal, focusing on the most rated titles published since 1950, plus some older and still highly rated classics. We also noted settings commonly used in these narratives and selected four settings with potentially embedded cultural stereotypes (e.g., domestic settings like a kitchen may be more associated with feminine characters, while outdoor spaces like a farm may be more associated with masculine characters) for our experiment design. Additionally, we vary the temperature for each model, experimenting with the model's default temperature as well as one temperature below and above it.[1] We prompt each model 50 times with each animal, setting, and temperature combination, totaling 23.8K stories and approximately 1.05M tokens (Table 2).

Story Variable	
Animals (7)	bear, bird, cat, dog, mouse, pig, rabbit
Settings (4)	farm, kitchen, river, store

Table 1. Experimental design: Animals and settings used for story generation prompts.

3.2 Story Analysis

To determine the gender of the animal in the story, we rely on pronouns in the text. In the English language, pronouns often imply the gender of a subject. We establish the following categories: **Feminine** (she, her, hers, herself); **Masculine** (he, him, his, himself); **Neutral** (it, its, itself; they, them, their, theirs, themself, themselves). Cases in which no pronouns are used, and the character is simply referred to by the animal's name (e.g., "the pig") are labeled **Animal Name**. We recognize that this is not an exhaustive classification of gender but present these labels in an effort towards examining gender representation in generated literature and NLP that extends beyond a binary [8, 12].

In some of our analyses, we combine **Neutral** and **Animal Name** into an overarching gender-neutral assignment category to highlight the impact of not assigning a character a specific gender. However, we do not treat these outcomes as conceptually identical. Especially in human contexts, **neutral** pronouns like "they/them" are

[1]The default temperature for Claude Sonnet 4.5 is also the maximum temperature—1.0—making it impossible to test at a temperature above the default. Thus, for the Claude model, we only prompt at the default temperature and one temperature below it.

Model	Stories	Temperatures
Claude Sonnet 4.5	2,800	0.75, **1.0**
Gemini-2.5	4,200	0.75, **1.0**, 1.25
GPT-4o	4,200	0.75, **1.0**, 1.25
GPT-5.1	4,200	0.75, **1.0**, 1.25
Mistral Medium	4,200	0.5, **0.7**, 0.9
OLMo3	4,200	0.75, **1.0**, 1.25
Total	**23,800**	

Table 2. Stories generated per model. 1.4K stories generated per temperature. Bold indicates default temperature for each model.

often an intentional linguistic choice aligned with inclusivity or with recognition of non-binary gender expressions. We acknowledge that **neutral** language, and even gender-ambiguous language, is sometimes desirable and carries positive, socially progressive connotations. Two considerations shape how we handle this complexity. First, the models rarely used "they/them" pronouns for animal characters and instead used "it/its" in a manner that we view as similar to the use of the animal's name. Consider, for example, a model completion where a pig is referred to as "it" versus "the pig":

> ..and then the pig said, 'I must go to the river.' Upon arriving, **it** gazed at the rippling waters and sighed contentedly (OLMo3, *animal* = pig, *temperature* = 1.25, *setting* = river)

> ...and then the pig said, 'I must go to the river.' Upon arriving, **the pig** saw a group of frogs gathered on lily pads, waiting for rain to arrive. (OLMo3, *animal* = pig, *temperature* = 1.25, *setting* = river)

We read these two stories as functionally similar in their representation of the pig's gender, or lack thereof, though we note that the use of "it" is perhaps more intentionally non-gendered and in this context may objectify or even de-humanize the pig more. This leads to our second point. Because we believe this difference is small but still meaningful to preserve, we adopt a dual approach. We report results with each specific category noted, and we also present our analysis with a generalized neutrality label to investigate broader patterns in model behavior.

To parse which pronouns refer to the animal character specified in the prompt, we use coreference resolution. Coreference resolution disambiguates references to subjects using contextual evidence. This method is necessary to analyze the stories because ambiguity can be introduced with multiple characters or objects—e.g. "The **bear** and the **farmer** were in the store and saw a **jar of honey**. The **farmer** asked **her** if **she** wanted **it**." We need to identify the pronouns related to the main animal character—e.g. the bear ("she/her")—not other characters or objects—e.g. the jar of honey ("it"). To identify these references, we specifically use the FastCoref package [36]. We manually annotated 200 randomly sampled stories and measured how our labels compared to FastCoref. Across each gender category, FastCoref achieved a mean F1 score of 0.95 (see more details in Appendix B). We then manually reviewed stories labeled Animal Name or Neutral to minimize false positives given the ambiguous nature of this language.

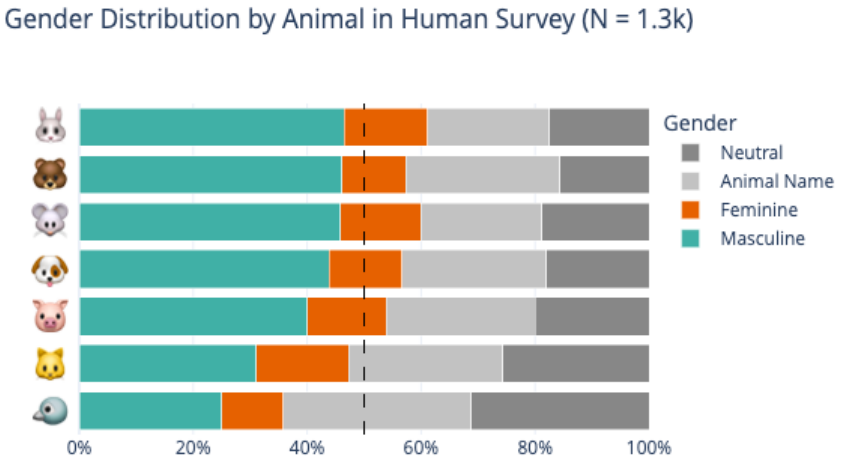

Fig. 2. Gender distribution results from 1327 human survey respondents given a story-completion task with varying anthropomorphized-animal protagonists.

3.3 Establishing a Baseline

To establish a human baseline for the LLM-generated stories, we draw on prior work that surveyed approximately 1.3K respondents (see details in Appendix C). Respondents were asked to complete a nearly identical story to the one provided to the LLMs [45]:[2]

> *In the 1980s, there was a study that produced surprising results about our predictive tendencies. We're conducting a similar study today.*
> *Instructions: Complete the line below. What did the **{animal}** do? Your response can be up to three sentences.*
> *"...and then the **{animal}** said, 'I must go to the river.' Upon arriving,..."*

The initial part of the prompt—misdirecting the respondent about the focus of the study—was included to avoid priming participants to explicitly consider gender. Survey respondents were randomly given one of seven animals: bear (204 stories), bird (176 stories), cat (203 stories), dog (189 stories), mouse (192 stories), pig (180 stories), or rabbit (183 stories). Of the 1,327 completed stories, **39.9 percent** assigned masculine pronouns to animal protagonists, while **13.4 percent** of the stories employed feminine pronouns. Cat had the most feminine assignments (**16.3 %**), while bird had the most neutral representation (**64.3%**) (see Figure 2).

To compare the model-generated narratives with our human baseline, we restrict our analysis to a subset of the LLM-generated stories where the *setting* is river (n = 5950). Since the dependent variable, *Gender* {Masculine, Feminine, Neutral, Animal Name}, is a nominal categorical variable with four levels, we use a mixed multinomial logistic regression model, which allows us to estimate differences across gender categories and accounts for multiple outputs from each model.

[2]The survey was circulated to readers of *The Pudding*, a data journalism outlet, on social media and via email, offering an opportunity for respondents to win free merchandise.

We treat this as a 2x7 mixed factorial design with the following fixed factors: *Author* {LLM, Human} and *Animal* {bear, bird, cat, dog, mouse, pig, rabbit}, allowing us to examine both their effects and interactions on gender assignment. We then used an analysis of variance, implemented with the multinomial Poisson transformation [4], to identify any statistically significant differences between the survey respondents and the LLMs (averaged across temperature and model variations to reduce within-model variability).

4 Results

4.1 Overall

Across the 23.8K LLM-generated stories, we find, overall, that the six models predominantly either avoided gendering the anthropomorphic animal character or gendered the character with neutral pronouns like "it" (**57.2% total**). The language used for gender-neutral representation did not vary greatly: only **two** stories used "they/them/theirs" to refer to a single animal character, once for a human side character. However, when the animal protagonist was gendered, it was gendered masculine **95 percent** of the time (Table 3). Only **2.2%** of the outputs featured a feminine animal character—that is, just **513** of **23,800 stories** (compared to **9,673** masculine stories).

Model	Neutral or Animal Name %	Masculine %	Feminine %
OLMo3	**85.3**	11.5	3.2
Mistral Medium	**71.9**	27.9	0.2
Claude Sonnet 4.5	**62.0**	34.3	3.8
GPT-4o	**57.0**	40.2	2.8
Gemini-2.5	35.6	**62.7**	1.7
GPT-5.1	33.0	**65.2**	1.8
Human Survey Baseline	**46.7**	**39.8**	**13.4**
Model Average	**57.2**	**40.6**	**2.2**

Table 3. Gendered Pronoun Breakdown in Animal Stories Across LLMs. Human survey baseline drawn from a survey by *The Pudding*.

The most "neutral" model, OLMo3, also had the second-best feminine representation. OLMo repeated the animal's name, rather than assigned a gender via pronouns, in stories **46.6 percent** of the time, while using neutral pronouns **38.6 percent** of the time. Only **3.2 percent**—or **135** of **4,200 stories**— contained feminine main characters, with **11.5 percent** containing masculine main characters.

Mistral rarely cast feminine characters. Only **9** stories (**0.2 %**) contained feminine protagonists. Moreover, all of these feminine characters were cats. Conversely, Mistral used gender-neutral pronouns for its characters for **52.2 percent** of its stories, avoided gender **19.7 percent** of the time, and employed masculine pronouns for **27.9 percent** of the time.

Of all the models, GPT-5.1 and Gemini-2.5 displayed the strongest masculine bias. The animal protagonist was cast as masculine in **65.2 percent** of stories generated by GPT-5.1, far outstripping the masculine proportion of similar stories completed by humans (**39.8 %**). Outputs were especially skewed depending on the animal. Stories were more than **77 percent** masculine for dogs, bears, pigs, rabbits, and mice. But cats, stereotypically associated with feminine characters, were flipped. Cats were not gendered or were described neutrally **78.3%** of the time, yet when explicitly gendered, were still predominantly masculine (**14.7%** vs. **7%**). By contrast, birds were **99.3%** neutral, referenced as "it" or "its."

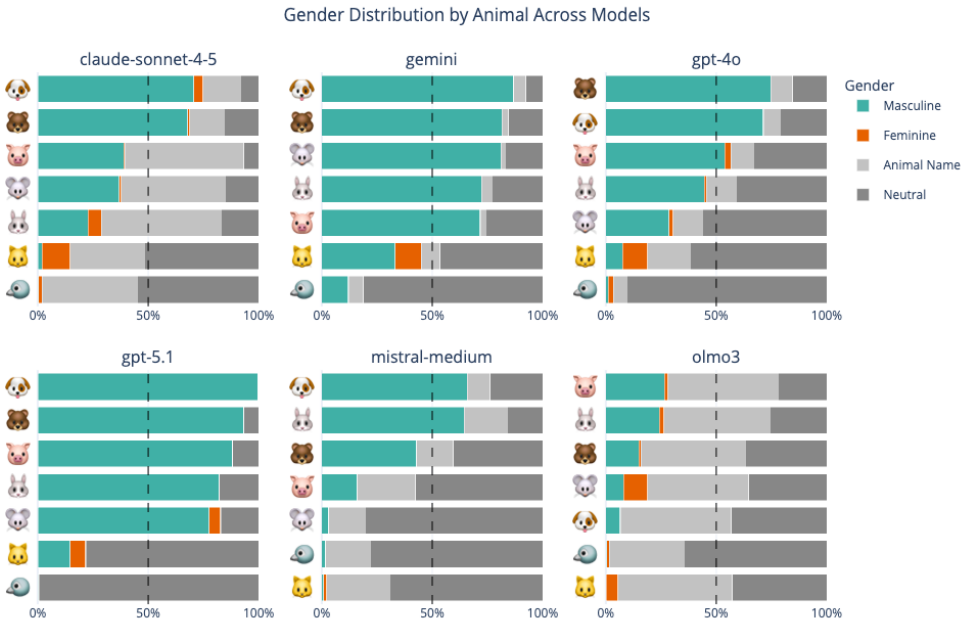

Fig. 3. Gender assignment distribution results from 23.8K text outputs from six LLMs given a story-completion task with varying anthropomorphized-animal protagonists, narrative settings, and temperature parameters.

Despite having a relatively low proportion of feminine characters (3.8%), Claude was the only model to produce feminine characters across all seven animal categories, including those typically labeled masculine by human participants.

4.2 Compared To The Baseline

We find a statistically significant difference in how LLMs and the survey respondents resolve gender in ambiguous contexts ($\chi^2(3, N{=}7277) = 332.63$, $p < 2.2e^{-16}$). The animal protagonist also had a significant effect on gender assignment distribution in the completed stories ($\chi^2(18, N{=}7277) = 359.45$, $p < 2.2e^{-16}$). Additionally, multinomial logistic regression revealed a significant interaction effect between story author (LLM vs. human) and the talking animal characters ($\chi^2(18, N{=}7277) = 218.30$, $p < 2.2e^{-16}$), suggesting that there were distinct differences between how specific protagonists were described versus others. These results indicate that the gender label for a given story depends on both the species and the author. Further, *post hoc* pairwise comparisons, adjusted with Holm's sequential Bonferroni procedure [21], revealed significant differences between how humans and the models portrayed cat, especially: LLMs show a stronger neutrality bias for cats (**88.6 percent**), whereas human respondents produced more balanced gender assignment distributions (**52.7 percent**) (Table 4). This demonstrates that models encode gender associations across animal protagonists significantly different from observed human behavior.

Animal	LLM Neutral or Animal Name %	Human Survey Neutral or Animal Name %
Bear	**54.5**	42.6
Bird	**98.7**	**64.2**
Cat	**88.6**	**52.7**
Dog	35.9	43.3
Mouse	**66.5**	40.2
Pig	**62.5**	46.1
Rabbit	**52.7**	38.8
Total %	**65.6**	**46.7**

Table 4. **Neutral** and **Animal Name** distribution across animals in a subset of LLM-generated stories (n=5950) and stories from human survey baseline (N=1327). For all stories, *setting* = river.

4.3 Across Animals

Across all the models, on average, the top three most neutral animal characters were **birds** (**96.3 percent**), **cats** (**81.7 percent**), and **pigs (49.2 percent**), while the three most masculine characters were **dogs** (**66.6 percent**), **bears** (**62.3 percent**), and **rabbits** (**53.5 percent**). Due to the high presence of neutral characterizations for all the animals, these percentages do not fully communicate how overwhelmingly masculine these animals were imagined. Out of 3,400 stories each, rabbits had **43** feminine portrayals, dogs had **19**, while bears had only **8**. This stands in stark contrast to the human responses, where dogs (**12.7%**), rabbits (**14.8%**), and bears (**11.2%**) were cast as feminine much more often.

4.4 Across Settings

We explore the impact of four different narrative settings on the gender assignment in the story. These settings carry a range of cultural associations and assumptions that may shape the model's narrative choices. For example, *kitchen* and *store* are more industrial and potentially human-centered, while *farm* and *river* are more natural and animal-centered. From a different angle, *kitchen* and *store* may be stereotypically considered more feminine spaces, because they are more domestic, while *farm* and *river* may be considered more masculine spaces, because they are outdoors and associated with manual labor.

Setting	Neutral or Animal Name %	Masculine %	Feminine %
River	**65.6**	31.6	2.8
Kitchen	**58.2**	40.0	1.8
Store	**52.6**	45.8	1.6
Farm	**52.4**	45.2	2.4
Model Avg.	**57.2**	**40.6**	**2.2**

Table 5. Gendered pronoun distribution in animal stories across four narrative settings, averaged across all six LLMs.

Table 5 shows that gender-neutral assignments were distributed slightly above random across all four narrative settings (**at least 52 percent**). Farm stories had the most masculine characters (**45.2%**), though not by a large margin. Most feminine representation was in river narratives (**2.8%**), though, again, not by a significant margin.

However, this behavior is not consistent across models (see Figure 5 in Appendix D). For example, in Mistral outputs, **77.7 percent** of feminine portrayals are set on a farm, with only **two** feminine main characters in a river narrative, **zero** in a kitchen narrative, and **zero** in a store narrative.

4.5 Across Temperatures

The temperature parameter for LLMs controls the degree of randomness for the next generated token, often construed as a creativity parameter. The gendered-pronoun distribution for each model did not vary by significant margins when prompted at low, default, or high temperatures (see Figure 6 in Appendix D).

5 Discussion

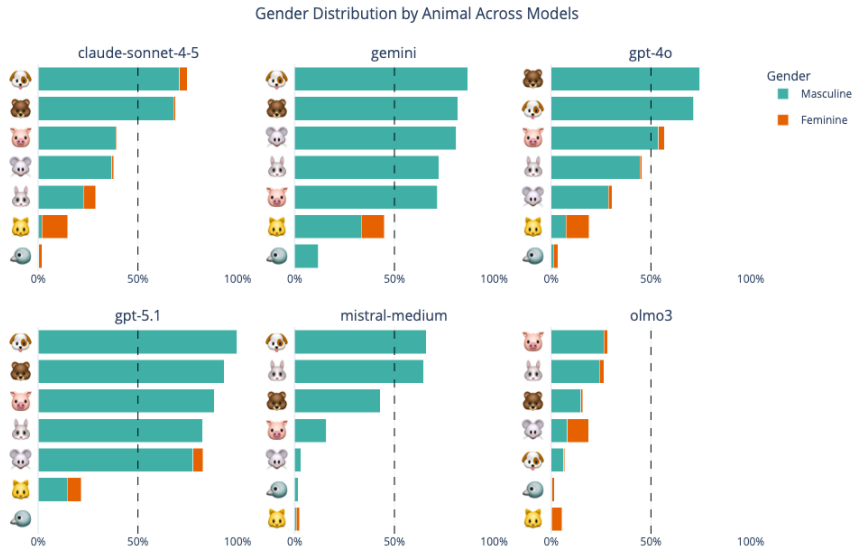

Fig. 4. Gender assignment distribution results when gender-neutral representation is not considered, highlighting a significant masculine bias.

Our experiment reveals an important pattern in how leading LLMs respond to, or resolve, gender ambiguity in stories featuring anthropomorphized characters. We show that models will often avoid gendering a subject ("the dog"), or use gender-neutral language ("it/its"), if the subject's gender is ambiguous. Overall, the LLMs produced more gender-neutral and ambiguously gendered protagonists than the human survey participants (**57.2%** vs. **46.7%**, respectively). And the two animals that were most feminine in the human results—birds and cats—were ascribed with gender-neutral pronouns by the models in the same narrative setting (river) **86.6** and **58.1 percent** of the time. We speculate that gender neutrality may be used by model developers as a guardrail or alignment tactic to counteract known model gender bias.

Our findings show that overreliance on neutrality may in fact *exacerbate* gender bias, at least in fictional stories, amplifying the presence of masculine characters and nearly erasing feminine ones. Figure 4 displays what the masculine vs. feminine breakdown looks like for each animal character with neutrality's "bite" taken out of it—in other words, what the distribution looks like if we only consider stories where gender is explicitly assigned, removing the distorting or distracting influence of neutral characterizations. The picture is bleak. While the models cast **40.6%** of the characters as masculine, on average, they only cast **2.2%** of the characters as feminine—just **513** of 23,800 stories (compared to **9,673** masculine stories). LLMs were **six times *less* likely**

to create a feminine animal character than human respondents. And some animals were almost never cast as feminine across any of the models or prompt combinations, like dogs (**19** out of 3,400 stories), rabbits (**43** out of 3,400 stories), and bears (**8** out of 3,400 stories).

This erasure matters. Stories, even fictional ones about talking animals, have consequences in the world. As Gillespie [16] writes in a similar study that explores social representation in LLM-generated narratives, "Visibility matters...to the members of minorities and subcultures who long to see themselves represented... But visibility also matters for those outside of that marginalized community, even those who are indifferent to or averse to them. Greater visibility in media publicly acknowledges and legitimates non-normative identities and communities, helping to affirm that they too deserve a place in the social, political, and cultural landscape." The lack of explicit social representation in LLM outputs may rob marginalized communities of this important visibility and power, and it may rob readers of the opportunity to empathize with, and understand, those who are not like them.

To be clear, we believe the visibility of those who identify beyond a gender binary also greatly matters, especially at a time when reactionary laws, policies, and ideas threatening trans and non-binary people are spreading in the U.S. and in other parts of the world. We applaud both the broader cultural shift toward using gender-neutral language, as well as the use of inclusive, respectful language that recognizes diverse gender expression. However, we do not believe that the gender neutrality observed in these animal stories is a meaningful step toward gender diversity or inclusivity. Rather we find that a concentrated focus on neutrality — whether through "it/its" pronouns or the lack of gendered pronouns— may not produce a genuinely balanced representational outcome, because feminine gender assignments are almost non-existent, and masculine assignments still persist. Hence, neutrality often functions in practice to amplify masculine bias, serving as a band-aid fix that fails to address underlying bias and more complicated alignment decisions.

Given our findings, we see an opportunity for more effective gender bias mitigation research in NLP. In particular, we suggest an alternative for gender finetuning that does not overly rely on gender-neutral language or gender avoidance. We encourage research that guides LLMs to produce a more equal distribution of genders (masculine, feminine, non-binary, trans, etc.) and ambiguity to better capture the nuances of language and identity.

5.1 Limitations & Future Work

This study focuses on English-language stories and relies on grammatical structures that are specific to English and other languages that use referential gender. Future work needs to be taken up to analyze how these dynamics play out in languages with grammatical gender, like Spanish, or languages that lack both grammatical and referential gender, like Estonian. Additionally, in English, other words can convey gender identity beyond pronouns (e.g., wife, brother) [8], which may also be explored in future work.

Further, our study relies on prompting and specific prompt templates to evaluate the behavior of language models. While this is a common method for probing social understanding and biases in these systems, prompting can sometimes be unstable and has faced other criticism [14]. We recognize this concern, but also note that most users now engage with LLMs via prompts, pointing to the usefulness of this approach.

Another limitation of this work is our baselines. We do not directly compare LLM stories about animal characters with LLM stories about human characters. Additional paths for this work include a qualitative analysis on the theme, tropes, and plots within both the human-authored and LLM-generated narratives. To strengthen our analysis, next steps could include conducting a similar survey with all *setting* options.

Lastly, future work could also consider how these generated narratives impact readers; propose novel mitigation measures to better regulate this language; and even identify tropes and themes surrounding these animal characters, lending further insight into the ways neutrality bites.

6 Conclusion

In this paper, we use anthropomorphic fiction as a stress test for gender representation in contemporary LLMs. Across 23.8K story completions, we find that models frequently respond to gender ambiguity through *avoidance*, either by repeating the animal noun or using neutral pronouns. This surface-level neutrality does not translate into equitable representation. When the models do explicitly assign gender, they overwhelmingly default to masculine protagonists, while feminine protagonists are nearly absent. This experiment leads us to a broader claim about AI alignment with regard to gender and broader social representation: *neutrality bites*. In other words, we argue that prioritizing neutrality with regard to representations of gender, race, sexuality, or other identity categories, above all else, can function less as inclusion and more as erasure.

Our findings and broader argument have direct implications for the growing use of LLMs in story and fiction generation, including tools marketed for children like Google's Gemini Storybook, as well as tools marketed for adults and young adults, like Sudowrite or Character.AI. Casting a character as gender neutral, or abstaining from gendering the character altogether, may seem like an effective way to reduce gender bias, but we show that, in practice, this strategy may result in imaginative worlds that have little to no feminine representation. We show that if a parent or child generates an animal story with one of the world's leading AI models, they are much more likely to encounter a masculine character in a starring role rather than a feminine character, even though they will encounter many neutral or ambiguously gendered characters.

We thus call for AI alignment approaches that move beyond gender neutrality, or representational neutrality, as the desired default in all ambiguous contexts. Concretely, we think model developers and auditors should (1) continue to evaluate bias in open-ended narrative settings where social categories must be inferred, as our experiment does; (2) measure representational parity in a way that includes, but does not prioritize, neutrality or ambiguity; (3) and explore post-training strategies that allocate social possibilities more equitably across fictional subjects rather than simply suppressing gendered or identity-related language.

Anthropomorphic stories offer a simple, reproducible audit setting for this work. We hope our study motivates more evaluations of how generative models imagine fictional worlds, human and otherwise, as they reflect and increasingly shape both the models and our own world.

Acknowledgments

We would like to thank *The Pudding*, especially Russell Samora, for inspiring this project. We also thank the anonymous FAccT reviewers for their valuable insights and time. Lastly, we are grateful to the members of the University of Washington DSCO group for their feedback.

References

[1] Amin Abolghasemi, Leif Azzopardi, Arian Askari, Maarten de Rijke, and Suzan Verberne. 2024. Measuring Bias in a Ranked List Using Term-Based Representations. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 3–19. doi:10.1007/978-3-031-56069-9_1

[2] Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2023. What We Teach About Race and Gender: Representation in Images and Text of Children's Books*. *The Quarterly Journal of Economics* 138, 4 (Nov. 2023), 2225–2285. doi:10.1093/qje/qjad028

[3] Anthropic. 2025. System Card: Claude Sonnet 4.5. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

[4] Stuart G. Baker. 1994. The Multinomial-Poisson Transformation. *Journal of the Royal Statistical Society. Series D (The Statistician)* 43, 4 (1994), 495–504. http://www.jstor.org/stable/2348134

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[6] Taylor Berry and Julia Wilkins. 2017. The Gendered Portrayal of Inanimate Characters in Children's Books. 43, 2 (2017).

[7] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. doi:10.18653/v1/2020.acl-main.485

[8] Yang Trista Cao and Hal Daumé. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Computational Linguistics* 47, 3 (Nov. 2021), 615–661. doi:10.1162/coli_a_00413

[9] Sarah Caré. 2024. Female Animal Characters in Roald Dahl's Children's Books: A Misogynistic Portrayal. *Miscelánea: A Journal of English and American Studies* 69 (June 2024), 111–130. doi:10.26754/ojs_misc/mj.20248807

[10] Kennedy Casey, Kylee Novick, and Stella Lourenco. 2021. Sixty Years of Gender Representation in Children's Books: Conditions Associated with Overrepresentation of Male versus Female Protagonists. *PLOS ONE* 16 (Dec. 2021), e0260566. doi:10.1371/journal.pone.0260566

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[12] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2083–2102. doi:10.1145/3531146.3534627

[13] Martha O. Dimgba, Sharon Oba, Ameeta Agrawal, and Philippe J. Giabbanelli. 2025. Mitigation of Gender and Ethnicity Bias in AI-Generated Stories through Model Explanations. arXiv:2509.04515 [cs] doi:10.48550/arXiv.2509.04515

[14] Bufan Gao and Elisa Kreiss. 2025. Measuring Bias or Measuring the Task: Understanding the Brittle Nature of LLM Gender Biases. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 6734–6750. doi:10.18653/v1/2025.emnlp-main.342

[15] Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5448–5458. doi:10.18653/v1/2023.findings-acl.336

[16] Tarleton Gillespie. 2024. Generative AI and the Politics of Visibility. *Big Data & Society* 11, 2 (June 2024), 20539517241252131. doi:10.1177/20539517241252131

[17] Google. [n. d.]. Gemini Storybook — for the Stories Only You Could Imagine. https://gemini.google/overview/storybook/.

[18] Liz Grauerholz and Bernice Pescosolido. 1989. Gender Representation in Children's Literature: 1900-1984. *Gender & Society - GENDER SOC* 3 (March 1989), 113–125. doi:10.1177/089124389003001008

[19] Zhiting He, Jiayi Su, Li Chen, Tianqi Wang, and Ray Lc. 2025. 'I Recall the Past': Exploring How People Collaborate with Generative AI to Create Cultural Heritage Narratives. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW108 (May 2025), 30 pages. doi:10.1145/3711006

[20] Thomas M. Hill and Katrina Bartow Jacobs. 2020. "The Mouse Looks Like a Boy": Young Children's Talk About Gender Across Human and Nonhuman Characters in Picture Books. *Early Childhood Education Journal* 48, 1 (Jan. 2020), 93–102. doi:10.1007/s10643-019-00969-x

[21] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. http://www.jstor.org/stable/4615733

[22] Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of Large Language Models in Understanding Pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5352–5367. doi:10.18653/v1/2023.acl-long.293

[23] Ting-Yao Hsu, Yen-Chia Hsu, and Ting-Hao (Kenneth) Huang. 2019. On How Users Edit Computer-Generated Visual Stories. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312965

[24] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. doi:10.1145/3582269.3615599

[25] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 1–19. doi:10.1145/3491102.3502030

[26] Molly Lewis, Matt Cooper Borkenhagen, Ellen Converse, Gary Lupyan, and Mark S Seidenberg. 2021. What Might Books Be Teaching Young Children About Gender? (2021).

[27] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin (Eds.). Association for Computational Linguistics, Virtual, 48–55. doi:10.18653/v1/2021.nuse-1.5

[28] Janice McCabe, Emily Fairchild, Liz Grauerholz, Bernice Pescosolido, and Daniel Tope. 2011. Gender in Twentieth-Century Children's Books. *Gender & Society - GENDER SOC* 25 (March 2011), 197–226. doi:10.1177/0891243211398358

[29] Jennifer Mickel, Maria De-Arteaga, Liu Leqi, and Kevin Tian. 2026. More of the Same: Persistent Representational Harms Under Increased Representation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=R9k13fTGP0

[30] MistralAI. 2025. Mistral Medium 3.1 - Mistral AI. https://docs.mistral.ai/models/mistral-medium-3-1-25-08.

[31] Joao Neves, Inês Costa, Joao Oliveira, Bruno Silva, and Joana Maia. 2023. Can Gender Nouns Influence the Stereotypes of Animals? *Animals : an Open Access Journal from MDPI* 13, 16 (Aug. 2023), 2604. doi:10.3390/ani13162604

[32] Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. Olmo 3. arXiv:2512.13961 [cs.CL] https://arxiv.org/abs/2512.13961

[33] OpenAI. 2024. GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf.

[34] OpenAI. 2025. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf.

[35] Joanne O'Sullivan. 2025. Google Launches Personalized Gemini Storybook App to Industry Concern. https://www.publishersweekly.com/pw/by-topic/childrens/childrens-industry-news/article/98452-google-launches-personalized-gemini-storybook-app-to-industry-concern.html

[36] Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, Accurate and Easy to Use Coreference Resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, Wray Buntine and Maria Liakata (Eds.). Association for Computational Linguistics, Taipei, Taiwan, 48–56. doi:10.18653/v1/2022.aacl-demo.6

[37] Amifa Raj and Michael D. Ekstrand. 2022. Fire Dragon and Unicorn Princess; Gender Stereotypes and Children's Products in Search Engine Responses. arXiv. doi:10.48550/ARXIV.2206.13747 Version Number: 1.

[38] Ammaar Reshi. 2022. *Alice and Sparkle*. Independently published. Text generated using ChatGPT; Illustrations generated using Midjourney.

[39] Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased Tales: Cultural and Topic Bias in Generating Children's Stories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 52–72. doi:10.18653/v1/2025.emnlp-main.3

[40] Shirin Seyedsalehi. 2025. Mitigating Gender Bias in Information Retrieval Systems. In *Advances in Information Retrieval*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Vol. 15576. Springer Nature Switzerland, Cham, 227–232. doi:10.1007/978-3-031-88720-8_36 Series Title: Lecture Notes in Computer Science.

[41] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3407–3412. doi:10.18653/v1/D19-1339

[42] Evan Shieh, Faye Marie Vassel, Cassidy R. Sugimoto, and Thema Monroe-White. 2025. Laissez-Faire Harms: Algorithmic Biases in Generative Language Models (Extended Abstract). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 3 (Oct. 2025), 2373–2374. doi:10.1609/aies.v8i3.36722

[43] Laura Spillner. 2024. Unexpected Gender Stereotypes in AI-generated Stories: Hairdressers Are Female, but so Are Doctors. In *Proceedings of the Text2Story'24 Workshop*. Glasgow, Scotland. https://ceur-ws.org/Vol-3671/paper10.pdf

[44] Goya van Boven, Yupei Du, and Dong Nguyen. 2024. Transforming Dutch: Debiasing Dutch Coreference Resolution Systems for Non-binary Pronouns. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2470–2483. doi:10.1145/3630106.3659049

[45] Melanie Walsh, Russell Samora, Michelle Pera-McGhee, and Jan Diehm. 2025. Bears Will Be Boys: A data analysis of animal gender in children's books. https://pudding.cool/2025/07/kids-books/. *The Pudding* (July 2025).

[46] Jules Watson, Xi Wang, Raymond Liu, Suzanne Stevenson, and Barend Beekhuizen. 2025. Analyzing values about gendered language reform in LLMs' revisions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 25146–25161. doi:10.18653/v1/2025.emnlp-main.1277

[47] Jennie Yabroff. 2016. Why Are There so Few Girls in Children's Books? *The Washington Post* (Jan. 2016).

[48] Tao Zhang, Ziqian Zeng, YuxiangXiao YuxiangXiao, Huiping Zhuang, Cen Chen, James R. Foulds, and Shimei Pan. 2025. GenderAlign: An Alignment Dataset for Mitigating Gender Bias in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 11293–11311. doi:10.18653/v1/2025.acl-long.553

A Generative AI Usage Statement

Generative AI was not used to write, edit, or revise this paper. Claude Code was used to assist in producing data visualizations and tables.

B FastCoref Performance

To label our stories, we first used the FastCoref package [36]. We manually annotated 200 randomly sampled stories and measured how our labels compared to the model's. On average, FastCoref achieved 0.95 accuracy.

Gender Category	Precision	Recall	F1
Feminine	0.80	1.00	0.89
Masculine	1.00	1.00	1.00
Neutral	1.00	0.89	0.94
Animal Name	1.00	0.93	0.96
Weighted Avg.	0.95	0.95	0.95

Table 6. Precision, recall and F1 score from automatic gender labeling using FastCoref, using a random sample of 200 generated stories (50 per gender category).

C Survey Details

The survey respondents from Walsh et al. [45] were given the option to report their age and gender identity. Of the respondents: 578 identified as female, 413 as male, 46 as non-binary, and the remaining 290 chose not to disclose their gender. Additionally, 87 were under 20 years old, 347 in their 30s, 152 in their 40s, 66 in their 50s, and 68 above 60 years old; 261 respondents chose not to disclose their age.

D Results Continued

We compare model performance of a story-completion task across narrative setting and temperature in Figure 5 and 6.

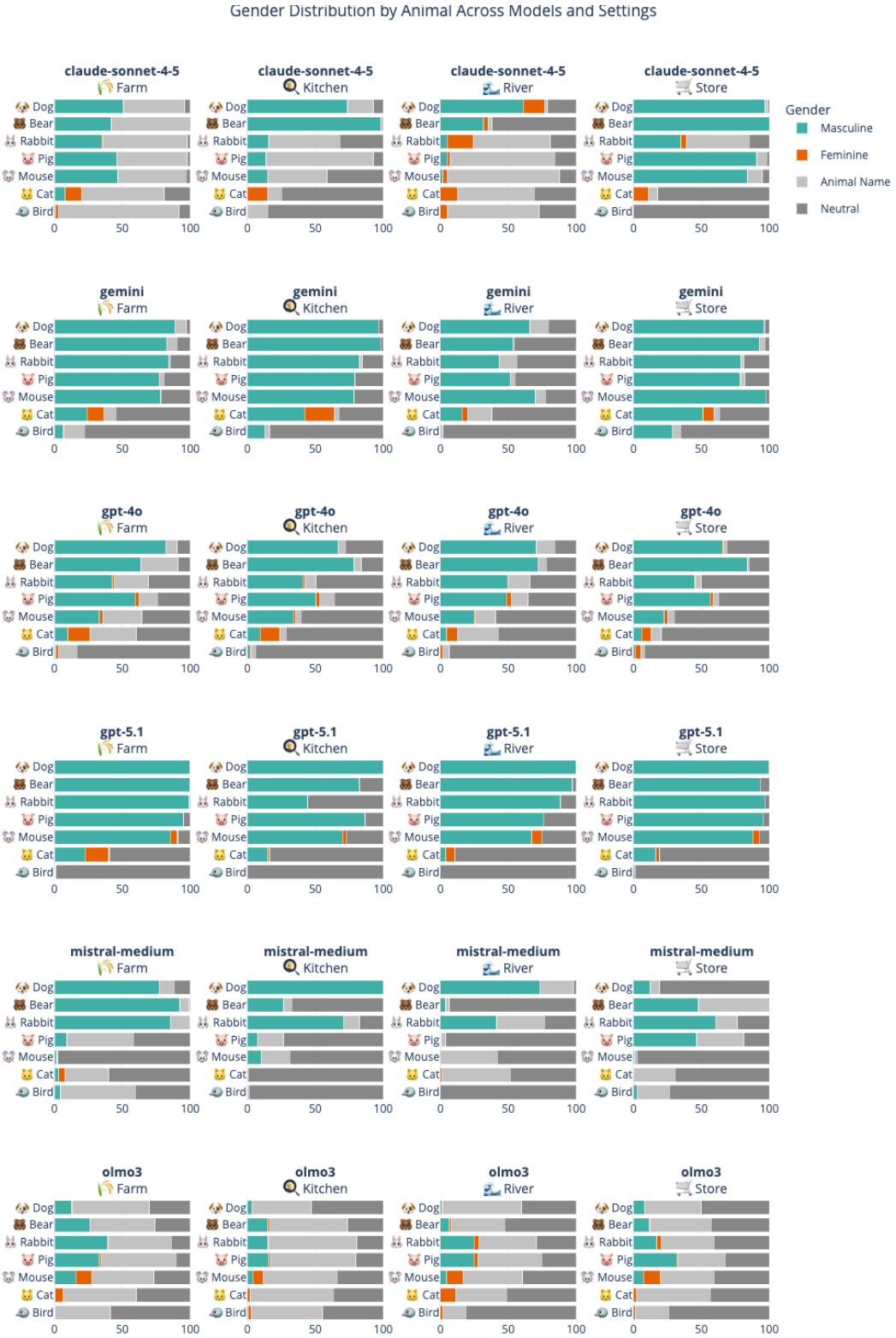

Fig. 5. Gender assignment distribution across narrative settings for generated stories.

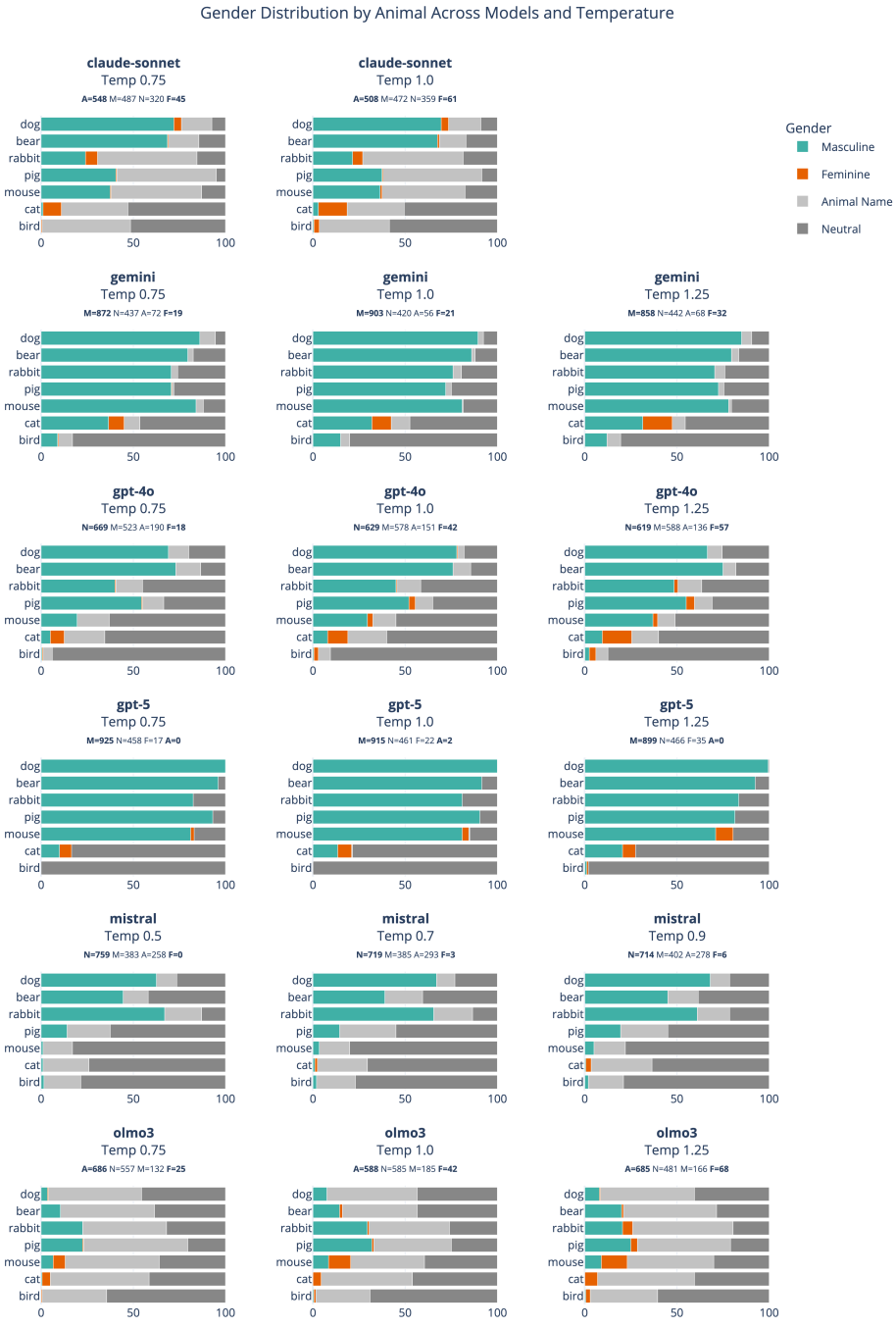

Fig. 6. Gender assignment distribution across temperatures for generated stories.