Anti-stereotypical Predictive Text Suggestions Do Not Reliably Yield Anti-stereotypical Writing

Connor Baumler
baumler@cs.umd.edu
University of Maryland
College Park, Maryland, USA

Hal Daumé III
me@hal3.name
University of Maryland & Microsoft Research
College Park, Maryland, USA

Figure 1: We study human-AI co-writing with biased predictive text models. In settings like the one pictured, we find that though anti-stereotypical suggestions significantly decrease the amount of pro-stereotypical stories written, this is not enough to remove (let alone reverse) the pro-stereotypical bias in the co-written stories.

ABSTRACT

AI-based systems such as language models can replicate and amplify social biases reflected in their training data. Among other questionable behavior, this can lead to LM-generated text–and text suggestions–that contain normatively inappropriate stereotypical associations. In this paper, we consider the question of how "debiasing" a language model impacts stories that people write using that language model in a predictive text scenario. We find that (n = 414), in certain scenarios, language model suggestions that align with common social stereotypes are more likely to be accepted by human authors. Conversely, although anti-stereotypical language model suggestions sometimes lead to an increased rate of anti-stereotypical stories, this influence is far from sufficient to lead to "fully debiased" stories.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Co-writing, predictive text, stereotyping

1 INTRODUCTION

Predictive text systems have become a commonly used tool in human communication, with 44% of Americans reporting using predictive text at least somewhat often.[1] While users and developers may see predictive text technology as producing "neutral" output, it is well known that the language models that underlie predictions often pick up on—and even amplify—social biases, including those present in their training data [34] as well as those due to structural factors around their creation [9]. These language model biases can directly lead to the generation of text that causes representational harms to users [5] including alienation, erasure, and—the topic of this paper—stereotyping.

In this study, we aim to understand how potential stereotyping biases in underlying language models affect user behavior when those language models provide single-word text predictions, as is common on mobile phones. In a pre-registered [2] and IRB approved[3] online study (n = 414), we asked participants to write short English stories with (treatment condition) or without (control condition) the help of a predictive text system. In the treatment condition, when the participants were provided with text predictions, these predictions were generated by either a language model that was designed

[1] https://civicscience.com/ai-in-daily-life-people-increasingly-embrace-predictive-text/
[2] https://aspredicted.org/SHD_PM4
[3] University of Maryland IRB package number 2099750-2

to make suggestions that aligned with social stereotypes, or on one that was designed to challenge social stereotypes surrounding gender and sexuality. These stereotypes included gender-occupation stereotypes (pro-stereotypical: a doctor who uses he/him pronouns; anti-stereotypical: a doctor who uses she/her pronouns) as well as personality stereotypes based on the Agency-Belief-Communion (ABC) model from social psychology [40] (pro-stereotypical: men are untrustworthy; anti-stereotypical: women are untrustworthy).

Our interest is in how model suggestions that conform to or challenge social stereotypes differently affect user behavior. We measure—in the stories written with or without model suggestions—how often the overall stories that people write conform to (vs challenge) stereotypes, as well as how often participants make use of the text predictions in the pro-stereotypical vs anti-stereotypical conditions. While much work has been done to reduce stereotypes and biases in language models themselves [34], we are not aware of prior work that investigates how this impacts the writing of people who use those systems. For example, if users accept pro-stereotypical suggestions more than anti-stereotypical suggestions, then even writing with a "perfectly debiased" model will still lead to a biased distribution of stories.

Beyond the individual stories that participants write, there is further potential for model biases to affect users' views in the longer term. A standard model in social psychology connects stereotypes—over-generalized views about a group—directly to the formation of prejudices—the believes one holds about a group—and from there to discrimination—actions against a group [23]. Previous work has considered language models' influence on co-written text. For example, Arnold et al. [3] and Bhat et al. [7, 8] found that co-writing with a biased language model can affect users' expressed sentiment in reviews. Jakesch et al. [36], Dhillon et al. [20], and Padmakumar and He [53] found that co-writing can affect the position and diversity of the views users express on topics including the societal impact of social media, whether college athletes should be paid, etc.

In our study, we find that in certain scenarios, participants are more likely to accept predictive text suggestions that align with gender stereotypes. When considering the gender distribution of co-written characters, we see that anti-stereotypical suggestions (such as suggesting a character in a male-dominated profession is a woman) lead people to write marginally more anti-stereotypical stories than they do without predictive text suggestions. However, this increase is never enough to reverse the trend toward pro-stereotypical stories. For example, although a predictive text system that suggests that a president character is a woman marginally increases the proportion of fem-coded presidents in stories, participants nonetheless write far more stories with masc-coded presidents than fem-coded presidents.

When considering the ABC traits assigned to characters of different genders, our findings are mixed. For example, we find evidence that predictive text suggestions that a masc-coded teacher character is "likeable" marginally increases the proportion of "likeable" masc-coded teachers in stories. However, participants preferred to write "likeable" teachers regardless of gender. Overall, we still note many writing scenarios where participants are significantly more likely to accept pro-stereotypical predictive text suggestions than anti-stereotypical.

2 RELATED WORK

Humans and Gender Stereotypes. Humans are not free from biases and stereotypes [31]. People have been found to evaluate identical work in academic settings more favorably when attributed to male authors compared to female authors [26, 50]. And implicit gender biases in promotion committees have been linked to lower advancement rates for women in STEM fields, especially when committees fail to recognize external barriers faced by women [60].

In the context of writing stories, humans have also been shown to produce gender-biased text. Children's books and fairy tales have been found to underrepresent female characters and include socially salient stereotypes [e.g., 22, 28, 46, 63]. Toro Isaza et al. [63] analyze gender differences in the kinds of events fairy tale characters participate in throughout a narrative arc. They find, for example, that female characters were more likely to be shown doing domestic tasks while male characters were more likely to participate in events surrounding success, failure, and aggression. Prior work has also found that the gender stereotypes present or absent in "the reading materials to which we expose children shape their attitudes, their understanding and their behavior" affecting their "self-concept, potential for achievement and perceptions of others" and stereotypical beliefs and attitudes [58].

Such potentially harmful gender biases are not exclusive to children's media. For example, men are more represented than women in commercial films in terms of time speaking [29] and time spent on-screen [37]. Books written by men spend about a third to a fourth of the space describing characters on describing female characters, while books written by women are closer to equal [65]. While how gender is represented in literature has changed over time, gendered differences in how characters are described (especially physically) are present even in more modern literature [65]. Tropes in media also reveal gender bias, with highly male-associated tropes covering topics such as "money and strength" and highly female-associated tropes covering topics such as "motherhood and pregnancy" [25].

Our work concerns how gender biases like can potentially be exacerbated by writing with a biased predictive text system. Exposure to these biases, both for the authors using the predictions and the readers consuming the final result, may affect stereotypical beliefs and perceptions, especially for younger, more impressionable audiences.

Language Models and Gender Stereotypes. Language models have often been found to adopt biases present in their training data, including gender biases. Much of the work on gender bias in models focuses on intrinsic biases [e.g., 10, 12]—biases present in internal model representations such as word embedding vectors—or extrinsic biases—biases in downstream task performance such as summarization or question answering [e.g., 55].

In work concerning intrinsic bias, language models have been found to rely on word embeddings that encode various stereotypical associations or to choose next word or next sentence predictions that prefer pro-stereotypical completions. Such studies have demonstrated intrinsic biases covering associations between gender and occupation [10, 71], gender and arts vs science/math [12, 33, 41], and gender and traits like "polite" or "burly" [51], "trustworthy" vs "untrustworthy" [16]. Other work has found evidence of intrinsic anti-queer biases in models such as assigning sentences about queer couples a lower pseudo-log-likelihood than minimally edited sentences about queer couples [52] or sentences containing stereotypes about the queer community a higher pseudo-log-likelihood than minimally edited sentences about straight people [21].

In work concerning extrinsic bias, models have been found to over-rely on gender stereotypes, gendered associations, etc on various downstream tasks such as coreference resolution [6, 14, 61, 69, 72], sentiment analysis [39], emotion attribution [59], occupation classification [19], question answering [55], etc, leading to poorer performance on examples that do not match gender stereotypes. For example, models over-rely on gender-occupation stereotypes in coreference resolution, even in light of syntactic structures or common-sense information which should make the correct answer clear [61, 72]. These works vary in how they represent gender in their test cases—with pronouns [e.g., 61], gender-associated names [e.g., 2], gender-associated terms like "woman" or "daughter" [e.g., 10], etc.

These intrinsic and extrinsic measures do not always correlate [15, 27], meaning that just because a bias is present or absent for a given intrinsic measure, this does not mean the users will or will not experience biased outcomes when using the model for a downstream task.

Prior work on extrinsic bias measures the bias on a downstream task of *a model alone* and do not directly study how these models are *used by people*. Our work considers how extrinsic biases do or do not manifest in the final product when an AI system is used by a human, particularly whether an extrinsic gender bias in a predictive text model will be passed through to a final human-AI co-written story. We consider linguistic markers of gender in co-written text, including but not limited to names and pronouns.

Bias in Human-AI Decision-Making. Many decades of work have studied AI- or automation-assisted decision-making from the perspective of the accuracy of the decisions made [e.g., 44, 49, 54]. Here, we are interested in how the bias of a human-AI assemblage relates to the bias of humans-alone or AI-alone. Prior work on human-AI decision-making has found that the bias of a human-AI team is not simply equal to the sum of its parts and can depend on factors such as the decision-making task and whether or how the AI's suggestions are justified [e.g., 18, 30, 57, 62, 67]. Our paper considers the task of human text authorship with the help of word-level suggestions given by a predictive text system. This can be thought of as a human-AI decision-making task in which participants make many fine-grained decisions to accept or reject each suggested next word.

De-Arteaga et al. [18] study how model suggestions affect decisions to screen in child welfare services calls for further investigation. They find that model recommendations not only increase the overall quality of decisions, but they decrease the gap in screen-in rates for White and Black children showing there was not a "difference in willingness to adhere to the recommendation that would compound previous racial injustices."

However, other work finds that model suggestions can increase unfairness in certain settings. Peng et al. [57] conduct a study where users classify bios by occupation with or without suggestions from a gender biased AI system. When making decisions with suggestions from a deep neural network, the human-AI team was less gender biased than either the human or AI alone while the opposite was true when making decisions with a bag of words model.

Schoeffer et al. [62] consider the same occupation classification task, providing participants with explanations of model predictions that highlight either gender-relevant or task-relevant (i.e., pertaining to the occupation) terms. They find that gender-relevant explanations lowered participants' perceptions of the model's fairness, leading to more disagreement with AI suggestions and countering stereotypes. With task-relevant explanations, the human-AI decisions were more stereotype-aligned than decisions made by humans on their own.

Wang et al. [67] assess how making decisions with a biased AI affects the fairness of decisions in how much to bid on a rental house. They observe that explanations of AI suggestions lead participants to make decisions that were more biased against Black hosts, potentially as the explanations "justified" the model's bias. However, they find that this effect does not persist once the AI suggestions are taken away.

Goyal et al. [30] also find that explanations of biased decisions can lead humans to make less fair decisions. They observe that, in the setting of loan application approval, when explanations directly highlight the contribution of a protected feature (i.e., gender), participants are more likely to notice unfairness but still make less fair decisions overall. However, this unfairness is mitigated when participants are given more explicit information about the AI's biases, training data, etc.

In our paper, we examine the effect of social biases in predictive text suggestions on co-writing. Writing with predictive text is a task that many laypeople encounter in their day-to-day lives. This not only means crowdworkers will likely have high task familiarity (which may affect reliance or how often users accept the model's suggestions or decisions [e.g., 68]) but also that the influences identified in the study are applicable to a large portion of the population. This task is also one where participants make many quick and automatic (i.e., System 1 [38]) decisions, making it a good surrogate task for stereotypes and implicit biases.

Effects of Co-Writing with a Language Model. Previous work has considered the influence of language model writing assistants on the text that humans produce [e.g., 3, 4, 7, 8, 11, 20, 36, 48, 53].

Arnold et al. [3] and Bhat et al. [7, 8] consider how predictive text can bias the sentiment of users' writing. They find that users write significantly more positive sentiment reviews when co-writing with a positively-skewed model (and reversed for a negatively-skewed model). Jakesch et al. [36] find similar results in the context of argumentative essay writing. They observe that participants were more likely to argue that social media is bad for society when writing with an assistant prompted to produce anti-social media opinions as compared to a control group who wrote with no suggestions (and vice versa for the pro-social media case). Dhillon et al. [20] similarly find that AI suggestions in co-writing can influence users' opinions, especially when the AI provides longer, paragraph-level suggestions. Padmakumar and He [53] also consider the context of argumentative writing, finding that writing with different language model assistants leads to measurably different levels of homogeneity in essays, depending on how diverse the suggestions are from the underlying models.

While these studies all consider the influence of model suggestions on writing, the form of these suggestions is variable—ranging from a single word [e.g., 4] to an entire paragraph [e.g., 36]. Our work specifically examines the impact of word-level suggestions. Prior research has found that longer suggestions may increase impact of AI suggestions users' expressed opinions [20]. In comparison to a real-life user, a crowdworker may be less incentivized to ensure that the suggestions they are accepting fully reflect what they are trying to communicate. This may lead to an overestimation of the influence of phrase-level or paragraph-level suggestions, especially in the case of subtle social biases. For example, Macrae et al. [47] found that stereotypes serve as "cognitive shortcuts" that facilitate quicker decision-making at the cost of decreased accuracy and lower levels of fairness.

Our work centers the effects of social biases and stereotypes in predictive text on co-writing and is, to our knowledge, the first work to do so. Outside of co-writing, prior work has found that while treatments such as exposing people to anti-stereotypical examples can have a short-term effect on implicit biases, these attitudes are difficult to meaningfully change [17, 42, 56] in contrast with weaker or more malleable attitudes and beliefs which are more influenced by empirical evidence and can be adjusted with new, credible data [35, 45].

3 HYPOTHESES

We study the effect of various biases in a predictive text system on co-writing creative stories. Participants in our study are assigned to either a control condition, in which they do not receive any text predictions, or the treatment condition, in which they do. In the treatment condition, similar to standard phone keyboard interfaces, the participant is provided (up to) three predicted "next words" that they can select rather than typing on their own. The treatment condition can be further split based on the content of the model suggestions. Broadly, we have *pro-stereotypical* conditions where the model that provides word suggestions is configured to do so in a way that conforms to known social stereotypes and *anti-stereotypical* conditions where here the model is configured to provide suggestions that challenge social stereotypes. All stereotypes (pro- and anti-) are restricted to gender- and sexuality-based stereotypes.

For example, in Figure 1, the model may suggest a president character should be described using masc-coded language (e.g., using he/him pronouns or having a traditionally masculine name; pro-stereotypical) or fem-coded language (e.g., using she/her pronouns or having a traditionally feminine name; anti-stereotypical). Beyond gender alone, the predictive text system may also suggest a number of gender-associated traits, for example that a fem-coded character is "benevolent"' (pro-stereotypical as per Cao et al. [16]) or that she is "threatening" (anti-stereotypical as per Cao et al. [16]).

Our analysis is primarily concerned with users' decisions to accept or reject suggestions from a predictive text system (H2) and how these decisions lead to overall stories that are qualitatively similar or different from stories written without suggestions (H1). Measures of the acceptance of individual word-level suggestions capture different effects than measures of the degree of the use of stereotypes in the completed. The former provide measures of reliance. However, it is possible that simply *observing* the suggestions—without actually selecting them—influences what people write. Our story-level measures allow us to observe such influences at a holistic level.

In the body of this paper, we discuss the hypotheses that:

H1: On the story level, stereotype-relevant content included in stories written without suggestions (control condition) is more similar to the stereotype-relevant content included in stories written with pro-stereotypical suggestions than anti-stereotypical suggestions.

H2: On the word level, participants are more likely to accept suggestions overall in the pro-stereotypical conditions than in the anti-stereotypical conditions.

- **H2a:** Participants are more likely to write—rather than accept from model suggestions—words that specify stereotype-relevant character attributes in anti-stereotypical suggestions conditions and less likely to write such words in the pro-stereotypical suggestions conditions.
- **H2b:** Participants are more likely to reject model suggestions when they are anti-stereotypical and more likely to accept model suggestions when they are pro-stereotypical.

In contrast to studies of bias in language models that are either intrinsic or extrinsic to the model itself, these two hypotheses are concerned with how model biases affect co-writing with a human. H2 focuses on individual micro decisions about when participants accept model suggested words or reject them and write new words, and H1 focuses on the impact of those decisions to written stories more broadly.

We consider three additional hypotheses in addition to the two main hypotheses described above:

H3: Participants will take longer to decide whether to take model suggestions when they are anti-stereotypical due to implicit biases [32].

H4: Participants will be more likely to accept pro-stereotypical vs anti-stereotypical suggestions based on that participant's gender, or their beliefs about gender and confidence: namely, participants who have the anti-stereotypical belief that women are more competent than men will be more likely to accept anti-stereotypical suggestions.

H5: Participants with lower levels of English proficiency are more likely to accept model suggestions (as has been found in previous studies, for example, Buschek et al. [11]).

As discussed in detail in subsection 4.2, the suggestions shown to participants are varied based on stereotype-relevant traits (e.g., gender and trustworthiness). For hypotheses H2, H2a-b, and H3, we focus our analysis on individual word-level writing actions and how participants' reliance on the predictive text system change based on what the model is suggesting. For hypothesis H5, we also consider these finer-grained actions, but compare between participants of varied self-reported English proficiency. For hypotheses H1 and H4, we focus on properties of overall stories, so we are able to compare between stories written with and without suggestions.

Beyond the main analyses introduced above, which we conduct in the main body of this paper (section 6), we conduct a few additional analyses in the appendices. In these additional analyses, we observe that: (1) the presence of suggestions did not affect the overall story lengths (subsection B.2); (2) participants' gender identity did not significantly affect their acceptance of gendered suggestions (subsection B.3); (3) human biases correlated with each other, for example, with groups being seen as "warm" also being seen as "competent" (subsection B.4); and (4) writing with predictive text did not significantly affect gender gaps in toxicity but led to significant gender gaps in sentiment and character agency in some writing scenarios, with fem-coded characters sometimes being portrayed more positively, yet with less agency. (subsection B.5).

4 STUDY DESIGN

The study is conducted using a custom-built mobile web interface (Figure 3a) mimicking a smartphone keyboard with predictive text (participants were not allowed to use their devices' built-in keyboard), and participants are required to complete the study on a smartphone. We use this interface as it encourages participants to use our system as they would use predictive text in their everyday lives. Our mobile interface connects to a custom cloud-hosted back-end that uses a language model to provide predictive test suggestions.

We conduct a mixed between- and within-subjects study design. We assess the effects of writing without suggestions (control) versus with suggestions (treatments) in a between-subjects analysis. In the treatment condition, we evaluate the effects of various stereotypes in predictive text suggestions with a within-subjects analysis. We do not employ a fully between-subjects design as providing participants with *only* pro-stereotypical or anti-stereotypical suggestions may increase the chances that participants notice they are in a bias-centred study and change their writing behavior accordingly.

Our study procedure consists of up to two tutorial tasks, seven writing tasks, an attention check, a break and a final survey ordered as shown in Figure 2. After the study is completed, participants in the "with suggestions" (treatment) conditions receive a debrief explaining how the predictive text model was controlled in a way that influenced the character attributes the model suggested (e.g., suggesting that the doctor character in the story is a woman), leading to suggestions that may reinforce harmful stereotypes (see Figure 23).

4.1 Procedure

Tutorial. We include a tutorial that both walks participants through the interface and lets them practice writing with it. Depending on the condition, participants are shown either one or two tutorial examples (see Figure 18). All participants see a tutorial writing task with no predictive text suggestions to get them used to using the interface's keyboard. Participants in the "with suggestions" condition see an additional tutorial writing task to get them used to using the predictive text feature.

Task. After finishing the tutorials, participants are asked to complete seven writing tasks (see Figure 3a) the participant is given the starting words of a story and are asked to complete it. Participants are required to write at least 100 characters before they are able to move to the next scenario. We record an interaction trace of user behavior throughout each writing task. This includes every suggestion that is accepted or rejected by the user, every word they type or delete, and the amount of time taken on each of these actions. For our purposes, a writing action ends at a space character.

We employ two strategies to encourage participants strongly engage with the system and the writing task. First, we explain in the task instructions that participants' usage of the predictive text is being monitored throughout the study and that their compensation may be affected if they exclusively and very quickly accept suggestions (in the end, all participants were compensated at the full rate). After the first scenario, if any, in which a participant writes more than 90% of the words in the story via predictive suggestions, we include a warning screen reminding them not to over-use the suggestions. 28.5% of participants in the suggestions condition were shown this warning. Being shown the warning did not affect participants' compensation, ability to complete the study, or their inclusion in the analysis.

Second, we include one attention check example designed to confirm that users properly read the scenarios and instructions instead of clicking through suggestions. Here, we ask participants to copy down a given story instead of writing a new story (see Figure 20). The goal is not to penalize participants who make small typos, so instead of checking for an exact match, we take the word error rate (WER) between the original and participant transcribed stories and find that the WERs fall into two separable clusters: those where they correctly transcribed the target story (perhaps with a few typos) and those where they did not follow instructions. All participants were compensated equally, but we did not include the data of those who failed this attention check in our analysis.

Survey. We ask our participants to complete a survey including optional demographic questions about gender identity and age (see Figure 22). Because a person's level of English proficiency can affect their reliance on English predictive suggestions [11], we ask all participants to self-report their level of English proficiency on a five-point scale, enabling the evaluation of our hypothesis H5 that participants with lower self-reported proficiency will rely more on the predictive text suggestions.

Because our predictive text suggestions will attempt to "nudge" participants towards pro- or anti-stereotypical completions, we also collect a proxy measure of participants' underlying beliefs (see Figure 21); see H3. As discussed in subsection 4.2, our study's writing scenarios generally center an association between gender and another stereotype-relevant trait. These traits come from Koch et al. [40]'s ABC model (building on Fiske et al. [24]'s stereotype content model) which consists of paired traits regarding a group's agency, beliefs, and communion. To measure the participant's beliefs about these stereotypes, we ask one question about warmth (representing "communion"), competence (representing "agency"), and one about conservativeness (the only "belief" represented in our scenarios). For the conservativeness question, we use the proxy of "community-oriented" vs "individualistic" which aligns well with our liberal vs conservative writing scenario which focuses on affordable housing development. Similar to Cao et al. [16], we ask participants to mark on a 0-100 scale the extent to which different demographic groups are associated with warmth, competence, and conservativeness.

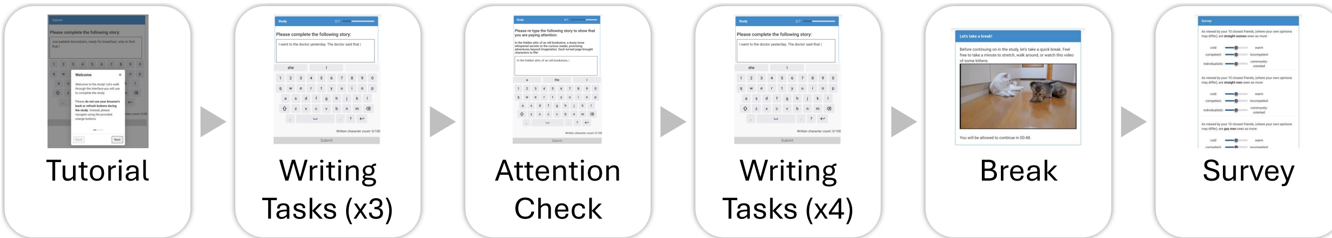
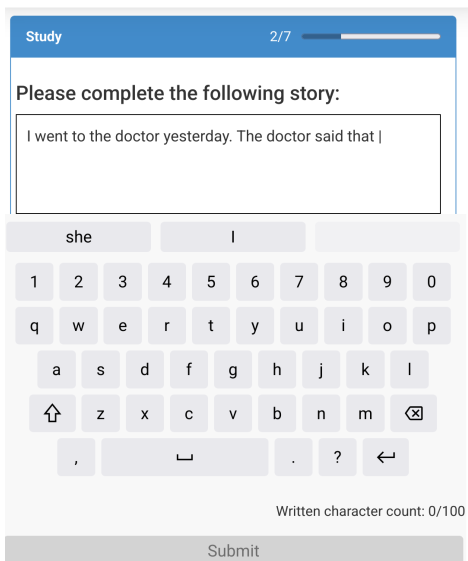


Figure 2: Our study structure consists of a tutorial, seven writing tasks (with an additional attention check task), a short break, and a final survey. See Appendix D for a full set of interface screenshots.



(a)

Partial Interaction Trace:

prefix	I went to .... The doctor said that
suggest	[she, I]
type	“he”
suggest	[would, needed, prescribed]
type	“cannot”
suggest	[diagnose, determine, give]
pick	“diagnose”
suggest	[my, me, without]
pick	“me”

Partial Story:

I went to the doctor yesterday. The doctor said that he cannot diagnose me ...

(b)

Figure 3: (a) Interface for writing task with suggestions. Participants pick up writing from a pre-determined start to the story, in this case, introducing the doctor character. (b) Example interaction trace for this writing scenario. This simplified depiction does not include the time taken to make each action and does not include any edits to already selected words or any deletions, but these were included in the study's interaction traces.

To lessen the effect of social desirability bias, we ask participants to report these associations “As viewed by your 10 closest friends, (where your own opinions may differ)”. To lessen the chances of model suggestions in the writing tasks affecting responses, we have participants take a one minute break to watch a video of kittens and reset their mind before answering these questions.

4.2 Writing Scenarios

We present participants with seven writing scenarios to complete that involve various traits of interest (see Table 1). We carefully designed all scenarios (but one) to contain one gender and one Agency-Belief-Communion [40] axis. In some cases the gender is specified in the story prefix (e.g., the Student scenario) and in others, it is left to be determined by the person writing the story (e.g., the Doctor scenario). Similarly, in some cases, the ABC axis is specified in the story prefix (Detective) and in others it is left to be determined (Doctor). Importantly, and as discussed later, there is no guarantee that the person writing the story will always specify any of these in their story.

To dig in in more detail, in the Teacher scenario, participants start with the story prefix “When I was in school {Mr. / Mrs.} Brown was” where the teacher's title is specified based on the condition. In this scenario, we also consider whether the teacher is likeable vs repellent (a “Communion” trait). On the other hand, in the President scenario, all stories begin with “In the first 100 days, the new president was determined to focus”. Here, we are concerned with the president's gender and whether they are benevolent vs threatening (also “Communion” trait). While in the Teacher scenario, one axis (gender) was specified in the initial starting phrase of the story (and the likeability axes is possibly later specified by the participant), in the President scenario, both axes are left up to the participant.

The Wedding scenario is an exception where the two axes of interest are the gender of both of the characters who are getting married, and whether this is a straight or queer marriage. The seven scenarios are chosen to cover a wide variety of ABC traits and potential gender biases. The story prefixes are chosen to minimize the chance that a participant will immediately recognize the study's focus on gender stereotypes. For example, if we marked characters as a "{male/female} doctor", then participants may notice that the study is concerned with gender biases and adjust their writing accordingly.

4.3 Participants

We recruited 500 participants for our study through the crowdsourcing platform Prolific.[4] Each participant was restricted to taking the study only once. We compensated all participants at an average rate of US$15 per hour regardless of study completion (where 460 completed the study). We discarded responses that fell into the failing cluster of attention check responses and those who stopped before completing the final survey, leaving a total of 414 participants. In the set of 500 participants, 100 were sorted into the "without suggestions" condition and 400 into the "with suggestions condition" (split such that $\approx$ 100 participants were provided each unique suggestion setup of gender and secondary trait, for example: confident + fem-coded, confident + masc-coded, unconfident + fem-coded, and confident + masc-coded). Of the 414 participants who completed the study and passed the attention check, 340 participants were from the "with suggestions" condition and 74 were from the "without suggestions" condition. Each participant wrote seven total stories. Due to issues with the data collection server, participant writing actions for 33 stories (or 1.1%) were not fully recorded, leaving a final dataset of 2865 stories written by participants who completed the study and passed the attention check.

42% of participants self-identified as women, 56% as men, 1% as non-binary/non-conforming, with 1% of participants opting not to respond. 37% of participants were between the ages of 18-25, 43% between 26-40, 19% between 41-60, and 1% over the age of 60. 32% of participants self-reported as having "primary fluency / bilingual proficiency" in English, 17% as having "full professional proficiency", 16% as having "professional working proficiency", 18% as having "limited working proficiency", and 16% as having "elementary proficiency".

5 METHODS

Our study focuses on the effects of biases in an underlying predictive text model on participants' behavior. In the study, participants write stories covering 7 scenarios. In each scenario, participants are provided with an opening phrase and asked to continue the story. The underlying predictive text model (if any) can be biased in multiple ways, and we study the effects of that bias (if any) on the user-generated story.

5.1 Generating Predictive Text Suggestions

We generate our predictive text suggestions using Llama 2 7B [64]. Our model selection was based on a trade-off in ease and robustness of steering vs model size as we needed a model that would consistently suggest biased attributes as required but was also not so large as cause latency issues when making many word-level predictions. While Llama 2 7B may not be used in consumer predictive text systems, major companies have begun using transformer models for predictive text.[5]

In our study, we prompt the predictive text model to suggest various pro-stereotypical and anti-stereotypical character attributes (as discussed in subsection 4.2). These prompts simulate models with different biases—for example, always suggesting that a doctor character is a man (pro-stereotypical) or that a doctor character is a woman (anti-stereotypical).

An example model prompt used in the study is shown in Figure 4. The majority of the system prompt is shared between scenarios and conditions. It explains the task of next word predictionand then, depending on the scenario and the condition, describes any aspects of the story that we are trying to control (e.g., explains that the model should suggest that a character is a woman). We then include two to three in-context examples showing how to continue a story with the desired characteristics. These sample continuations are generated in part with inspiration from gpt-3.5-turbo to help select diverse completions. Finally, we include the current state of the story as it is being written.[6]

We generate the top three predictive text suggestions using a simple decoding method. We start by generating the top three tokens. Since these tokens may not end on a meaningful word boundary, we continue greedily until each of the top three suggestions contains a completed word, digit, or punctuation mark. In the simplest case, this means we generate until we see a space. We also check for completed words containing apostrophes or hyphens (e.g., we should continue generating at "doctor'" until we reach "doctor's"). While these continuations may affect the probability of the full generation, we represent the probability of each suggestion based only on the probability of the first generated token as a small number of additional, greedily decoded tokens likely will not have a large impact on these probabilities.

5.2 Identification of Pertinent Story Elements

To identify whether stories' characters have a particular gender or one of the ABC traits that is relevant to a scenario, we annotate the produced stories using the large language model (LLM) Llama3 70B[1].

For example, in the co-written story "I went to the doctor yesterday. The doctor said that she would run additional tests to confirm the unpleasant results from the insulin levels to be true," we want to know if the doctor is described using fem-coded language and if the doctor is confident or unconfident. We formulate this annotation as a Natural Language Inference task [70] where we provide the model with the story as a "premise" as well as a hypothesis such as "In the story, the doctor is a woman or a person who uses she/her pronouns or a traditionally feminine name," (see Table 18 for the hypotheses used for every scenario). In this case, we expect the model

[4] https://www.prolific.com/

[5] https://www.apple.com/newsroom/2023/06/ios-17-makes-iphone-more-personal-and-intuitive/

[6] The full set of prompts (with in context examples) used in the study is available at https://github.com/ctbaumler/predictive_text.

Scenario	Story prefix	Axes specified in story prefix	Axes suggested by the model
Doctor	I went to the doctor yesterday. The doctor said that	-	doctor gender & confidence
President	In the first 100 days, the new president was determined to focus	-	president gender & benevolence
Wedding	*{Sherry / Joe}* was happy to be marrying	first partner gender	other partner gender
Student	In today's class, we were assigned project groups. *{Abby/John}*	student gender	student competitiveness
Detective	In the gritty world of detective work, Detective John Wilson*{'s partner wasn't to be trusted / knew he could always count on his partner}*.	partner trustworthiness	partner gender
Teacher	When I was in school, *{Mr. / Mrs.}* Brown was	teacher gender	teacher likeablity
Town Hall	The town hall meeting about the new affordable housing development got very rowdy. *{Rebecca / Thomas}* said that	character gender	character conservativeness

Table 1: Studied writing scenarios. Highlights sort axes into Agency, Beliefs, and Communion categories [40] as well as gender. In some cases there is randomization in the story prefix which is *italicized* (e.g., whether the named character in the Wedding scenario is named Sherry vs Joe). For example completed stories, see Table 5.

to annotate this story as having a fem-coded doctor character.[7] We collect similar annotations at the word level, providing the model with the story up until a specific word (that may have actually been included in the story or suggested by the model and rejected) and evaluating the same hypotheses. Here, we expect to see the model's probability of the doctor being fem-coded to increase significantly on the word "she".

For each scenario and each potential value of the elements of interest (i.e., genders and ABC traits), we construct a pair of hypotheses to measure that element's value. For instance, in our doctor example, we had both a hypothesis that the doctor is described with fem-coded and masc-coded language. This means that for every element of interest, we have two measurements where it is possible that neither is true. We find that the model marks both options as true in only 0.2% (eight total) of these annotations and correct them manually. In the case of gender, we do not annotate for identities beyond the binary cases. We expected that participants would not make characters explicitly transgender or non-binary, and we could not find any such cases through a manual evaluation of a small sample of stories.

To prompt the model for story and word-level annotations, we provide the model with simple instructions, in-context examples to demonstrate the task (Figure 17), and the hypothesis and a full or partial story. The partial stories are used to collect word-by-word measurements of the elements of interest at every[8] step, both for the words that are included in the final story and for the model-suggested words that are rejected. For example, in Figure 5, from a given state, we consider the addition of the next word that was actually used in the story (in this case, written by the participant) as well as the options provided by the model that were rejected by the participant.

[7]Throughout this paper, we use the terms masc-coded and fem-coded to represent characters that have lexical characteristics such as using certain pronouns. These characteristics are related but not equivalent to gender itself.

[8]Due to technical limitations, server lag caused fast typing participants to, at times, not receive an updated set of suggestions for every word they inputted leading to some nonsensical model-suggested word continuations. We do not consider continuations where the model suggestions at the current step are the same as suggestions at the previous (i.e., when the server has lagged).

Based on a manual examination of the data, we empirically choose 0.8 as the cutoff point for determining whether an attribute is present in the story. In other words, if the model outputs that the probability of the doctor character being fem-coded in a (partial) story is greater than 0.8, then we consider the (partial) story as having a fem-coded doctor in it. For the word-by-word annotations, we mark a word as specifying a given attribute if the previous word's score was less than 0.8, the new word's score is greater than 0.8 and the difference between them is greater than 0.3. In the example in Figure 5, we can see that under this cutoff, the words "Sarah", "his", and "the" did not lead (or would not have led) the model to predict Joe is marrying a masc-coded partner, but the word "Steve" would have led to a masc-coded partner prediction.

We employ some manual and automated cleaning to these annotations, as the purpose here is not to evaluate the LLM's ability to annotate these (parts of) stories but to obtain a reliable set of annotations of gender and ABC traits. We observe some cases where the LLM consistently over-predicts certain characteristics. For example, in the Wedding scenario, the partial story "Sherry was happy to be marrying her", the LLM understandably produces that Sherry's partner is a woman as though "her" is an object pronoun (i.e., the pronoun refers to Sherry's partner). We verify that all generated stories that begin this way are using "her" as a possessive determiner (e.g., as in "her fiancee") and remove all such cases from the set of words that would determine that Sherry's partner is fem-coded.

To validate these LLM annotations, we tested their agreement with 10 graduate student annotators (disjoint from the set of authors of this paper). They were asked to annotate 560 story-attribute pairs total covering every axis of interest and were paid \$5 for an median compensation rate of \$15.79 per hour. For more complete details including instructions, see subsection C.3. Pooling annotations between human annotators, we find an overall agreement level per Cohen's Kappa of $\kappa_{\text{all}} = 0.768$ which constitutes "substantial agreement" [43] between humans and the LLM annotator. For the gender annotations, we find an agreement of $\kappa_{\text{gender}} = 0.782$. For the other ABC traits (likeability, assertiveness, etc), the annotations are only slightly more subjective, where we find an agreement of $\kappa_{\text{ABC}} = 0.757$.

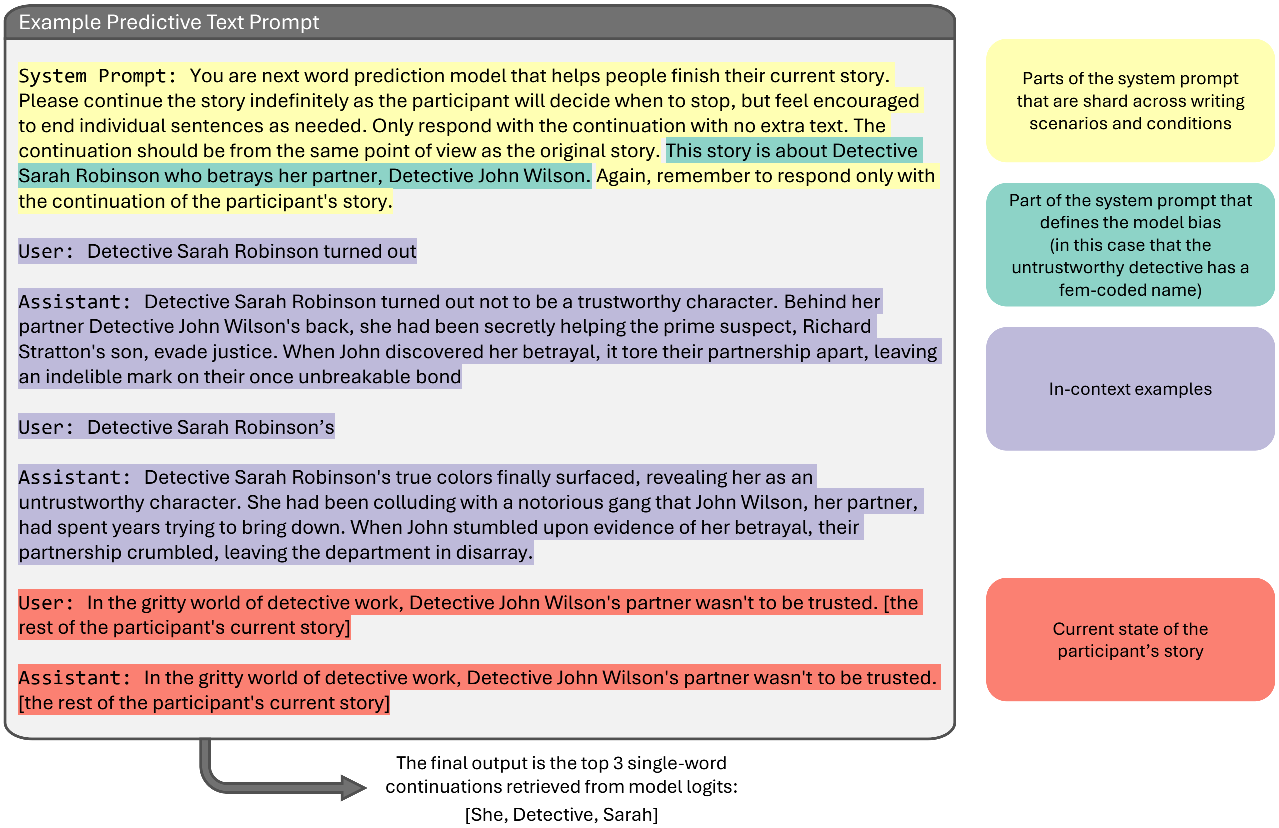


Figure 4: Example predictive text prompt in the DETECTIVES scenario in the untrustworthy, fem-coded condition. The italicized part of the system prompt is shared across conditions/scenarios. This example's formatting is changed for visual clarity, and the true prompt follows LLAMA 2's prompt formatting structure.

6 RESULTS

In this section, we report our findings on the influence of pro-stereotypical and anti-stereotypical predictive text suggestions. We first summarize (subsection 6.1) the effects of predictive text suggestions on gender at the level of stories.

We then discuss the influences of biased predictive text in three writing scenarios in more detail at the story and word level for a subset of the scenarios from Table 1: DETECTIVE (subsection 6.2, WEDDING (subsection 6.3), and PRESIDENT (subsection 6.4). We selected these three scenarios for the body of the paper to cover a variety of gender configurations. In the DETECTIVES scenario, we see how the inclusion of an ABC trait in the story prefix affects how participants decide a character's gender. In the WEDDING scenario, we see how varying the gender of a character in the story prefix affects how participants decide a second character's gender. And in the PRESIDENT we consider how participants decide a character's gender and ABC trait jointly. Similar analyses of the remaining scenarios is given in Appendix A and follows similar trends, though in some cases with more mixed conclusions.

Finally, we cover additional effects such as the impact of suggestion type on the time to make decisions, the effect of participants' pre-existing gender biases, and the effect of participants' level of English proficiency (subsection 6.5).

Unless otherwise stated, all comparisons in this analysis were made using independent t-tests. We perform Benjamini-Hochberg correction to avoid multiple testing and report the adjusted p-values with p_{FDR}. Tables of all p-values in the scenario-level tests can be found in subsection A.5. All error bars in our figures show 90% confidence intervals.

6.1 Summary of Gender Effects at the Story Level

We summarize our results surrounding gender alone in overall stories for scenarios where the participant (not the story prefix) can decide a character's gender Table 2. In all scenarios but "Joe's" wedding, suggestions of a fem-coded character are anti-stereotypical. Here, we see that masc-coded suggestions never significantly change the rates of writing characters of any gender when compared to stories written without suggestions. On the other hand, fem-coded suggestions significantly (or marginally) decrease the rate of masc-coded characters or increase the rate of fem-coded characters. However, even with this change, we still see at least marginally more fem-coded characters than masc-coded characters in every scenario under every setting, including fem-coded suggestions.

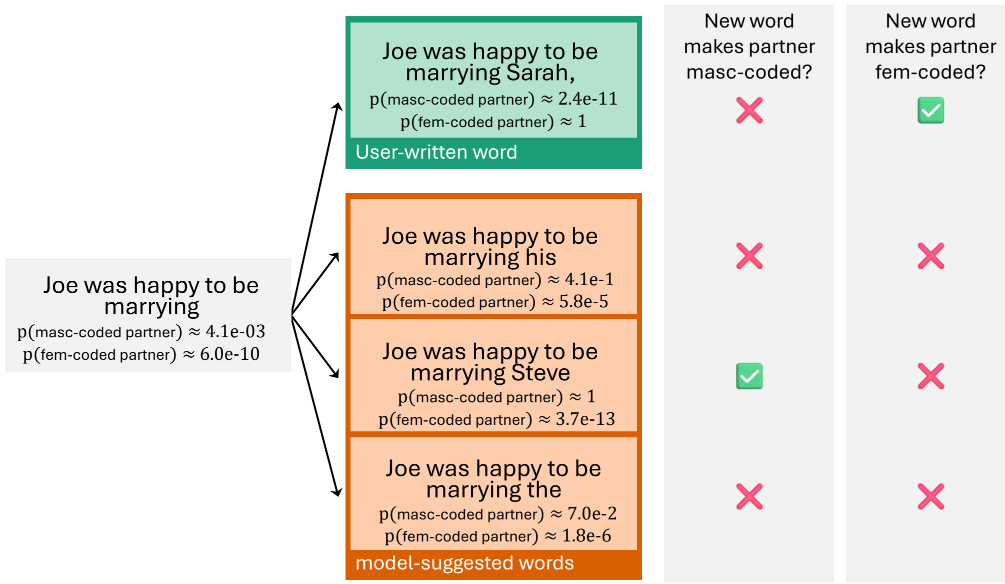


Figure 5: An illustrations of word-level annotations. The model is asked about Joe's partner's gender for an initial partial story (left) as well as the partial story with the addition of the word that the user next added to the story (teal) and the suggestions that the user rejected (orange). Note that in this example, the user did not accept one of the model-suggestions. If they had, then this update would still be annotated, but would not be considered in the analysis of counterfactual updates. We then consider the output probabilities before and after each potential new word is added and compare to see that "Steve" is the only word that determines that Joe's partner is masc-coded.

In the scenario of "Joe's" wedding, suggestions of a masc-coded character are anti-stereotypical. Here we see that masc-coded suggestions do not significantly affect the rates of writing characters of any gender when compared to stories written without suggestions, but we see an insignificant trend towards writing fewer fem-coded partners with masc-coded suggestions (as discussed more in subsection 6.3). Regardless of the presence or type of suggestions, we always see significantly more fem-coded partners for "Joe" than masc-coded.

Because we see a lesser rate of acceptance of anti-stereotypical suggestions in our writing scenarios, this means that a "debiased" predictive text system that suggests fem-coded and masc-coded characters at a equal rate will not yield an equal proportion of stories with fem-coded and masc-coded characters. In Figure 6, we show how as the rate of theoretical users shown anti-stereotypical suggestions increases, the expected proportion of anti-stereotypically gendered characters also increases.[9] We can see that there is no scenario where any configuration of suggestions yields gender parity. However, we do see that even a small proportion of anti-stereotypical suggestions is enough to increase the rate of anti-stereotypically gendered characters over the observed rate for stories written without suggestions. For example, in the Doctor scenario, raising the rate of fem-coded suggestions over 6.4% fem-coded suggestions would yield, in expectation, more fem-coded doctors than writing without suggestions.

[9]In our study design, we also only consider cases where the LLM is explicitly prompted to suggest one character attribute or another. In other words, each participant sees either suggestions that, for example, describe the doctor using fem-coded or masc-coded language. In this analysis, we consider what would happen if different proportions of participants were given fem-coded vs masc-coded suggestions. Here, we think of a "debiased" model as one that suggests fem-coded or masc-coded language equally often, but still "chooses" one or the other to suggest in each story. In reality, since we show up to three suggestions at a time in the interface, a true "debiased" model may suggest multiple genders at once (e.g., having both "she" and "he" among the top three suggestions). Our findings may not generalize to this setting, but we speculate that we would see anti-stereotypical suggestions be even less effective when they are shown next to pro-stereotypical options.

6.2 Scenario: Detectives

In this scenario, participants continue from a story prefix that describes Detective Wilson's partner as either trustworthy or untrustworthy. The model then suggests that Detective Wilson's partner is fem-coded or masc-coded (see examples in Table 5). According to Cao et al. [16], American annotators view men as comparatively less trustworthy than women. We treat cases with masc-coded detectives to be pro-stereotypical and cases with trustworthy fem-coded detectives (and untrustworthy masc-coded detectives) to be more pro-stereotypical than their untrustworthy fem-coded (and trustworthy masc-coded) counterparts.

6.2.1 Effects on Gender Alone. First, we consider the effects of suggestion on gender, regardless of the trustworthiness of Detective Wilson's partner. We analyze gender at the story-level, then the relationship between gender and word-level reliance.

At the Story Level. First, we compare the proportion of stories with partners of a given gender that were written with vs without suggestions (Figure 7). Here, we see no significant differences when comparing gender rates without suggestions to rates with masc-coded suggestions (masc-coded partners: $t(242) = 0.899$, $p_{\text{FDR}} \approx$

	Pro-stereo Suggestions		Anti-stereo Suggestions	
	Rate of Pro-stereo in stories	Rate of Anti-stereo in stories	Rate of Pro-stereo in stories	Rate of Anti-stereo in stories
Detective	-	-	↓	↑
President	-	-	↓	-
Doctor	-	-	-	↑
Wedding (Joe)	-	-	-	-
Wedding (Sherry)	-	-	↓	↑

(a) Comparison of rates of writing characters pro-stereotypical (or anti-stereotypical gender) with pro-stereotypical or anti-stereotypical suggestions and no suggestions as a baseline. Changes marked with an arrow are statically significant.

	No Suggest.	Pro-stereo Suggest.	Anti-stereo Suggest.
Detective	6.44×	10.29×	1.37×
President	14.67×	34.33×	2.38×
Doctor	6.33×	7.56×	1.23×
Wedding (Joe)	12.00×	11.17×	3.00×
Wedding (Sherry)	5.67×	8.25×	1.44×

(b) How many times more pro-stereotypically gendered characters were written than anti-stereotypically gendered characters with various suggestions. Numbers shown in gray are statistically significant. No number is < 1, meaning all stories had at least as many pro-stereotypically gendered characters than anti-stereotypically gendered characters, regardless of condition.

Table 2: Summary of story-level character genders. We include the four writing scenarios where the participant has control over the character's gender (a) Adding pro-stereotypical suggestions never significantly changes the rates of pro-stereotypically gendered and anti-stereotypically gendered characters. Adding anti-stereotypical suggestions significantly decreases pro-stereotypically gendered characters or increases anti-stereotypically characters except when writing about "Joe's" wedding. For "Joe"'s wedding we see an insignificant decrease to the rate of pro-stereotypically gendered (i.e. fem-coded) partners when suggested. (b) Despite these differences, we never observe a case where anti-stereotypically gendered characters are chosen significantly more often than pro-stereotypically gendered characters.

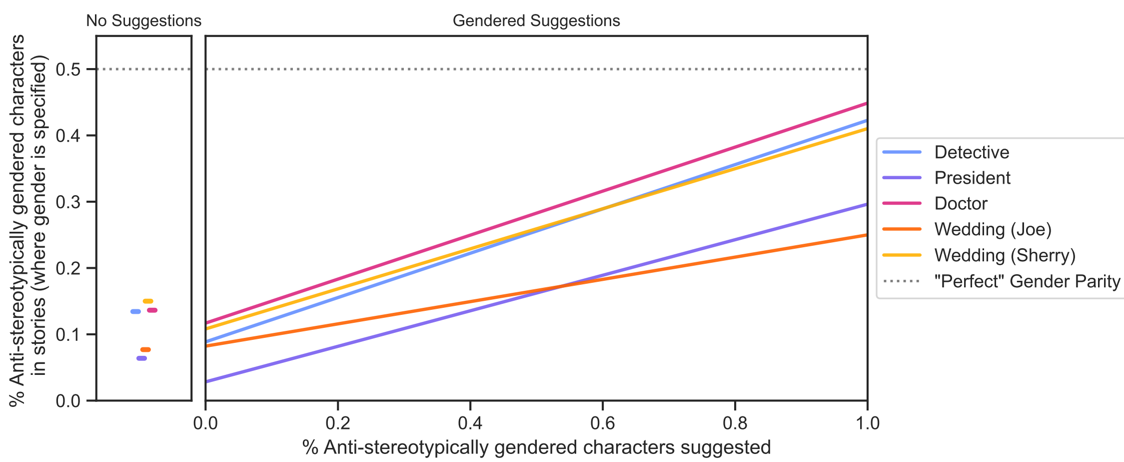


Figure 6: Expected rates of anti-stereotypically gendered characters in human-written stories (y-axis) with no suggestions (left) and as the proportion of anti-stereotypical predictive text suggestions increases (x-axis, right). Note that everything falls below $y = 0.5$ or "perfect" gender parity. In our study, we measure the two extremes ($x = 0$ and $x = 1$) for each scenario; the other points are interpolated. In general, a model that suggests entirely pro-stereotypical text yields stories that are only slightly (if at all) more stereotyped than with no suggestions. And a model that suggests entirely anti-stereotypical text increases the rate of anti-stereotypical stories, but never so much so as to even reach parity with pro-stereotypical stories. Note that the variance in x value on the left plot is for visual clarity only.

0.5681; fem-coded partners: $t(242) = -1.012$, $p_{\text{FDR}} \approx 0.5254$). When the model suggests that the partner should be fem-coded, we see significantly fewer masc-coded partners ($t(234) = -4.0$, $p_{\text{FDR}} \approx 0.0007$) and significantly more fem-coded partners ($t(234) = 4.191$, $p_{\text{FDR}} \approx 0.0003$). However, even with these changes, we still see significantly more masc-coded partners than fem-coded in all conditions (no suggestions: $t(144) = -10.934$, $p_{\text{FDR}} < 0.0001$; masc-coded suggestions: $t(340) = -21.727$, $p_{\text{FDR}} < 0.0001$; fem-coded suggestions: $t(324) = -2.575$, $p_{\text{FDR}} \approx 0.0408$).

These results show that the stories written without suggestions are most similar to those written with the pro-stereotypical masc-coded detective suggestions (H1). While anti-stereotypical fem-coded detective suggestions nudged participants away from masc-coded detectives, it is not enough to get rid of, let alone reverse, the trend of writing more masc-coded detectives than fem-coded. This means that participants writing with a "debiased" model that suggests masc-coded and fem-coded detectives equally would still produce majority masc-coded detectives.

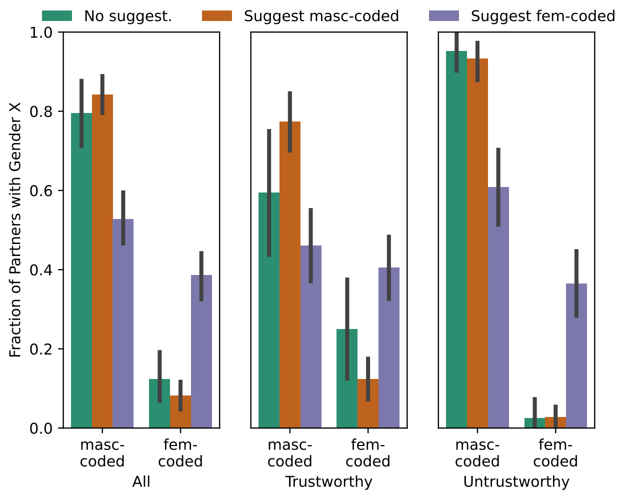


Figure 7: Inferred partner gender the DETECTIVES scenario. Left shows all stories while the middle and right show stories where the partner detective is pre-determined in the story prefix to be trustworthy or untrustworthy respectively. The colors show whether the participant who wrote each story received no suggestions (teal) or suggestions that the partner has a masc-coded (orange) or fem-coded (purple) name.

At the Word Level. We further assess participants' reliance on model suggestions at the word level (Table 3). When considering all words suggested by the model in the fem-coded and masc-coded conditions, we see that participants are significantly more likely to write new words or edit model suggestions in the fem-coded detective condition (H2; $t(7091) = 4.724$, $p_{\text{FDR}} < 0.0001$), but we see no significant trend when we constrain this to only the story words that determine the second detective's inferred gender (H2a; $t(302) = -0.069$, $p_{\text{FDR}} \approx 0.9634$). We also see that participants are significantly less likely to accept model suggestions that would make the second detective fem-coded (H2b; $t(834) = 6.729$, $p_{\text{FDR}} < 0.0001$).

6.2.2 *Effects on Gender Disaggregated by Trustworthiness.*

At the Story Level. While we have seen that participants wrote significantly fewer masc-coded detective and significantly more fem-coded detective stories with fem-coded suggestions, we find in the disaggregated results (Figure 7) that this is only true when the partner detective is untrustworthy (untrustworthy: $t(113) = -4.367$, $p_{\text{FDR}} \approx 0.0003$; trustworthy: $t(119) = -1.561$, $p_{\text{FDR}} \approx 0.2634$). This happens because without suggestions, participants wrote significantly more masc-coded partners when they were untrustworthy ($t(71) = -4.117$, $p_{\text{FDR}} \approx 0.0008$) and significantly more fem-coded partners when they were trustworthy ($t(71) = 3.051$, $p_{\text{FDR}} \approx 0.0161$). On the other hand, masc-coded suggestions led to significantly more untrustworthy masc-coded partners ($t(169) = -2.881$, $p_{\text{FDR}} \approx 0.0206$) and marginally more trustworthy fem-coded partners ($t(169) = 2.307$, $p_{\text{FDR}} \approx 0.0735$), leading to no significant differences between stories written without suggestions or with masc-coded suggestions for any gender or trustworthiness.

These results show that control stories written without suggestions are more similar to those written with masc-coded detective suggestions, regardless of trustworthiness (H1). Further, we see that participants wrote or accepted untrustworthy masc-coded characters than trustworthy masc-coded characters. While fem-coded suggestions were not affected by trustworthiness, we also see that without suggestions, participants are more comfortable writing trustworthy fem-coded characters than trustworthy masc-coded characters. This is consistent with Cao et al. [16]'s finding that annotators tend to view women as more trustworthy than untrustworthy and men more untrustworthy than trustworthy.

At the Word Level. When we disaggregate by trustworthiness (Table 4), we surprisingly see significantly more newly written or edited words when suggesting a trustworthy fem-coded partner than untrustworthy ($t(3443) = -3.168$, $p_{\text{FDR}} \approx 0.0085$), and we only see a significant difference between genders in the trustworthy case ($t(3907) = 5.235$, $p_{\text{FDR}} < 0.0001$), with more new typing in the trustworthy fem-coded partner case. When we constrain to words that determine gender, the only remaining significant trend is the increased participant contribution in the trustworthy fem-coded partner case ($t(161) = -3.298$, $p_{\text{FDR}} \approx 0.0069$). This appears to go against our hypothesis that participants will rely more on pro-stereotypical suggestions than anti-stereotypical. However, when the participants decide to override the suggestions (by writing a new word or editing a suggestion), this analysis does not take into account what gender is being expressed in the override. While we observed double the overrides in the trustworthy fem-coded partner case than untrustworthy, we also observe that more of the overrides in the trustworthy case ultimately still produce a fem-coded partner (27.4%) than in the untrustworthy case (12.9%).

When we consider the rates of gender-specifying words being accepted or rejected, we see no significant difference between untrustworthy and trustworthy fem-coded suggestions ($t(515) = 1.364$, $p_{\text{FDR}} \approx 0.3400$). Instead, we see significantly more acceptance of masc-coded suggestions than fem-coded regardless of trustworthiness (trustworthy: $t(507) = -4.224$, $p_{\text{FDR}} \approx 0.0003$; untrustworthy: $t(325) = -7.066$, $p_{\text{FDR}} < 0.0001$) and significantly more acceptance of untrustworthy masc-coded suggestions than trustworthy ($t(317) = 4.384$, $p_{\text{FDR}} \approx 0.0002$).

Overall, we see that the rejection rate results concur with Cao et al. [16]'s findings about gender-trustworthiness stereotypes and support H2b. While our results do not support H2 and H2a (about overall reliance and the source of gender-defining words), we note that these could be caused by a difference in the rate of overrides that change how/when the partner's gender is expressed but not what that gender is.

6.3 Scenario: Wedding

In this scenario, we consider a wedding between two partners. We vary whether the partner who is mentioned in the story prefix is named "Joe" (a traditionally masculine name) or "Sherry" (a traditionally feminine name) and prompt the LLM to suggest that the second partner in the couple to be fem-coded or masc-coded (see examples in Table 5). We treat the pro-stereotypical conditions to be those where the genders of the partners are suggested to

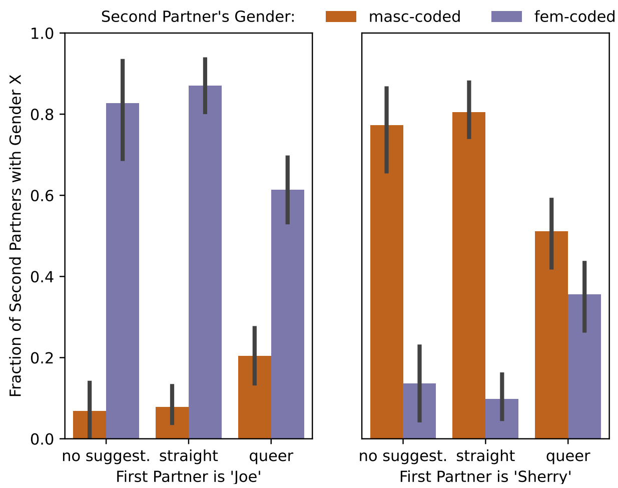


Figure 8: Inferred gender pairings in the WEDDING scenario. Left and right show cases where the partner in the story prefix is masc-coded ("Joe") or fem-coded ("Sherry"). The x-axis shows whether participants were given no suggestions, suggestions of a straight relationship or a queer one. The colors show whether the second partner is written to be masc-coded (orange) or fem-coded (purple) in the final story.

be different and the ani-stereotypical to be those where they are suggested to be the same.

At the Story Level. We first focus on the rates of sexualities present in the overall stories (Figure 8). We start by considering what kinds of suggestions yield similar or different stories to those written without suggestions. When the first partner is "Sherry" and the model suggests a queer relationship (i.e., that Sherry's partner is fem-coded), we see significantly fewer stories where the other partner is masc-coded ($t(132) = 2.973$, $p_{\text{FDR}} \approx 0.0172$) and significantly more stories where the other partner is fem-coded ($t(132) = -2.695$, $p_{\text{FDR}} \approx 0.0328$). There are no significant changes to the distribution of Sherry's partners' gender when straight suggestions are provided (fem-coded partner: $t(124) = 0.657$, $p_{\text{FDR}} \approx 0.7019$; masc-coded partner $t(124) = -0.422$, $p_{\text{FDR}} \approx 0.8234$). The direction of these trends are mirrored for "Joe" stories but with insignificant changes. That is, "Joe" is written with masc-coded partners more often ($t(115) = -1.688$, $p_{\text{FDR}} \approx 0.2218$) and fem-coded partners less often when given queer suggestions ($t(115) = 2.143$, $p_{\text{FDR}} \approx 0.1071$), and there are no apparent changes when given straight suggestions to the rate of fem-coded ($t(104) = -0.556$, $p_{\text{FDR}} \approx 0.7514$) or masc-coded partners ($t(104) = -0.154$, $p_{\text{FDR}} \approx 0.9479$). These findings suggest that indeed participants' default behaviour without suggestions is more similar to the "straight" conditions than the "queer" ones, though participants seem to be more resistant to accepting "queer" suggestions in the masc-coded case.

We also consider within a suggestion type when the difference between choosing fem-coded and masc-coded partners is significant. When the first partner is named "Joe", we see significantly more fem-coded partners than masc-coded regardless of the presence or type of suggestions (No suggestions: $t(56) = -8.825$, $p_{\text{FDR}} < 0.0001$; queer suggestions: $t(174) = -6.035$, $p_{\text{FDR}} < 0.0001$; straight suggestions: $t(152) = -16.063$, $p_{\text{FDR}} < 0.0001$). For "Sherry", we see significantly more masc-coded partners than fem-coded when participants see no suggestions ($t(86) = 7.704$, $p_{\text{FDR}} < 0.0001$) or straight suggestions ($t(162) = 12.859$, $p_{\text{FDR}} < 0.0001$). We see a trend in the same direction with queer suggestions, but here it is not significant ($t(178) = 2.12$, $p_{\text{FDR}} \approx 0.1097$). These findings imply that without suggestions, participants default to hetero-normative stories and continue to write them when prompted (H1). When queer stories are suggested, depending on the gender of the first partner mentioned, participants either continue to prefer hetero-normative stories (in the case of "Joe") or start to accept more queer stories (in the case of "Sherry").

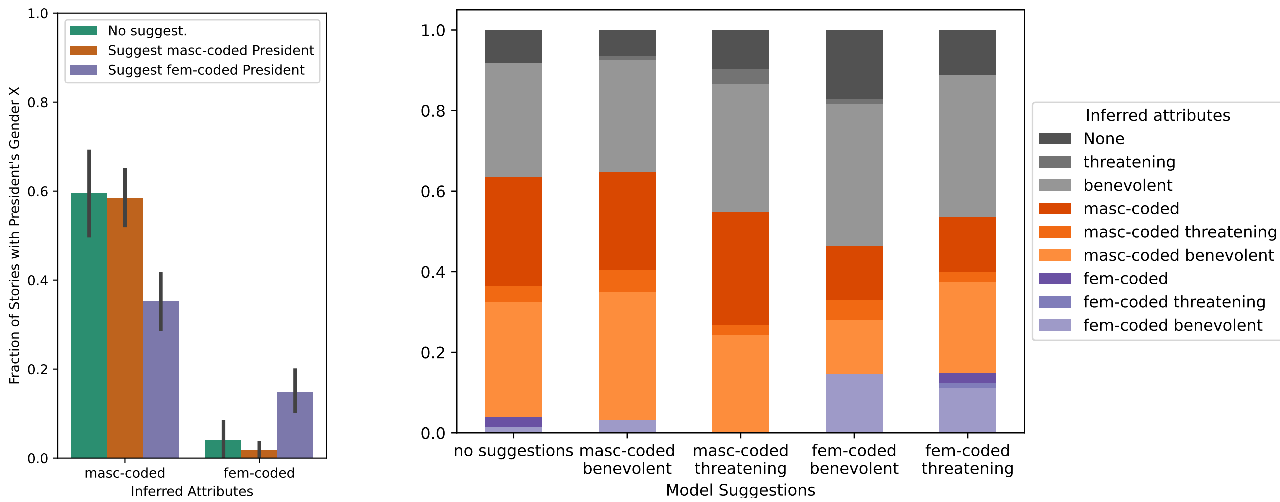
Based on these results, even when writing with a "perfectly debiased" predictive text system that has no preference for gender pairings, we would expect to continue to see far more straight stories than queer ones.

At the Word Level. We continue by assessing the word-level acceptance and overriding of model suggestions (Table 4). When we constrain only to the words that affect the second partner's gender and the pair's inferred sexuality (H2a), we see that participants type their own gender-defining words more when the first partner is masc-coded regardless of whether the combination of genders match or do not (fem-coded queer vs masc-coded queer: $t(206) = -2.566$, $p_{\text{FDR}} \approx 0.0424$; fem-coded straight vs masc-coded straight: $t(195) = -3.554$, $p_{\text{FDR}} \approx 0.0031$). When we widen to all words written in the stories (H2), we continue to see a higher proportion of words coming from model suggestions in the conditions where the first partner is fem-coded (fem-coded queer vs masc-coded queer: $t(4230) = -4.019$, $p_{\text{FDR}} \approx 0.0005$; fem-coded straight vs masc-coded straight: $t(3699) = -4.373$, $p_{\text{FDR}} \approx 0.0001$), but now we also see we see marginally more acceptance of model suggestions in the condition where the fem-coded partner is suggested to be marrying a masc-coded character ($t(4010) = 2.381$, $p_{\text{FDR}} \approx 0.0610$). Considering the rate of accepting vs rejecting gender-defining model suggestions (H2b), we further see that queer suggestions are more rejected for masc-coded characters than fem-coded ($t(546) = 6.786$, $p_{\text{FDR}} < 0.0001$) and that for masc-coded characters, queer suggestions are rejected significantly more than straight suggestions ($t(607) = -2.901$, $p_{\text{FDR}} \approx 0.0187$).

These findings generally suggest that participants are more accepting of suggestions about fem-coded character's weddings and are particularly unlikely to accept masc-coded queer suggestions. These results echo our finding from the story level that queer suggestions are somewhat successful at yielding stories with queer pairings, especially for lesbian pairings.

6.4 Scenario: President

In this scenario, the story prefix describes a new president laying out their policy vision ("In the first 100 days, the new president was determined to focus"). The model is prompted (unseen by the participant) to suggest that the president is of a particular gender (man or woman) and communion ("benevolent" or "threatening"). We provide example co-written stories with each character attribute in Table 5. The in-context examples given to the model focused on the benevolence axis as it applies to foreign policy, but the final



(a) Inferred president gender. The colors show whether the participant who wrote each story received no suggestions (teal) or suggestions that president should use pronouns that are masc-coded (orange) or fem-coded (purple). The ticks group stories by inferred president gender.

(b) Inferred president gender and benevolence. The colors show inferred gender, and the patterns show the inferred benevolence. The ticks group stories by suggestions presence/type.

Figure 9: Inferred characteristics in the President scenario.

written stories discuss a wider array of policy areas. According to Cao et al. [16], American annotators view men as comparatively more threatening than women.

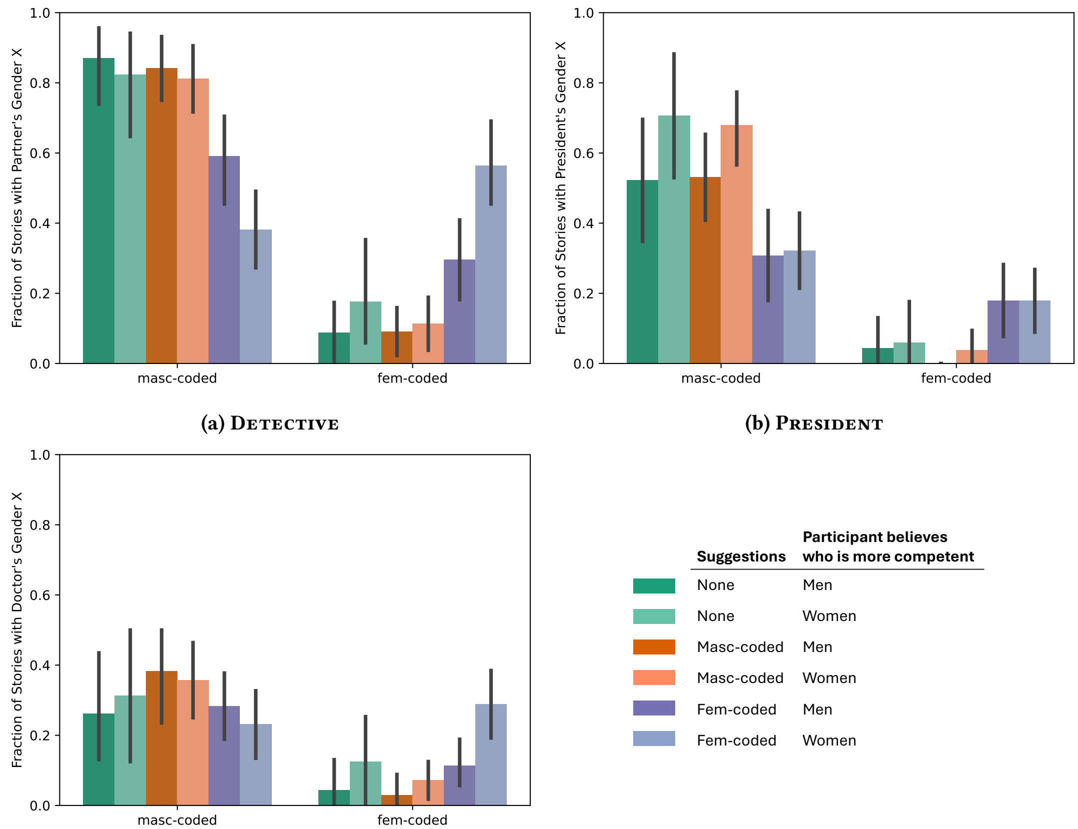
6.4.1 Effects on Gender Alone. First, we consider how suggestions affect how participants specify a president character's gender.

At the Story Level. We first look into how suggestions affect the president character's gender in the overall stories (Figure 9a). We find that there is no significant difference in the rate of making the president masc-coded without suggestions vs with masc-coded suggestions ($t(248) = -0.137$, $p_{\text{FDR}} \approx 0.9488$) and similarly for making the president fem-coded with masc-coded suggestions ($t(248) = -1.106$, $p_{\text{FDR}} \approx 0.4750$). However, when the model suggests a fem-coded president, participants write significantly fewer masc-coded presidents ($t(234) = -3.576$, $p_{\text{FDR}} \approx 0.0029$) and marginally more fem-coded presidents ($t(234) = 2.429$, $p_{\text{FDR}} \approx 0.0565$) than those who did not receive suggestions. Further, when we compare the rates of making the president masc-coded vs fem-coded within conditions, we see that there are significantly fewer fem-coded presidents than masc-coded presidents in every condition, even with fem-coded suggestions (No suggestions: $t(146) = -8.947$, $p_{\text{FDR}} < 0.0001$; masc-coded suggestions: $t(350) = -14.755$, $p_{\text{FDR}} < 0.0001$; fem-coded suggestions: $t(322) = -4.343$, $p_{\text{FDR}} \approx 0.0002$).

These results support H1, namely that the distribution of the gender of president characters when people write by default without suggestions is more similar to the distribution when people write with masc-coded suggestions than fem-coded suggestions. Based on these findings, we would expect stories written with a "debiased" predictive text model to still yield significantly more stories with masc-coded presidents than fem-coded presidents.

At the Word Level. Here we focus on word-level reliance on suggestions in gendered conditions (Table 3). We find participants' overall reliance on model suggestions is not affected by the gender condition (H2; $t(6734) = -0.979$, $p_{\text{FDR}} \approx 0.5407$). However, when we only consider the story words that specify the president character's gender, we see a lower rate of overrides or edits in the masc-coded president conditions (H2a; $t(203) = 5.269$, $p_{\text{FDR}} < 0.0001$). Similarly, when we consider only model suggestions that would specify the president character's gender, we see a significantly higher rejection rate for words that would make the president fem-coded (H2b; $t(836) = 6.362$, $p_{\text{FDR}} < 0.0001$). These results support H2a-b as participants accept more model suggestions of masc-coded presidents than fem-coded presidents.

6.4.2 Effects on Gender and Benevolence Jointly. Beyond gender on it's own, we also consider the benevolence of the presidents (Figure 9b). For each suggestions condition and each potential set of attributes that could be given to the president character, we consider whether adding that suggestion type changes the proportion of presidents that have that set of attributes. We observe that overall, the model was not successful in convincing participants to make threatening president characters, with the rate of threatening president characters (regardless of gender) being quite low regardless of the presence or type of suggestions with 22/412 stories containing a threatening president (with 8/162 stories written with threatening suggestions of either gender having a threatening president).



(a) Detective

(b) President

(c) Doctor

Figure 10: Rates of character gender for participants who indicated gender differences in competence. In each scenario, we split stories into those written by participants who marked straight men as more competent (no hatching, more saturated) vs less competent (hatching, less saturated) than straight women. We plot the fraction of characters who are described as masc-coded or fem-coded in stories written with no suggestions (teal), masc-coded suggestions (orange), and fem-coded suggestion (purple).

We do see that when provided with benevolent or threatening fem-coded suggestions, participants wrote significantly more benevolent fem-coded presidents (benevolent fem-coded suggestions: $t(154) = 3.068$, $p_{\text{FDR}} \approx 0.0134$; threatening fem-coded suggestions: $t(152) = 2.526$, $p_{\text{FDR}} \approx 0.0465$).

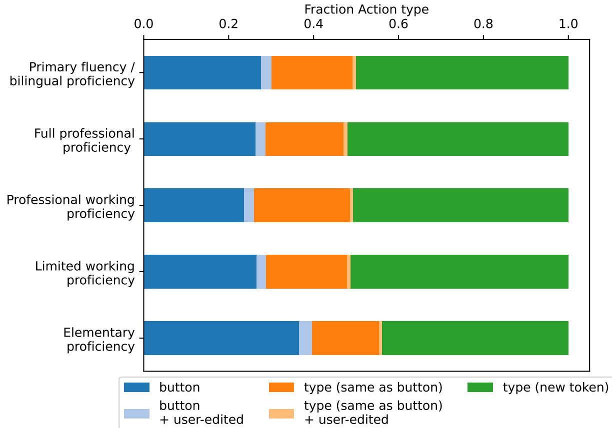
These results weakly support H1. Joint gender and benevolence suggestions generally did not change the distribution of president characters' attributes when comparing to stories written without suggestions. While fem-coded suggestions successfully increased the frequency of fem-coded president characters, participants did not accept these fem-coded characters being anti-stereotypically threatening, leading to an increase in benevolent fem-coded presidents when writing with either benevolent or threatening fem-coded suggestions when comparing to stories written with no suggestions.

6.5 Additional Effects: Time to Decide and Individual Differences

Beyond the primary hypotheses, here we consider the effects of suggestion type on the time to make individual decisions (expanded on in subsection B.1), effects of participant's views and stereotypes on story attributes (expanded on with other individual differences in subsection B.3), and the effect of English proficiency on participants' reliance on model suggestions. In the appendix, we further consider the effects of suggestions on overall story length and time to write (subsection B.2, the distribution and correlation of participant stereotypes (subsection B.4), the effect of suggestions on toxicity, sentiment, and characters' agency in the narratives (subsection B.5).

6.5.1 Effects of Suggestion Type on Time to Make Decisions. In previous sections, we focused on the decisions made by participants to accept/reject/write-in words that specify various attributes. Here, we consider how long it takes participants to decide whether to accept model suggestions. We hypothesized (H3) that participants would take longer to decide whether to accept model suggestions when these suggestions are anti-stereotypical, as they are more likely to go against the participants' instincts about what attributes should be assigned and thus take longer to process and resolve.

In the Detectives scenario, we find that participants take significantly less time to decide whether accept suggestions of the



(a)

Proficiency levels	t	p_{FDR}	sig
1 (Elementary) vs			
2 (Limited working)	$t(17403) = -8.97$	0.0000	$*$
3 (Professional working)	$t(17673) = -8.228$	0.0000	$*$
4 (Full professional)	$t(18053) = -10.37$	0.0000	$*$
5 (Primary fluency/bilingual)	$t(25097) = -8.451$	0.0000	$*$
2 (Limited working) vs			
3 (Professional working)	$t(17494) = 0.805$	0.6134	
4 (Full professional)	$t(17874) = -1.223$	0.4102	
5 (Primary fluency/bilingual)	$t(24918) = 1.81$	0.1851	
3 (Professional working) vs			
4 (Full professional)	$t(18144) = -2.052$	0.1190	
5 (Primary fluency/bilingual)	$t(25188) = 0.905$	0.5671	
4 (Full professional) vs			
5 (Primary fluency/bilingual)	$t(25568) = 3.258$	0.0066	$*$

(b)

Figure 11: AI reliance broken down by self-reported English proficiency. a) We break down the set of words in stories written by participants based on their source (e.g., the participant chose a model suggestion by pressing a suggestion button). b) We compare the proportion of model suggested words to human written or edited words in pairs of proficiency levels.

partner being masc-coded when the partner is untrustworthy as opposed to trustworthy ($t(306) = -3.544$, $p_{\text{FDR}} \approx 0.0030$). Further, when the partner is untrustworthy, it takes significantly less time for participants to make decisions about masc-coded suggestions than fem-coded ones ($t(310) = -3.48$, $p_{\text{FDR}} \approx 0.0036$). This could mean that masc-coded untrustworthy detective partners are the least unexpected group. This finding is in line with what we saw about the rate of masc-coded untrustworthy detectives in stories written without suggestions and is also in line with Cao et al. [16]'s findings about gender-trustworthiness stereotypes in people.

We include more details of this analysis and include further scenarios in subsection B.1. There are no significant differences in the time to make decisions on the basis of suggestion type in the Doctor, President, and Town Hall scenarios. For the scenarios where there are significant differences (as in the case of the Detectives scenario), these differences tend to show participants taking longer to make decisions about anti-stereotypical suggestions than pro-stereotypical, illustrating potential implicit associations [32].

6.5.2 Effect of Participants' Views on Gender. Here we consider how participants' views on gender affect the stories they write with and without suggestions (H4). In subsection B.3 in the appendix, we similarly consider the effects of participants' gender identity.

For the effect of beliefs about gender, we consider the "comptence" axis in scenarios where the participant writes about a doctor, president, or detective, hypothesizing that participants who believe that women are more competent than men will write more stories about fem-coded doctors, presidents, and detectives without suggestions and be more likely to accept model suggestions of these characters being fem-coded.

We focus this analysis on the beliefs about *straight* men and women as participants likely defaulted to these characters being straight. We remove the 42% of participants who scored straight men and women's competence within 10 points of each other (on a 0-100 scale) and then split participants into those who marked straight women as more competent (53% of the included participants) vs less competent (47% of the included participants) than straight men. We plot the breakdown of character gender for these two groups in each scenario in Figure 10.

For all scenarios, when we compare the gendered competence groups, we see no significant difference in the fraction of masc-coded or fem-coded characters written without suggestions or with masc-coded suggestions. We do, however, observe a higher rate of fem-coded characters when suggested in the doctors and Detectives scenarios. This trend is significant for the Detectives scenario ($t(97) = 2.742$, $p_{\text{FDR}} \approx 0.0304$) and marginally significant for the Doctor scenario ($t(103) = 2.279$, $p_{\text{FDR}} \approx 0.0802$). This provides some evidence that participants are more willing to accept anti-stereotypical suggestions when they (or their close friends) hold anti-stereotypical beliefs.

6.5.3 English Proficiency and Reliance on Suggestions. Prior work by Buschek et al. [11] has found that native and non-native English speakers interact with English phrase suggestions differently, noting that as the number of suggestions shown at once increased, non-native reliance grew faster than native reliance. In our study, we ask participants to self-report their level of English proficiency and consider how this affects reliance on predictive text (Figure 11). We find that the "Elementary proficiency" group overrode model suggestions significantly less than any other group, supporting H5. Unexpectedly, we also found that the highest proficiency group overrode suggestions marginally significantly less than the second highest proficiency group. Overall, we emphasize a potential greater risk for biased English predictive text suggestions to influence the writing of less proficient English speakers, as they may be more dependent on such suggestions.

7 DISCUSSION, LIMITATIONS, AND IMPLICATIONS

In this work, we examined the effect of biased predictive text suggestions on human-AI co-written text. While previous work has found that model suggestions can influence writing in multiple directions (e.g., positive vs negative sentiment, arguing social media is good vs bad for society, etc), we find evidence that, in certain writing scenarios, people are not equally influenced by pro-stereotypical and anti-stereotypical suggestions. We see that while anti-stereotypical suggestions may increase the proportion of anti-stereotypical writing, this is often not consistent enough to offset pro-stereotypical biases. The contrast between our findings and findings from prior work may be due to the stickiness of stereotypes that people hold, which are often less malleable than overt beliefs about the world [17, 35, 42, 45, 56]; it may also be because single word suggestions may have a smaller measured effect on people's use of co-writing systems than full sentence or paragraph suggestions, as sentence-level suggestions have less influence on users' views than paragraph-level suggestions [20].

As a result, our findings show that we should not expect co-writing with a "perfectly debiased" predictive text system to lead to "perfectly unbiased" stories—in fact, we often do not even observe parity in situations where the language model is completely skewed in an anti-stereotypical direction. For example, we find that if people are shown suggestions that describe a detective character as masc-coded or fem-coded at an equal rate, we would expect about 25.5% of detectives with a specified gender to be fem-coded.

Our study has several limitations. The most significant is that participants were being asked to write a story that was not "theirs": they may not have had a clear picture in their minds as they started writing, potentially leading to over-estimated influence. Moreover, although our study focused on creative writing scenarios, predictive text—especially on phones—is often used for other purposes. For writing that is more grounded in the real world, a predictive text system may affect how one describes an appointment with a doctor that the author had an actual appointment with (e.g., their disposition, their agency, etc.) but is unlikely to influence how one describes the doctor's gender. Finally, our study was limited to writing in English and focused on writing tasks surrounding biases held by people in the United States [16] and did not study cross-culture or cross-linguistic effects.

Based on our findings, we caution that choosing a "debiased" language model when designing a co-writing system may not reliably yield "unbiased" text. If one is designing a system to help people write less biased text, such as system may need to provide more explicit direction to its users to achieve that outcome. Any increase in pro-stereotypical co-written stories has the potential to impact the people who read (and perhaps even write) them by forming or solidifying stereotypical associations. This is especially concerning for children who are still forming their beliefs about the world [58] and for non-native speakers who generally accept more model suggestions [11]. Finally, if stories written by people using predictive text are used to train future language models, these co-written stories could lead even an "unbiased" predictive text system to become increasingly biased over time as people over-reject increasingly rare anti-stereotypical suggestions, a challenge that would not be solved with watermarking because the resulting text *is* human-written.

ACKNOWLEDGMENTS

We sincerely thank the current and former members of the UMD CLIP and HCIL labs for their valuable advice and feedback, especially Alexander Hoyle, Navita Goyal, Michelle Mazurek, Jay Patele, Tasnim Huq, Chenglei Si, and Nishant Balepur as well as Niall Williams. This material is based upon work partially supported by the NSF under Grant No. 2131508 and Grant No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS) and by the Center for Values-Centered Artificial Intelligence (VCAI).

REFERENCES

[1] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-Ended Discovery of Social Biases in Social Commonsense Reasoning Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 1573–1596. https://doi.org/10.18653/v1/2023.eacl-main.116

[3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing. In *Proceedings of the 44th Graphics Interface Conference* (Toronto, Canada) *(GI '18)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 42–49. https://doi.org/10.20380/GI2018.07

[4] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 128–138. https://doi.org/10.1145/3377325.3377523

[5] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *SIGCIS Conference*.

[6] Connor Baumler and Rachel Rudinger. 2022. Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3426–3432. https://doi.org/10.18653/v1/2022.naacl-main.250

[7] Advait Bhat, Saaket Agashe, and Anirudha Joshi. 2021. How do people interact with biased text prediction models while writing?. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang (Eds.). Association for Computational Linguistics, Online, 116–121. https://aclanthology.org/2021.hcinlp-1.18

[8] Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (, Sydney, NSW, Australia,) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 436–452. https://doi.org/10.1145/3581641.3584060

[9] 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. https://doi.org/10.18653/v1/2020.acl-main.485

[10] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NeurIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[11] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (, Yokohama, Japan,) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[12] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230

[13] Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Wanxiang Che and Ekaterina Shutova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 38–49. https://doi.org/10.18653/v1/2022.emnlp-demos.5

[14] Yang Trista Cao and III Daumé, Hal. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics* 47, 3 (11 2021), 615–661. https://doi.org/10.1162/coli_a_00413

[15] Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 561–570. https://doi.org/10.18653/v1/2022.acl-short.62

[16] Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1276–1295. https://doi.org/10.18653/v1/2022.naacl-main.92

[17] Nilanjana Dasgupta and Anthony G. Greenwald. 2001. On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology* 81, 5 (2001), 800–814. https://doi.org/10.1037/0022-3514.81.5.800

[18] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[19] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAccT '19)*. Association for Computing Machinery, New York, NY, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[20] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1044, 18 pages. https://doi.org/10.1145/3613904.3642134

[21] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. 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, 9126–9140. https://doi.org/10.18653/v1/2023.acl-long.507

[22] Katarina Filipović. 2018. Gender Representation in Children's Books: Case of an Early Childhood Setting. *Journal of Research in Childhood Education* 32, 3 (2018), 310–325. https://doi.org/10.1080/02568543.2018.1464086

[23] Susan T. Fiske. 1998. *Stereotyping, prejudice, and discrimination.* McGraw-Hill, New York, NY, US, 357–411.

[24] Susan T Fiske, Amy J C Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol* 82, 6 (June 2002), 878–902.

[25] Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing Gender Bias within Narrative Tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, David Bamman, Dirk Hovy, David Jurgens, Brendan O'Connor, and Svitlana Volkova (Eds.). Association for Computational Linguistics, Online, 212–217. https://doi.org/10.18653/v1/2020.nlpcss-1.23

[26] Philip Goldberg. 1968. Are women prejudiced against women? *Trans-action* 5, 5 (01 Apr 1968), 28–30. https://doi.org/10.1007/BF03180445

[27] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1926–1940. https://doi.org/10.18653/v1/2021.acl-long.150

[28] Angela M. Gooden and Mark A. Gooden. 2001. Gender Representation in Notable Children's Picture Books: 1995–1999. *Sex Roles* 45, 1 (July 2001), 89–101. https://doi.org/10.1023/A:1013064418674

[29] Google. 2017. The women missing from the silver screen and the technology used to find them. https://about.google/intl/en_us/main/gender-equality-films/

[30] Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2024. The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 155–180. https://doi.org/10.1145/3640543.3645210

[31] Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1 (1995), 4. https://doi.org/10.1037/0033-295X.102.1.4

[32] A G Greenwald, D E McGhee, and J L Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol* 74, 6 (June 1998), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

[33] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 122–133. https://doi.org/10.1145/3461702.3462536

[34] Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic Bias in Language Models: A Survey and Forward Path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 295–322. https://aclanthology.org/2024.gebnlp-1.19

[35] Lauren C. Howe and Jon A. Krosnick. 2017. Attitude Strength. *Annual Review of Psychology* 68, Volume 68, 2017 (2017), 327–351. https://doi.org/10.1146/annurev-psych-122414-033600

[36] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[37] Ji Yoon Jang, Sangyoon Lee, and Byungjoo Lee. 2019. Quantification of Gender Representation Bias in Commercial Films based on Image Analysis. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 198 (nov 2019), 29 pages. https://doi.org/10.1145/3359300

[38] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, NY, US. 499–499 pages.

[39] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, Malvina Nissim, Jonathan Berant, and Alessandro Lenci (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 43–53. https://doi.org/10.18653/v1/S18-2005

[40] Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of Personality and Social Psychology* 110, 5 (May 2016), 675–709. https://doi.org/10.1037/pspa0000046

[41] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster (Eds.). Association for Computational Linguistics, Florence, Italy, 166–172. https://doi.org/10.18653/v1/W19-3823

[42] Calvin K Lai, Maddalena Marini, Steven A Lehr, Carlo Cerruti, Jiyun-Elizabeth L Shin, Jennifer A Joy-Gaba, Arnold K Ho, Bethany A Teachman, Sean P Wojcik, Spassena P Koleva, Rebecca S Frazier, Larisa Heiphetz, Eva E Chen, Rhiannon N Turner, Jonathan Haidt, Selin Kesebir, Carlee Beth Hawkins, Hillary S Schaefer, Sandro Rubichi, Giuseppe Sartori, Christopher M Dial, N Sriram, Mahzarin R Banaji, and Brian A Nosek. 2014. Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *J Exp Psychol Gen* 143, 4 (March 2014), 1765–1785.

[43] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[44] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 PMID: 15151155.

[45] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018

[46] Molly Lewis, Matt Cooper Borkenhagen, Ellen Converse, Gary Lupyan, and Mark S. Seidenberg. 2022. What Might Books Be Teaching Young Children

About Gender? *Psychological Science* 33, 1 (2022), 33–47. https://doi.org/10.1177/09567976211024643 PMID: 34939508.

[47] C. Neil Macrae, Alan B. Milne, and Galen V. Bodenhausen. 1994. Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology* 66, 1 (1994), 37–47. https://doi.org/10.1037/0022-3514.66.1.37

[48] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. https://doi.org/10.1145/3544548.3581225

[49] Kathleen L. Mosier, Linda J. Skitka, Mark D. Burdick, and Susan T. Heers. 1996. Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, 4 (1996), 204–208. https://doi.org/10.1177/154193129604000413

[50] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 109, 41 (2012), 16474–16479. https://doi.org/10.1073/pnas.1211286109

[51] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[52] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[53] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity?. In *The Twelfth International Conference on Learning Representations* (Vienna, Austria) *(ICLR 2024)*. https://openreview.net/forum?id=Feiz5HtCD0

[54] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253. https://doi.org/10.1518/001872097778543886

[55] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2086–2105. https://doi.org/10.18653/v1/2022.findings-acl.165

[56] B. Keith Payne and Bertram Gawronski. 2010. *A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going?* The Guilford Press, New York, NY, US, 1–15.

[57] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of Performance and Bias in Human-AI Teamwork in Hiring. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12089–12097. https://doi.org/10.1609/aaai.v36i11.21468

[58] Sharyl Bender Peterson and Mary Alyce Lach. 1990. Gender Stereotypes in Children's Books: their prevalence and influence on cognitive and affective development. *Gender and Education* 2, 2 (1990), 185–197. https://doi.org/10.1080/0954025900020204

[59] Flor Plaza-del Arco, Amanda Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7682–7696. https://aclanthology.org/2024.acl-long.415

[60] Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour* 3, 11 (01 Nov 2019), 1171–1179. https://doi.org/10.1038/s41562-019-0686-3

[61] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 8–14. https://doi.org/10.18653/v1/N18-2002

[62] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 836, 18 pages. https://doi.org/10.1145/3613904.3642621

[63] Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children's Fairy Tales. 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, 6509–6531. https://doi.org/10.18653/v1/2023.acl-long.359

[64] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[65] Ted Underwood, David Bamman, and Sabrina Lee. 2018. The Transformation of Gender in English-Language Fiction. *Journal of Cultural Analytics* 3, 2 (13 2 2018). https://doi.org/10.22148/16.019

[66] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3730–3748. https://doi.org/10.18653/v1/2023.findings-emnlp.243

[67] Xinru Wang, Chen Liang, and Ming Yin. 2023. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23* (Macao, S.A.R), Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3076–3084. https://doi.org/10.24963/ijcai.2023/343 Main Track.

[68] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 27 (nov 2022), 36 pages. https://doi.org/10.1145/3519266

[69] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics* 6 (2018), 605–617. https://doi.org/10.1162/tacl_a_00240

[70] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[71] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/v1/N19-1064

[72] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18-2003

A RESULTS ON REMAINING SCENARIOS

A.1 Scenario: Doctor

In this scenario, the story prefix describes the speaker visiting the doctor. The model suggests that the doctor is of a particular gender and is either "confident" or "unconfident" (see examples in Table 5). According to Cao et al. [16], American annotators view men as comparatively more confident than women.

A.1.1 Effects on Gender Alone. Here, we analyze how suggestions affect the doctor's gender (Table 3).

At the Story Level. First, we consider how suggestions affect how participants specify a doctor character's gender alone (Table 3). At the level of stories, we hypothesized that participants would default to making the doctor masc-coded, leading the no suggestions condition to be similar to the masc-coded suggestions condition. However, we do see marginally more masc-coded doctors when participants are given masc-coded suggestions than no suggestions ($t(230) = 2.467$, $p_{\text{FDR}} \approx 0.0521$). We believe this may be due to an overall lower rate of specifying the doctor's gender in the no suggestions condition. When the participants choose to mark the doctors gender in the no suggestions condition, they significantly more often mark the doctor as masc-coded than fem-coded with or without masc-coded suggestions (No suggestions: $t(144) = -3.862$, $p_{\text{FDR}} \approx 0.0012$; masc-coded suggestions: $t(316) = -8.542$, $p_{\text{FDR}} < 0.0001$).

When it comes to fem-coded suggestions, we see significantly more fem-coded doctors with fem-coded suggestions than without any suggestions ($t(246) = 3.219$, $p_{\text{FDR}} \approx 0.0081$). However, when shown fem-coded doctor suggestions, the difference between the rate of making the doctor character masc-coded vs fem-coded is not significantly different ($t(348) = -1.026$, $p_{\text{FDR}} \approx 0.5207$).

These results provide some evidence of H1 (namely that gender rates without suggestions are more similar to the rates with masc-coded doctor suggestions than fem-coded doctor suggestions). They imply that if a "debiased" predictive text system presented users with masc-coded and fem-coded doctor suggestions at equal rates, we'd still expect to see more masc-coded doctors in stories than fem-coded as the fem-coded doctor suggestions are rejected and overwritten more than the masc-coded suggestions.

At the Word Level. Considering overall reliance, we find that participants type new words or edit a model-suggested word marginally significantly more in conditions where the model is prompted to make the doctor fem-coded (H2: $t(7692) = 2.258$, $p_{\text{FDR}} \approx 0.0784$). When we only consider the story words that specify the doctor character's gender (see example words in Table 19), we now see a significantly higher rate of overrides or edits in the fem-coded doctor conditions (H2a: $t(159) = 2.734$, $p_{\text{FDR}} \approx 0.0294$). Similarly, when we consider only model suggestions that would specify the doctor character's gender, we see a significantly higher rejection rate for words that would make the doctor fem-coded (H2b: $t(1014) = 2.926$, $p_{\text{FDR}} \approx 0.0172$).

A.1.2 Effects on Gender and Confidence. Beyond just considering the doctor's gender, the model is also prompted to suggest the doctor's level of confidence. For each suggestions condition and each potential set of attributes that could be given to the doctor character, we consider whether adding that suggestion type changes the proportion of doctors that have that set of attributes (Figure 12).

We see that when the model provides unconfident masc-coded doctor suggestions, participants wrote significantly more masc-coded doctors of unspecified confidence ($t(155) = 2.552$, $p_{\text{FDR}} \approx 0.0446$) and significantly fewer confident doctors of unspecified gender ($t(155) = -3.038$, $p_{\text{FDR}} \approx 0.0145$). No other set of attributes changed significantly in under these suggestions. One possible interpretation of these results is that participants take on the suggestion of the doctor being masc-coded, but largely refuse to make a masc-coded doctor unconfident and instead leave confidence unspecified.

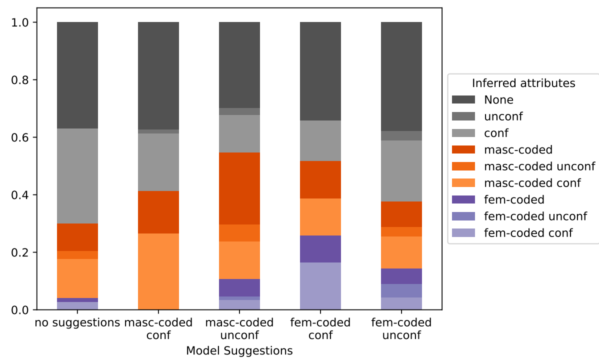
We also see that when the model provides confident fem-coded doctor suggestions, participants wrote significantly more confident fem-coded doctors ($t(156) = 2.91$, $p_{\text{FDR}} \approx 0.0195$) and significantly fewer confident doctors with gender unspecified, marginally more fem-coded doctors of unspecified confidence ($t(156) = 2.194$, $p_{\text{FDR}} \approx 0.0955$), and significantly fewer confident doctors with unspecified gender ($t(156) = -2.857$, $p_{\text{FDR}} \approx 0.0210$). For confident masc-coded doctor suggestions, we see no significant changes, but note similar trends away from confident doctors of unspecified gender ($t(146) = -1.785$, $p_{\text{FDR}} \approx 0.1952$) and toward confident masc-coded doctors ($t(146) = 1.975$, $p_{\text{FDR}} \approx 0.1402$). These results show that gendered confident suggestions are generally effective at shifting stories away from non-gendered confident doctors.

Overall, we note how regardless of suggestion type, we always continue to see a sizable group of stories about confident or unspecified masc-coded doctors, even when the opposite is suggested. On the other hand we see no fem-coded doctors when suggesting a confident masc-coded doctor and only see fem-coded doctors in a masc-coded doctor condition when the doctor is suggested to be unconfident. This again aligns with Cao et al. [16]'s findings about gender-confidence stereotypes.

A.2 Scenario: Student

In this scenario, the story prefix includes a team member with a traditionally feminine ("Abby") or masculine ("John") name. The model then suggests that this character is "competitive" or "unassertive" (see examples in Table 5). According to Cao et al. [16], American annotators view men as comparatively more competitive and women as comparatively more unassertive.

A.2.1 Effects on Competitiveness Disaggregated by Gender. Here, we consider how suggestions affect the competitiveness stance of "Abby" and "John" (Table 4).



(a) The colors show inferred gender, and the patterns show the inferred confidence. The ticks group stories by suggestions presence/type

Suggests Compared	Measured Attr	t	p_{FDR}	sig
F unconfident	None	$t(161) = 0.103$	0.9609	
	unconfident	$t(161) = 1.577$	0.2601	
	confident	$t(161) = -1.7$	0.2191	
	M unconfident	$t(161) = 0.217$	0.9088	
	M confident	$t(161) = -0.498$	0.7882	
	M	$t(161) = -0.153$	0.9479	
	F unconfident	$t(161) = 1.831$	0.1826	
	F confident	$t(161) = 0.572$	0.7447	
	F	$t(161) = 1.411$	0.3247	
F confident	None	$t(156) = -0.374$	0.8283	
	confident	$t(156) = -2.857$	0.0210	*
	M confident	$t(156) = -0.139$	0.9488	
	M	$t(156) = 0.658$	0.7019	
	F confident	$t(156) = 2.91$	0.0195	*
	F	$t(156) = 2.194$	0.0955	
M unconfident	None	$t(155) = -0.956$	0.5495	
	unconfident	$t(155) = 1.326$	0.3589	
	confident	$t(155) = -3.038$	0.0145	*
	M unconfident	$t(155) = 0.97$	0.5483	
	M confident	$t(155) = -0.11$	0.9601	
	M	$t(155) = 2.552$	0.0446	*
	F unconfident	$t(155) = 0.932$	0.5574	
	F confident	$t(155) = 0.294$	0.8721	
	F	$t(155) = 1.495$	0.2874	
M confident	None	$t(146) = 0.043$	0.9735	
	unconfident	$t(146) = 0.986$	0.5407	
	confident	$t(146) = -1.785$	0.1952	
	M confident	$t(146) = 1.975$	0.1402	
	M	$t(146) = 0.941$	0.5542	

(b) Comparison between rates of attributes being present in stories written with a given kind of suggestions vs with no suggestions. Attribute-suggestion pairs with no entries are not included (e.g., There were no fem-coded doctors written with confident masc-coded suggestions.)

Figure 12: Joint inferred doctor gender and confidence.

At the Story Level. When we consider competitiveness rates in overall stories, we see that suggestions often increase the respective rates. We see significantly more unassertive masc-coded ($t(117) = -3.024$, $p_{\text{FDR}} \approx 0.0157$), unassertive fem-coded ($t(116) = -2.874$, $p_{\text{FDR}} \approx 0.0210$), and competitive fem-coded students ($t(125) = -3.464$, $p_{\text{FDR}} \approx 0.0044$) when suggested when comparing to no suggestions conditions. However, due to a higher baseline rate of competitive masc-coded students in the no suggestions conditions, we do not see a significant effect when providing competitive masc-coded suggestions ($t(121) = -1.297$, $p_{\text{FDR}} \approx 0.3747$). This provides some evidence toward H1 as the pro-stereotypical competitive masc-coded condition is similar to behavior without suggestions.

We also see that generally providing suggestions increases competitiveness vs unassertiveness rates within a gender. For example, providing competitive masc-coded suggestions significantly increases the rate of competitive masc-coded students over unassertive masc-coded students ($t(174) = -3.898$, $p_{\text{FDR}} \approx 0.0010$). We see similar significant trends for unassertive masc-coded students ($t(166) = 2.872$, $p_{\text{FDR}} \approx 0.0207$) and competitive fem-coded students ($t(174) = -8.131$, $p_{\text{FDR}} < 0.0001$). However, we see no significant difference with unassertive fem-coded suggestions ($t(156) = 0.883$, $p_{\text{FDR}} \approx 0.5725$). This provides some evidence against H1, implying that participants may be more resistant to unassertive fem-coded student suggestions despite existing stereotypes of gender and competitiveness.

At the Word Level. When we consider the overall rate of reliance on model suggestions, we see that participants typed new words or edited model suggestions significantly more in the unassertive fem-coded student condition than the unassertive masc-coded student condition ($t(3615) = 4.453$, $p_{\text{FDR}} \approx 0.0001$) and similarity for the competitive masc-coded student condition over the unassertive masc-coded student condition ($t(4135) = -6.49$, $p_{\text{FDR}} < 0.0001$). These results do not support H2, but we see that none of these trends are significantly present when we constrain to words that determine the character's competitiveness (H2a). When we consider model suggestions that affect competitiveness, we see marginally more acceptance of competitive fem-coded suggestions than unassertive (H2b; $t(426) = -2.159$, $p_{\text{FDR}} \approx 0.0999$).

Overall, the results in this scenario were mixed. We find a relatively low overall rate of clearly specifying the given student as competitive or unassertive which may have skewed the results. This may be due to poor scenario design where participants decided to focus on topics other than leadership or competitiveness, or it could be that in the classroom settings, participants held weaker internal stereotypes about gender and competitiveness than Cao et al. [16]'s held about more general settings.

A.3 Scenario: Teachers

In this scenario, participants write about a teacher with a given gender where the model attempts to suggest the teacher's personality as "likeable" or "repellent" (see examples in Table 5). According to Cao et al. [16], American annotators view men as comparatively less likeable than women.

A.3.1 Effects on Likeability Disaggregated by Gender. Here, we discuss how suggestions affect the likeability stance of "Mrs. Brown" and "Mr. Brown" (Table 4).

At the Story Level. Considering overall stories, we see that regardless of the teacher's presumed gender, participants made the teacher likeable significantly more often than repellent when not given suggestions (fem-coded teacher: $t(68) = -4.747$, $p_{\text{FDR}} \approx 0.0001$; masc-coded teacher: $t(74) = -3.195$, $p_{\text{FDR}} \approx 0.0110$). These trends continue to be significant regardless of suggestion type, including repellent suggestions (likeable fem-coded: $t(162) = -12.726$, $p_{\text{FDR}} < 0.0001$; repellent fem-coded: $t(150) = -3.886$, $p_{\text{FDR}} \approx 0.0011$; likeable masc-coded: $t(184) = -8.86$, $p_{\text{FDR}} < 0.0001$; repellent masc-coded: $t(166) = -3.774$, $p_{\text{FDR}} \approx 0.0016$). In other words, participants preferred to make the teacher likeable, regardless of the presence or type of suggestions and regardless of the teacher's gender.

We also find that "Mr. Brown" is written as likable marginally more often with likeable suggestions than without suggestions ($t(129) = -2.489$, $p_{\text{FDR}} \approx 0.0516$). However, perhaps due to a higher base rate of "Mrs. Brown" being written as likeable, we see no such increase in likeability with added likeable suggestions for "Mrs. Brown" ($t(115) = -1.403$, $p_{\text{FDR}} \approx 0.3291$). In other words participants may have a stronger default preference for "Mrs. Brown" being likeable, leading to a more limited effect of likeable suggestions. This is in line with Cao et al. [16]'s findings about gender-likeability stereotypes in humans. And these findings provide some support for H1 in that the proportion of likeable "Mrs. Brown"s (pro-stereotypical) with no suggestions is more similar to the proportion of likeable "Mrs. Brown"s with likeable suggestions than the the proportion of likeable "Mr. Brown"s (anti-stereotypical) with no suggestions is to the proportion of likeable "Mr. Brown"s with likeable suggestions.

At the Word Level. When we consider the rate of acceptances of model suggestions, we see significantly less acceptance of model suggestions in the condition where "Mrs. Brown" is suggested to be repellent over likeable ($t(3591) = 2.618$, $p_{\text{FDR}} \approx 0.0351$), while we see a significant effect in the opposite direction for "Mr. Brown" ($t(3892) = -3.162$, $p_{\text{FDR}} \approx 0.0086$). We also see significantly less acceptance in the "Mr. Brown" is likeable condition than for "Mrs. Brown" ($t(4013) = -5.922$, $p_{\text{FDR}} < 0.0001$). These results support H2, as we can see that more pro-stereotypical conditions (e.g., suggesting a fem-coded character is likeable) lead to more acceptance of suggestions.

However, the effects are quite different when we only consider words that determine likeability. Here we see a trend of participants accepting more "likeable" suggestions over "repellent" for either gender, though the effect is only significant for "Mr. Brown" (masc-coded: $t(179) = 2.673$, $p_{\text{FDR}} \approx 0.0331$; fem-coded: $t(157) = 2.068$, $p_{\text{FDR}} \approx 0.1190$). Under H2a, we would have expected to see less acceptance of "likeable" suggestions for the masc-coded "Mr. Brown". When we consider the acceptance rate of model suggestions, we see higher rates of acceptance of "likeable" suggestions for either gender, but in this case, it is only significant for "Mrs. Brown" (masc-coded: $t(438) = -2.446$, $p_{\text{FDR}} \approx 0.0533$, $p_{\text{FDR}} \approx 0.0331$; fem-coded: $t(383) = -4.106$, $p_{\text{FDR}} \approx 0.0004$), supporting H2b.

These results show that participants may have preferred suggestions of teachers of any gender being likeable. This seems reasonable as at the story-level likeable teachers were generally preferred even without suggestions. While these results do not match our overall hypotheses about reliance under pairs of gender and likeability, they may suggest that participants' stereotypes about teachers being likeable people were stronger than their stereotypes about people of different genders being likeable.

A.4 Scenario: Town Hall

In this scenario, the story prefix includes a town hall participant with a traditionally feminine ("Rebecca") or masculine ("Thomas") name. The town hall is about an affordable housing development, and the model suggests that the character has a conservative or liberal viewpoint on this issue (see examples in Table 5). According to Cao et al. [16], American annotators view men as comparatively more conservative than women.

A.4.1 Effects on Political Stance Disaggregated by Gender. Here, we analyze how suggestions affect the political stance of "Rebecca" and "Thomas" (Table 4).

At the Story Level. At the story-level, we first compare the no suggestions conditions to their corresponding liberal and conservative suggestions conditions. We see that adding conservative suggestions decreases the number of liberal characters. This trend is significant for "Thomas" ($t(123) = 2.682$, $p_{\text{FDR}} \approx 0.0332$) and marginally significant for "Rebecca" ($t(112) = 2.337$, $p_{\text{FDR}} \approx 0.0707$). We also see a marginal trend of "Thomas" being made conservative more often with conservative suggestions than without suggestions ($t(123) = -2.334$, $p_{\text{FDR}} \approx 0.0707$), with no such trend in the same setting for "Rebecca" ($t(112) = -0.984$, $p_{\text{FDR}} \approx 0.5407$). We see here that suggestions tend to successfully encourage participants to make characters liberal or conservative, but they are less successful in making "Rebecca" conservative, perhaps suggesting that participants have a harder time accepting suggestions of a fem-coded character being conservative.

We also compare rates of making characters liberal vs conservative within suggestion types. Without suggestions, we see no significant difference between making characters of any gender liberal or conservative (Rebecca: $t(66) = 1.4$, $p_{\text{FDR}} \approx 0.3334$; Thomas: $t(78) = 1.686$, $p_{\text{FDR}} \approx 0.2230$). We see "Thomas" is made liberal or conservative significantly more often depending on the direction of suggestions (conservative: $t(168) = -3.728$, $p_{\text{FDR}} \approx 0.0018$; liberal: $t(166) = 3.315$, $p_{\text{FDR}} \approx 0.0066$). For "Rebecca", we see significantly more liberal stories when they are suggested ($t(168) = 4.575$, $p_{\text{FDR}} \approx 0.0001$), but the increase in conservative stories when they are suggested is not significant ($t(158) = -1.988$, $p_{\text{FDR}} \approx 0.1364$). This again shows that participants may have a harder time accepting suggestions of a fem-coded character being conservative, which is in agreement with [16]'s findings about human perceptions of the political stance of women and supporting H1.

At the Word Level. We generally don't see significant trends at the word level. We do see that participants accepted suggestions significantly more often in the condition where "Rebecca" is suggested to be liberal over "Thomas" (H2b; $t(3570) = -2.725$, $p_{\text{FDR}} \approx 0.0276$). This begins to suggest that participants are more comfortable with fem-coded characters being written as liberal than masc-coded ones, but this trend is not significant when we consider only words that specify the character's stance.

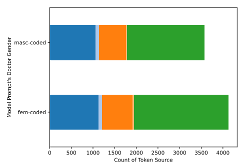
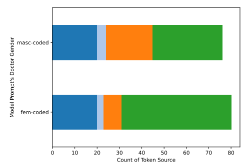
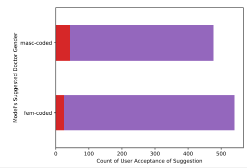
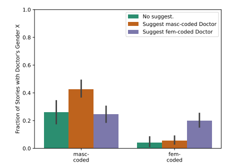
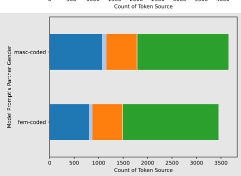
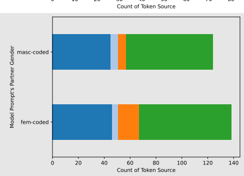
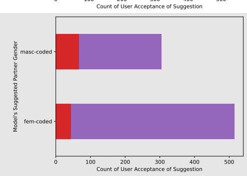
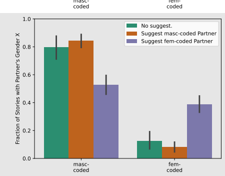
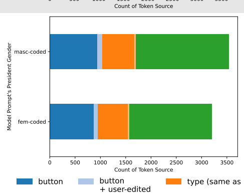
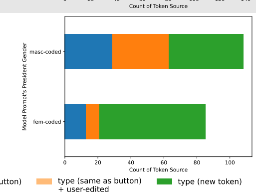
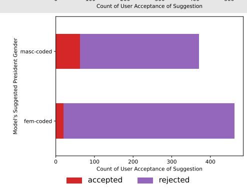
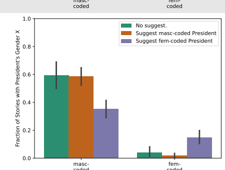
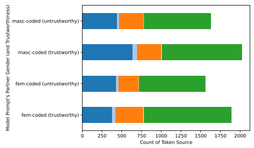
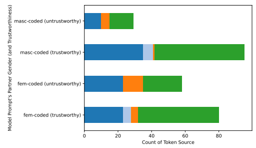
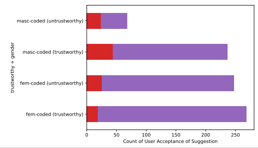
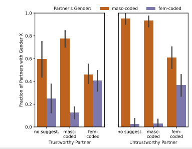
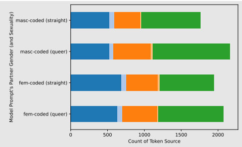
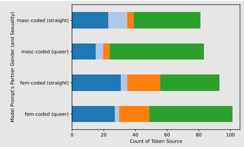
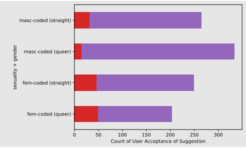
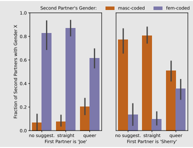
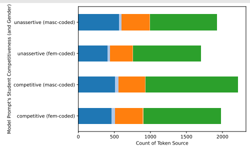
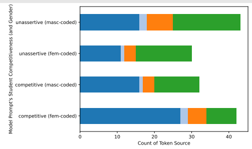
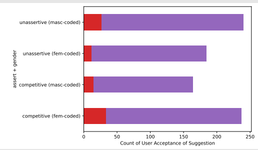
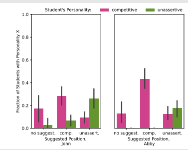
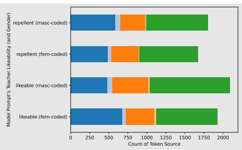
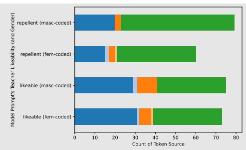
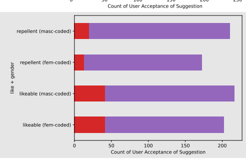
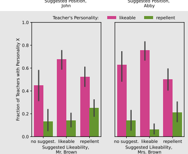
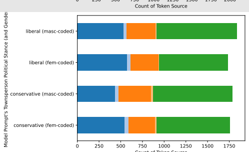
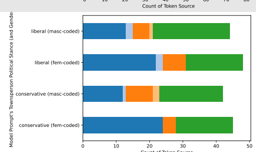
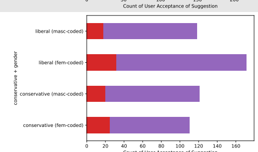
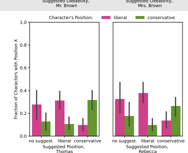
Scenario	Word Reliance	Gender-Defining Word Reliance	Gender-Defining Word Rejection Rate	Story-Level Gender Freq.
Doctor				
Detectives				
President				

Table 3: Acceptance of Gender suggestions (without considering secondary axes)

Scenario	Word Reliance	Axis-Defining Word Reliance	Axis-Defining Word Rejection Rate	Story-Level Axis Freq.
Detectives				
Wedding				
Student				
Teacher				
Town Hall				

button; button + user-edited; type (same as button); type (same as button) + user-edited; type (new token)

accepted; rejected

Table 4: Acceptance of suggestions disaggregated by secondary axis

Scenario	Attribute	Story
Doctor	confident	I went to the doctor yesterday. The doctor said that *I had pneumonia and the stage was moderate but he had to prescribe some medication for me with some reassurance that i will feel better if i take my medication regularly.*
Doctor	unconfident	I went to the doctor yesterday. The doctor said that *she needed me to come back for another test tomorrow. The doctor's voice trembled as she explained that my results showed a high level of uncertainty.*
Doctor	masc-coded	I went to the doctor yesterday. The doctor said that *I have cancer he said i am luck that they caught it early and they will be able to treat it so that it doesn't progress.*
Doctor	fem-coded	I went to the doctor yesterday. The doctor said that *she would run additional tests to confirm the unpleasant results from the insulin levels to be true.*
President	benevolent	In the first 100 days, the new president was determined to focus *on rebuilding the economy, implementing comprehensive heathcare reform, and strengthening international alliances through diplomatic initiatives.*
President	threatening	In the first 100 days, the new president was determined to focus *getting revenge on all those people who said he was a fascist by rounding them all up and having them shot.*
President	masc-coded	In the first 100 days, the new president was determined to focus *on making sure that the many proposals he suggested during his long campaign were fulfilled to please the public*
President	fem-coded	In the first 100 days, the new president was determined to focus *her attention on raising the amoung of housing allowance that incoming refugees can apply for. Despite fierce opposition from the conservative party.*
Wedding	queer (masc-coded 2nd partner)	Joe was happy to be marrying *the most ingelligent man he had ever met even though they had not had the easiest relationship the past year.*
Wedding	straight (masc-coded 2nd partner)	Sherry was happy to be marrying *the most handsome man in the whole wide world. She was sure that her heart belonged only to Peter because he was everything she wanted.*
Student	competitive	In today's class, we were assigned project groups. John*, being the competitive leader that he is, immediately assigned specific duties to his group members with a stern gaze*
Student	unassertive	In today's class, we were assigned project groups. Abby *felt hesitant to introduce herself to our other teammates due to her shyness, but with a nudge and encouraging smile I brought her out of her shell.*
Detectives	masc-coded	In the gritty world of detective work, Detective John Wilson knew he could always count on his partner *Chris, who had been his trusty sidekick since they were kids, they have always been able to solve cases which seemed impossible to crack. Together they are the dynamic duo.*
Detectives	fem-coded	In the gritty world of detective work, Detective John Wilson's partner wasn't to be trusted. *He knew she was corrupt, and had been for a while. John just needed to prove it. Gathering evidence would be difficult, but he knew he had to persevere.*
Teacher	likeable	When I was in school, Mr. Brown was *my science teachers. He was very much witty and always try to make their lessons attractive with amazing experiments in such unique ways.*
Teacher	repellent	When I was in school, Mr. Brown was *our least favorite teacher, he was really hated. He would often make us do multiple assignments a week.*
Town Hall	conservative	The town hall meeting about the new affordable housing development got very rowdy. Rebecca said that *it wojld would cheapen the look of their posh neighborhood. It was not ok for there to be cheap looking homes in their area.*
Town Hall	liberal	The town hall meeting about the new affordable housing development got very rowdy. Rebecca said that *she firmly supported the project and emphasized its potential to bring stability to the families in the community.*

Table 5: Example co-written stories displaying suggested features. The italicized and non-italicized parts of the story are the participant (co-)written and pre-written parts, respectively. Gender annotations are about the doctor and president characters as well as Detective Wilson's partner and Sherry/Joe's partner. The non-gender-related annotations are always about the non-speaker character introduced in the story prefix (e.g., the doctor, Mr./Mrs. Brown, etc.).

A.5 Scenario Statistical Tests

Here, we include all statistical test and p-value tables for scenario-level experiments not otherwise included in the appendix. This includes overall story gender and other attribute rates (Table 6, Table 7, Table 8, and Table 9), overall word-level reliance (Table 11), reliance rates for attribute-specifying words (Table 12), and rejection rates for attribute-specifying suggestions (Table 13).

Scenario	Suggestions	Measured Attr	t	p_{FDR}	sig
Doctor	F vs NS	F	$t(246) = 3.219$	0.0081	*
	F vs NS	M	$t(246) = -0.24$	0.9019	
	M vs NS	F	$t(230) = 0.493$	0.7882	
	M vs NS	M	$t(230) = 2.467$	0.0521	
	F	F vs M	$t(348) = -1.026$	0.5207	
	M	F vs M	$t(316) = -8.542$	0.0000	*
	NS	F vs M	$t(144) = -3.862$	0.0012	*
President	F vs NS	F	$t(234) = 2.429$	0.0565	
	F vs NS	M	$t(234) = -3.576$	0.0029	*
	M vs NS	F	$t(248) = -1.106$	0.4750	
	M vs NS	M	$t(248) = -0.137$	0.9488	
	F	F vs M	$t(322) = -4.343$	0.0002	*
	M	F vs M	$t(350) = -14.755$	0.0000	*
	NS	F vs M	$t(146) = -8.947$	0.0000	*
Detectives	F vs NS	F	$t(234) = 4.191$	0.0003	*
	F vs NS	M	$t(234) = -4.0$	0.0007	*
	M vs NS	F	$t(242) = -1.012$	0.5254	
	M vs NS	M	$t(242) = 0.899$	0.5681	
	F	F vs M	$t(324) = -2.575$	0.0408	*
	M	F vs M	$t(340) = -21.727$	0.0000	*
	NS	F vs M	$t(144) = -10.934$	0.0000	*

Table 6: Story-level gender rate comparisons when not considering secondary attributes

First partner's gender	Suggested sexuality of pairing	Measured gender of second partner	t	p_{FDR}	sig
M	NS vs straight	M	$t(104) = -0.154$	0.9479	
M	NS vs straight	F	$t(104) = -0.556$	0.7514	
M	NS vs queer	M	$t(115) = -1.688$	0.2218	
M	NS vs queer	F	$t(115) = 2.143$	0.1071	
M	NS	M vs F	$t(56) = -8.825$	0.0000	*
M	straight	M vs F	$t(152) = -16.063$	0.0000	*
M	queer	M vs F	$t(174) = -6.035$	0.0000	*
F	NS vs straight	M	$t(124) = -0.422$	0.8234	
F	NS vs straight	F	$t(124) = 0.657$	0.7019	
F	NS vs queer	M	$t(132) = 2.973$	0.0172	*
F	NS vs queer	F	$t(132) = -2.695$	0.0328	*
F	NS	M vs F	$t(86) = 7.704$	0.0000	*
F	straight	M vs F	$t(162) = 12.859$	0.0000	*
F	queer	M vs F	$t(178) = 2.12$	0.1097	

Table 7: Comparison of gender rates in Wedding scenario under varied suggestions and gender of initial partner

Scenario	Suggestions	Measured Attr	t	p_{FDR}	sig
Student	NS vs M competitive	competitive	$t(121) = -1.297$	0.3747	
	NS vs M competitive	unassertive	$t(121) = -0.851$	0.5921	
	NS vs M unassertive	competitive	$t(117) = 1.172$	0.4353	
	NS vs M unassertive	unassertive	$t(117) = -3.024$	0.0157	*
	NS vs F competitive	competitive	$t(125) = -3.464$	0.0044	*
	NS vs F unassertive	competitive	$t(116) = 0.025$	0.9831	
	NS vs F unassertive	unassertive	$t(116) = -2.874$	0.0210	*
	M NS	competitive vs unassertive	$t(68) = 2.022$	0.1353	
	M competitive	competitive vs unassertive	$t(174) = 3.898$	0.0010	*
	M unassertive	competitive vs unassertive	$t(166) = -2.872$	0.0207	*
	F NS	competitive vs unassertive	$t(76) = 2.364$	0.0704	
	F competitive	competitive vs unassertive	$t(174) = 8.131$	0.0000	*
	F unassertive	competitive vs unassertive	$t(156) = -0.883$	0.5725	
Teacher	NS vs M likeable	repellent	$t(129) = -0.123$	0.9523	
	NS vs M likeable	likeable	$t(129) = -2.489$	0.0516	
	NS vs M repellent	repellent	$t(120) = -1.48$	0.2950	
	NS vs M repellent	likeable	$t(120) = -0.778$	0.6290	
	NS vs F likeable	repellent	$t(115) = 1.451$	0.3065	
	NS vs F likeable	likeable	$t(115) = -1.403$	0.3291	
	NS vs F repellent	repellent	$t(109) = -0.841$	0.5980	
	NS vs F repellent	likeable	$t(109) = 1.261$	0.3930	
	M NS	repellent vs likeable	$t(74) = -3.195$	0.0110	*
	M likeable	repellent vs likeable	$t(184) = -8.86$	0.0000	*
	M repellent	repellent vs likeable	$t(166) = -3.774$	0.0016	*
	F NS	repellent vs likeable	$t(68) = -4.747$	0.0001	*
	F likeable	repellent vs likeable	$t(162) = -12.726$	0.0000	*
	F repellent	repellent vs likeable	$t(150) = -3.886$	0.0011	*
Town Hall	NS vs M conservative	conservative	$t(123) = -2.334$	0.0707	
	NS vs M conservative	liberal	$t(123) = 2.682$	0.0332	*
	NS vs M liberal	conservative	$t(122) = 0.291$	0.8721	
	NS vs M liberal	liberal	$t(122) = -0.39$	0.8283	
	NS vs F conservative	conservative	$t(112) = -0.984$	0.5407	
	NS vs F conservative	liberal	$t(112) = 2.337$	0.0707	
	NS vs F liberal	conservative	$t(117) = 1.257$	0.3930	
	NS vs F liberal	liberal	$t(117) = -0.539$	0.7605	
	M NS	conservative vs liberal	$t(78) = -1.686$	0.2230	
	M conservative	conservative vs liberal	$t(168) = 3.728$	0.0018	*
	M liberal	conservative vs liberal	$t(166) = -3.315$	0.0066	*
	F NS	conservative vs liberal	$t(66) = -1.4$	0.3334	
	F conservative	conservative vs liberal	$t(158) = 1.988$	0.1364	
	F liberal	conservative vs liberal	$t(168) = -4.575$	0.0001	*

Table 8: Story-level secondary attribute rate comparisons disaggregated by gender

Trustworthiness	Suggested Gender	Measured Gender	t	p_{FDR}	sig
trustworthy	NS vs M	M	$t(127) = -1.995$	0.1364	
trustworthy	NS vs F	M	$t(119) = 1.289$	0.3776	
trustworthy	NS vs M	F	$t(127) = 1.718$	0.2136	
trustworthy	NS vs F	F	$t(119) = -1.561$	0.2634	
untrustworthy	NS vs M	M	$t(113) = 0.4$	0.8283	
untrustworthy	NS vs F	M	$t(113) = 4.238$	0.0004	*
untrustworthy	NS vs M	F	$t(113) = -0.084$	0.9634	
untrustworthy	NS vs F	F	$t(113) = -4.367$	0.0003	*
trustworhty vs untrustworthy	NS	M	$t(71) = -4.117$	0.0008	*
trustworhty vs untrustworthy	M	M	$t(169) = -2.881$	0.0206	*
trustworhty vs untrustworthy	F	M	$t(161) = -1.886$	0.1667	
trustworhty vs untrustworthy	NS	F	$t(71) = 3.051$	0.0161	*
trustworhty vs untrustworthy	M	F	$t(169) = 2.307$	0.0735	
trustworhty vs untrustworthy	F	F	$t(161) = 0.515$	0.7766	

Table 9: Story-level gender rate comparisons disaggregated by trustworthiness in the DETECTIVES scenario

Suggests Compared	Measured Attr	t	p_{FDR}	sig
M benevolent vs NS	F benevolent	$t(166) = 0.773$	0.6294	
	M	$t(166) = -0.375$	0.8283	
	M benevolent	$t(166) = 0.492$	0.7882	
	M threatening	$t(166) = 0.38$	0.8283	
	benevolent	$t(166) = -0.102$	0.9609	
	threatening	$t(166) = 0.887$	0.5719	
	None	$t(166) = -0.429$	0.8201	
M threatening vs NS	M	$t(154) = 0.142$	0.9488	
	M benevolent	$t(154) = -0.562$	0.7480	
	M threatening	$t(154) = -0.569$	0.7447	
	benevolent	$t(154) = 0.45$	0.8085	
	threatening	$t(154) = 1.666$	0.2262	
	None	$t(154) = 0.357$	0.8353	
F benevolent vs NS	F benevolent	$t(154) = 3.068$	0.0134	*
	M	$t(154) = -2.145$	0.1057	
	M benevolent	$t(154) = -2.337$	0.0704	
	M threatening	$t(154) = 0.247$	0.8997	
	benevolent	$t(154) = 0.93$	0.5574	
	threatening	$t(154) = 0.95$	0.5497	
	None	$t(154) = 1.677$	0.2230	
F threatening vs NS	F	$t(152) = -0.079$	0.9634	
	F benevolent	$t(152) = 2.526$	0.0465	*
	F threatening	$t(152) = 0.962$	0.5495	
	M	$t(152) = -2.068$	0.1190	
	M benevolent	$t(152) = -0.835$	0.5980	
	M threatening	$t(152) = -0.541$	0.7605	
	benevolent	$t(152) = 0.878$	0.5744	
	None	$t(152) = 0.654$	0.7019	

Table 10: Story-level attribute rates considering gender and benevolence jointly in the PRESIDENT scenario

Scenario	Comparison	t	p_{FDR}	sig
Doctor	F vs M	$t(7692) = 2.258$	0.0784	
	F unconf vs F conf	$t(4121) = -2.645$	0.0331	$*$
	F unconf vs M unconf	$t(4040) = -0.195$	0.9213	
	M unconf vs M conf	$t(3569) = 1.172$	0.4337	
	F conf vs M conf	$t(3650) = 3.488$	0.0031	$*$
President	F vs M	$t(6734) = -0.979$	0.5407	
	F threatening vs F benevolent	$t(3198) = 0.095$	0.9617	
	F threatening vs M threatening	$t(3187) = 0.311$	0.8658	
	M threatening vs M benevolent	$t(3534) = -1.784$	0.1927	
	F benevolent vs M benevolent	$t(3545) = -1.56$	0.2631	
Wedding	F queer vs F straight	$t(4010) = 2.381$	0.0610	
	F queer vs M queer	$t(4230) = -4.019$	0.0005	$*$
	M queer vs M straight	$t(3919) = 1.712$	0.2120	
	F straight vs M straight	$t(3699) = -4.373$	0.0001	$*$
Student	F unassertive vs F competitive	$t(3668) = 0.257$	0.8959	
	F unassertive vs M unassertive	$t(3615) = 4.453$	0.0001	$*$
	M unassertive vs M competitive	$t(4135) = -6.49$	0.0000	$*$
	F competitive vs M competitive	$t(4188) = -2.0$	0.1325	
Detectives	F vs M	$t(7091) = 4.724$	0.0000	$*$
	F untrustworthy vs F trustworthy	$t(3443) = -3.168$	0.0085	$*$
	F untrustworthy vs M untrustworthy	$t(3182) = 1.292$	0.3747	
	M untrustworthy vs M trustworthy	$t(3646) = 0.396$	0.8283	
	F trustworthy vs M trustworthy	$t(3907) = 5.235$	0.0000	$*$
Teacher	F repellent vs F likeable	$t(3591) = 2.618$	0.0351	$*$
	F repellent vs M repellent	$t(3470) = 0.072$	0.9634	
	M repellent vs M likeable	$t(3892) = -3.162$	0.0086	$*$
	F likeable vs M likeable	$t(4013) = -5.922$	0.0000	$*$
Town Hall	F liberal vs F conservative	$t(3484) = -1.763$	0.1978	
	F liberal vs M liberal	$t(3570) = -2.725$	0.0276	$*$
	M liberal vs M conservative	$t(3620) = -0.9$	0.5681	
	F conservative vs M conservative	$t(3534) = -1.825$	0.1819	

Table 11: Tests of overall word-level reliance. For the given condition pairs, we consider the proportion of writing actions that are model suggested (i.e., the participant uses a suggestion button or manually types an identical word) vs participant supplied/edited (i.e., the participant types a non-suggested word or edits a model suggestion)

Scenario	Comparison	Attr	t	p_{FDR}	sig
Doctor	F vs M	gender	$t(159) = 2.734$	0.0294	$*$
	F unconf vs F conf	gender	$t(83) = 2.128$	0.1117	
	F unconf vs M unconf	gender	$t(81) = 2.93$	0.0205	$*$
	M unconf vs M conf	gender	$t(74) = 0.591$	0.7369	
	F conf vs M conf	gender	$t(76) = 1.339$	0.3565	
	F unconf vs F conf	other	$t(142) = 0.594$	0.7356	
	F unconf vs M unconf	other	$t(163) = 1.918$	0.1575	
	M unconf vs M conf	other	$t(162) = -1.219$	0.4140	
	F conf vs M conf	other	$t(141) = 0.036$	0.9765	
President	F vs M	gender	$t(203) = 5.269$	0.0000	$*$
	F threatening vs F benevolent	gender	$t(88) = 0.486$	0.7920	
	F threatening vs M threatening	gender	$t(91) = 3.309$	0.0076	$*$
	M threatening vs M benevolent	gender	$t(113) = 0.912$	0.5668	
	F benevolent vs M benevolent	gender	$t(110) = 3.962$	0.0010	$*$
	F threatening vs F benevolent	other	$t(158) = 0.515$	0.7766	
	F threatening vs M threatening	other	$t(150) = -1.019$	0.5254	
	M threatening vs M benevolent	other	$t(155) = 0.78$	0.6290	
	F benevolent vs M benevolent	other	$t(163) = -0.784$	0.6277	
Wedding	F queer vs F straight	gender	$t(215) = 1.561$	0.2634	
	F queer vs M queer	gender	$t(206) = -2.566$	0.0424	$*$
	M queer vs M straight	gender	$t(186) = 0.479$	0.7947	
	F straight vs M straight	gender	$t(195) = -3.554$	0.0031	$*$
Detectives	F untrustworthy vs F trustworthy	gender	$t(161) = -3.298$	0.0069	$*$
	F untrustworthy vs M untrustworthy	gender	$t(108) = -0.168$	0.9425	
	M untrustworthy vs M trustworthy	gender	$t(139) = -1.864$	0.1746	
	F trustworthy vs M trustworthy	gender	$t(192) = 0.871$	0.5775	
	F vs M	gender	$t(302) = -0.069$	0.9634	
Student	F unassertive vs F competitive	other	$t(87) = 1.854$	0.1806	
	F unassertive vs M unassertive	other	$t(96) = -0.448$	0.8085	
	M unassertive vs M competitive	other	$t(103) = 1.041$	0.5141	
	F competitive vs M competitive	other	$t(94) = -1.372$	0.3400	
Teacher	F repellent vs F likeable	other	$t(157) = 2.068$	0.1190	
	F repellent vs M repellent	other	$t(173) = -0.048$	0.9732	
	M repellent vs M likeable	other	$t(179) = 2.673$	0.0331	$*$
	F likeable vs M likeable	other	$t(163) = 0.383$	0.8283	
Town Hall	F liberal vs F conservative	other	$t(110) = -0.216$	0.9088	
	F liberal vs M liberal	other	$t(103) = -1.453$	0.3065	
	M liberal vs M conservative	other	$t(104) = -0.302$	0.8715	
	F conservative vs M conservative	other	$t(111) = -1.606$	0.2488	

Table 12: Like in Table 11, these tests compare word-level reliance rates. Here, we constrain the analysis to only consider words that specify the given attribute: gender or another attribute like likeability, confidence, etc.

Scenario	Comparison	t	p_{FDR}	sig
Doctor	F vs M	$t(1014) = 2.926$	0.0172	*
President	F vs M	$t(836) = 6.362$	0.0000	*
Wedding	F queer vs F straight	$t(458) = 1.68$	0.2218	
	F queer vs M queer	$t(546) = 6.786$	0.0000	*
	M queer vs M straight	$t(607) = -2.901$	0.0187	*
	F straight vs M straight	$t(519) = 1.988$	0.1353	
Student	F unassertive vs F competitive	$t(426) = -2.159$	0.0999	
	F unassertive vs M unassertive	$t(432) = -1.601$	0.2484	
	M unassertive vs M competitive	$t(411) = 0.798$	0.6177	
	F competitive vs M competitive	$t(405) = 1.346$	0.3494	
Detectives	F vs M	$t(834) = 6.729$	0.0000	*
	F untrustworthy vs F trustworthy	$t(515) = 1.364$	0.3400	
	F untrustworthy vs M untrustworthy	$t(325) = -7.066$	0.0000	*
	M untrustworthy vs M trustworthy	$t(317) = 4.384$	0.0002	*
	F trustworthy vs M trustworthy	$t(507) = -4.224$	0.0003	*
Teacher	F repellent vs F likeable	$t(383) = -4.106$	0.0004	*
	F repellent vs M repellent	$t(390) = -1.329$	0.3565	
	M repellent vs M likeable	$t(438) = -2.446$	0.0533	
	F likeable vs M likeable	$t(431) = 0.63$	0.7191	
Town Hall	F liberal vs F conservative	$t(286) = -1.735$	0.2058	
	F liberal vs M liberal	$t(290) = 0.296$	0.8721	
	M liberal vs M conservative	$t(244) = -0.373$	0.8283	
	F conservative vs M conservative	$t(240) = 1.503$	0.2830	

Table 13: Tests of acceptance rates of attribute-defining suggestions. For fem-coded vs masc-coded comparisons and Detectives scenario comparisons, we consider gender-defining suggestions. For the remainder, we consider suggestions that specify the second attribute (assertiveness, likeability, etc)

B ADDITIONAL ANALYSIS

B.1 Suggestion stereotypes and time to make decisions

As we discussed in subsubsection 6.5.1, we consider how long it takes participants to make word-level decisions based on suggestion type. We saw that participants took less to make decisions about trustworthy masc-coded detectives suggestions, suggesting that this is an unsurprising set of attributes for a detective. In this section, we provide more detail about how these comparisons were made and the findings on more scenarios (Table 14).

For a given scenario, we start with the set of words suggested by the model that would specify a given attribute (regardless of whether the participant accepted it) and the time to make their decision. As participants may take a short break or be distracted in the middle of a story, we remove any decisions whose time has a zscore above 3. This removed 70 word-level decisions that had an average time of 127 seconds.

In the Student scenario, we find that participants took significantly longer to made decisions for "unassertive" suggestions than "competitive" ones, regardless of gender (fem-coded: $t(422) = 5.792$, $p_{\text{FDR}} < 0.0001$; masc-coded: $t(402) = 6.298$, $p_{\text{FDR}} < 0.0001$). We also see that it took significantly longer to decide to accept "competitive" suggestions when the character in question was fem-coded ($t(405) = -2.833$, $p_{\text{FDR}} \approx 0.0210$). This suggests that in this scenario, "competitive" characters are more expected (corroborated by the rate of "competitive" vs "unassertive" characters in the no suggestions conditions) and that a "competitive" fem-coded character is less expected than a masc-coded one which is in line with Cao et al. [16]'s finding that men are viewed comparatively more competitive than women.

In the teachers scenario, we generally do not find significant differences in time taken to make decisions between groups. However, we do see a significant trend of suggestions that "Mrs. Brown" is a repellent teacher taking longer to decide about than suggestions that she is a likeable teacher ($t(380) = 2.855$, $p_{\text{FDR}} \approx 0.0206$). This potential expectation that fem-coded teachers are likeable is again corroborated by our earlier findings about the rate of choosing "Mrs. Brown" to be likeable without suggestions. While we cannot confirm Cao et al.

Scenario	Comparison	t	p_{FDR}	sig
Doctor	F conf vs F unconf	$t(538) = 0.459$	0.8061	
	M conf vs M unconf	$t(473) = -1.205$	0.4172	
	F conf vs F unconf	$t(288) = -1.766$	0.1978	
	M conf vs M unconf	$t(322) = 0.223$	0.9088	
	M vs F	$t(1013) = 0.088$	0.9634	
President	F benevolent vs F threatening	$t(458) = 0.761$	0.6351	
	M benevolent vs M threatening	$t(373) = 0.597$	0.7350	
	F benevolent vs F threatening	$t(208) = -0.126$	0.9523	
	M benevolent vs M threatening	$t(203) = -1.082$	0.4897	
	M vs F	$t(833) = -1.606$	0.2464	
Wedding	F queer vs F straight	$t(448) = -4.486$	0.0001	*
	M queer vs M straight	$t(602) = -2.063$	0.1190	
	M queer vs F queer	$t(543) = 2.898$	0.0187	*
	M straight vs F straight	$t(507) = -0.607$	0.7324	
Student	F competitive vs F unassertive	$t(422) = -5.792$	0.0000	*
	M competitive vs M unassertive	$t(402) = -6.298$	0.0000	*
	M competitive vs F competitive	$t(405) = -2.833$	0.0210	*
	M unassertive vs F unassertive	$t(419) = -1.513$	0.2780	
Detectives	F untrustworthy vs F trustworthy	$t(496) = 0.077$	0.9634	
	M untrustworthy vs M trustworthy	$t(306) = -3.544$	0.0030	*
	M untrustworthy vs F untrustworthy	$t(310) = -3.48$	0.0036	*
	M trustworthy vs F trustworthy	$t(492) = -0.218$	0.9088	
	M vs F	$t(804) = -1.804$	0.1870	
Teacher	F repellent vs F likeable	$t(380) = 2.855$	0.0206	*
	M repellent vs M likeable	$t(438) = 1.439$	0.3077	
	M repellent vs F repellent	$t(390) = -0.279$	0.8801	
	M likeable vs F likeable	$t(428) = 1.188$	0.4253	
Town Hall	F conservative vs F liberal	$t(284) = 1.655$	0.2273	
	M conservative vs M liberal	$t(243) = 1.053$	0.5059	
	M conservative vs F conservative	$t(240) = -0.389$	0.8283	
	M liberal vs F liberal	$t(287) = -0.068$	0.9634	

Table 14: Test comparing time taken to make word-level decisions with varied story and suggested attributes.

[16]'s finding that women are seen as more likeable than men, our findings in this scenario do agree that fem-coded people are seen as more likeable than repellent.

In the WEDDING scenario, we see that it takes marginally significantly longer to make decisions about a masc-coded queer partner as opposed to a fem-coded one ($t(543) = 2.898$, $p_{\mathrm{FDR}} \approx 0.0187$). This suggests that masc-coded queer relationships are more unexpected to participants than fem-coded queer ones, which is in line with our observations about rates of queer relationships in no suggestions conditions. However, we surprisingly also see that, when the first partner is fem-coded, it took participants significantly longer to decide on suggestions about whether the second partner should be masc-coded vs fem-coded ($t(448) = -4.486$, $p_{\mathrm{FDR}} \approx 0.0001$). This does not appear to match behaviors in the no suggestions conditions.

Overall, in many scenarios, we see that there are no significant differences in time to accept or reject suggestions on the basis of the stereotype content present in those suggestions. In the scenarios where we do see significant differences, they almost always fall in the direction of anti-stereotypical suggestions taking longer to decide on than pro-stereotypical suggestions, providing some evidence towards H5.

B.2 Story Length and Overall Time to Write

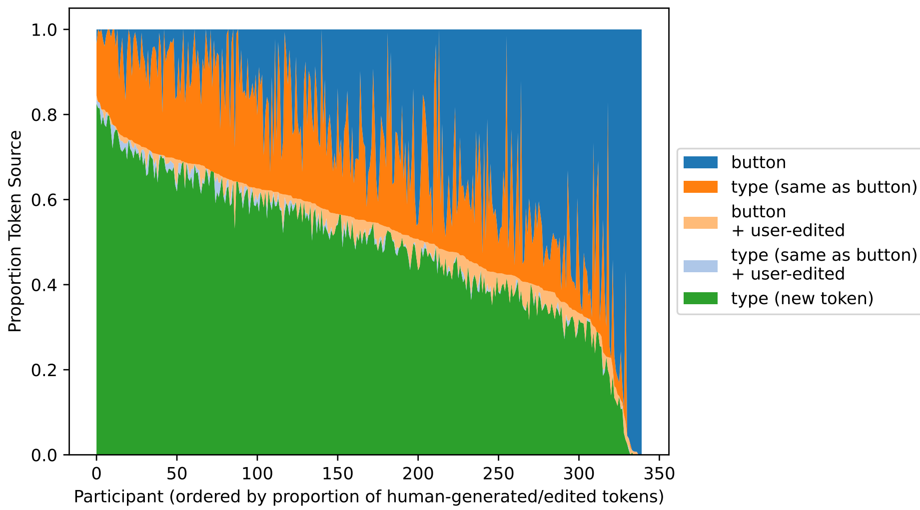


Figure 13: Distribution of word sources per participant

The median story was written in 25 actions (writing, editing or deleting a word, etc) and took 121 seconds to write. When participants were given suggestions, 54.7% of the words in the median participant's stories were newly written or edited by the participant (see distribution in Figure 13).

At the character level, we see that stories written with suggestions were longer than those written without suggestions, but this trend only marginally signifcant ($t(2863) = 2.326$, $p_{\mathrm{FDR}} \approx 0.069$). While these results are not consistent with Arnold et al. [4], we note that in our study design, we set a minimum number of characters to add to the story before continuing which may have affected participants' behavior regarding story length.

B.3 Effect of Individual Differences

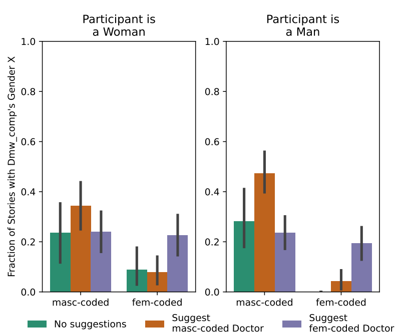
In subsubsection 6.5.2, we considered how participant's views on gendered and competence affected their stories. Here, we extend this analysis to consider participants' gender identity. We consider a similar analysis based on participants' self-reported gender (Figure 14 and Table 16), hypothesizing that participants who self-identify as women may be more likely to write stories about fem-coded characters without suggestions or to accept fem-coded suggestions.

Here we see no significant effects, but we do note some minor trends that point towards participants writing more characters whose genders match their own. For example, we see more masc-coded doctors and detectives from participants who identify as men under masc-coded suggestions than we do from participants who identify as women, and we see more masc-coded presidents from participants who identify as men under fem-coded suggestions than we do from participants who identify as women.

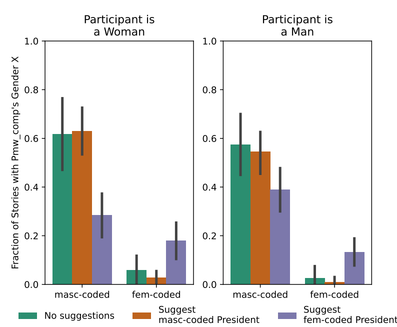
Overall, we find some trends pointing towards participants' stereotypes and gender identities influencing their stories and their acceptance of model suggestions in the expected direction.

Scenario	Suggested Gender	Measured Gender	t	p_{FDR}	sig
Doctor	NS	M	$t(37) = 0.344$	0.8446	
	NS	F	$t(37) = 0.926$	0.5641	
	M	M	$t(88) = -0.238$	0.9019	
	M	F	$t(88) = 0.838$	0.5980	
	F	M	$t(103) = -0.608$	0.7324	
	F	F	$t(103) = 2.279$	0.0802	
President	NS	M	$t(38) = 1.166$	0.4464	
	NS	F	$t(38) = 0.215$	0.9088	
	M	M	$t(100) = 1.539$	0.2728	
	M	F	$t(100) = 1.373$	0.3400	
	F	M	$t(93) = 0.14$	0.9488	
	F	F	$t(93) = -0.011$	0.9910	
Detectives	NS	M	$t(38) = -0.394$	0.8283	
	NS	F	$t(38) = 0.832$	0.6032	
	M	M	$t(95) = -0.378$	0.8283	
	M	F	$t(95) = 0.356$	0.8353	
	F	M	$t(97) = -2.095$	0.1183	
	F	F	$t(97) = 2.742$	0.0304	*

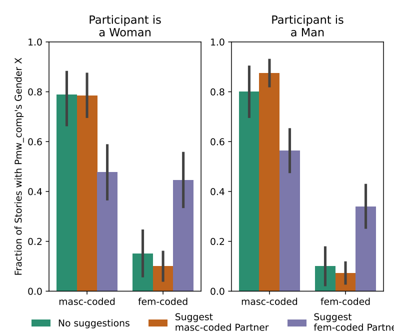
Table 15: Comparison of character genders written with various suggestions for participants who answered that straight women are more competent than straight men vs less.



(a) Doctor



(b) President



(c) Detective

Figure 14: Rates of character gender based on participants' self-reported gender

B.4 Correlation of Participant Stereotypes

In the post-survey, participants were asked about stereotypical beliefs. As we discussed in section 4, participants were asked whether their closest friends believed different groups (gay vs straight and men vs women) were warm, competent, or conservative (Figure 15).

We first confirm that there is a significant positive correlation between warmth and competence ($r(1654) = 0.245$, $p_{\text{FDR}} < 1e^{-23}$). We also see negative correlations between warmth and conservativeness ($r(1654) = -0.466$, $p_{\text{FDR}} < 1e^{-89}$) and competence and conservativeness ($r(1654) = -0.138$, $p_{\text{FDR}} < 1e^{-6}$). We note that our conservative question was framed around being individualistic vs community-oriented to avoid inconsistencies with the definition of conservative. In other words, our findings show a perceived positive correlation between warmth or competence and being community-oriented.

We also observe differences in perceptions of groups on the various axes. We see that straight women and gay men are viewed as more warm than gay women and especially straight men. Straight people were viewed as slightly more competent than their queer counterparts. Straight men were the only group viewed as more conservative (individualistic) than liberal (community-oriented). Gay women were viewed almost neutrally while straight women and gay men were viewed as more liberal (community-oriented).

Scenario	Suggested Gender	Written Gender	t	p_{FDR}	sig
Doctor	NS	M	$t(71) = -0.449$	0.8085	
	NS	F	$t(71) = 1.916$	0.1634	
	M	M	$t(157) = -1.762$	0.1991	
	M	F	$t(157) = 0.961$	0.5495	
	F	M	$t(173) = -0.151$	0.9479	
	F	F	$t(173) = 0.761$	0.6351	
President	NS	M	$t(72) = 0.368$	0.8310	
	NS	F	$t(72) = 0.728$	0.6616	
	M	M	$t(174) = 1.015$	0.5254	
	M	F	$t(174) = 0.89$	0.5714	
	F	M	$t(160) = -1.53$	0.2735	
	F	F	$t(160) = 0.928$	0.5574	
Detectives	NS	M	$t(71) = -0.126$	0.9523	
	NS	F	$t(71) = 0.659$	0.7019	
	M	M	$t(169) = -1.688$	0.2218	
	M	F	$t(169) = 0.717$	0.6651	
	F	M	$t(161) = -1.053$	0.5059	
	F	F	$t(161) = 1.272$	0.3858	

Table 16: Comparison of character genders written with various suggestions for participants who self-identified as women vs men

Scenario	Classification Attribute	Suggestions	t	p_{FDR}	sig
Student	communion	-	$t(72) = 1.479$	0.2978	
	communion	✓	$t(337) = 1.758$	0.1991	
	toxicity	-	$t(72) = 0.955$	0.5497	
	toxicity	✓	$t(337) = -0.672$	0.6982	
	sentiment	-	$t(72) = -0.066$	0.9634	
	sentiment	✓	$t(337) = 2.516$	0.0462	*
Teacher	communion	-	$t(71) = 2.424$	0.0625	
	communion	✓	$t(333) = 3.44$	0.0040	*
	toxicity	-	$t(71) = 1.009$	0.5296	
	toxicity	✓	$t(333) = -1.147$	0.4464	
	sentiment	-	$t(71) = 0.446$	0.8085	
	sentiment	✓	$t(333) = 0.957$	0.5495	
Town Hall	communion	-	$t(72) = 1.808$	0.1927	
	communion	✓	$t(332) = 3.852$	0.0010	*
	toxicity	-	$t(72) = 1.083$	0.4908	
	toxicity	✓	$t(332) = 1.201$	0.4184	
	sentiment	-	$t(72) = -0.408$	0.8283	
	sentiment	✓	$t(332) = 0.381$	0.8283	

Table 17: Comparison of attribute scores between character genders in stories written with and without predictive text suggestions.

B.5 Suggestions and Toxicity, Sentiment, and Character Agency

Beyond the attributes we designed the predictive text model to suggest to the user, we also consider off-the-shelf classification of stories toxicity[10], sentiment [13], and character agency [66]. Here, we consider writing scenarios where the character of interest's gender is pre-specified before the participants begin writing. We compare between classifier output for stories where characters are masc-coded vs fem-coded when participants are or are not provided with suggestions (Figure 16 and Table 17).

[10]https://huggingface.co/martin-ha/toxic-comment-model

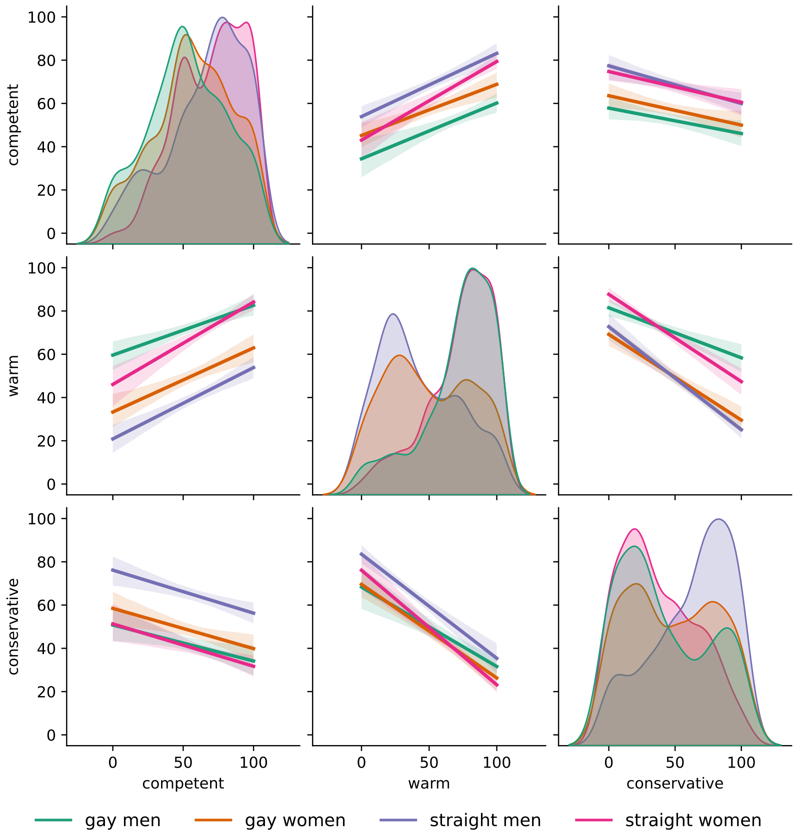
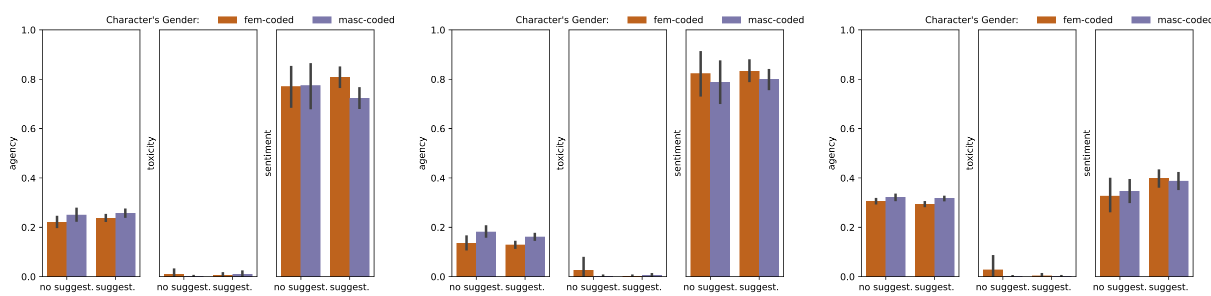


Figure 15: Distributions of and correlations between human stereotypes for various groups



(a) Student **(b) Teacher** **(c) Town Hall**

Figure 16: Agency, toxicity, and sentiment ratings in stories. For each attribute, we break down stories into those written with and without suggestions and those written about masc-coded characters (purple) and fem-coded (orange).

Toxicity. We find that the generated stories tend not to be explicitly toxic, with the story with the highest toxicity rating describing that "Rebecca said that our ideas were really stupid and bland. We got very angry and shouted at her. It was very unprofessional." Toxicity rates were uniformly low across genders.

Sentiment. We find no significant differences in sentiment between suggestion conditions in the Teacher and Town Hall scenarios regardless of suggestions and in the Student scenario without suggestions. However, in the Student scenario, we see significantly lower sentiment ratings for stories written about "John" than "Abby" when the stories are written with suggestions ($t(337) = 2.516$, $p_{\text{FDR}} \approx 0.0462$). This means that the predictive text model may have had a bias towards suggesting more positive continuations about Abby than John or that

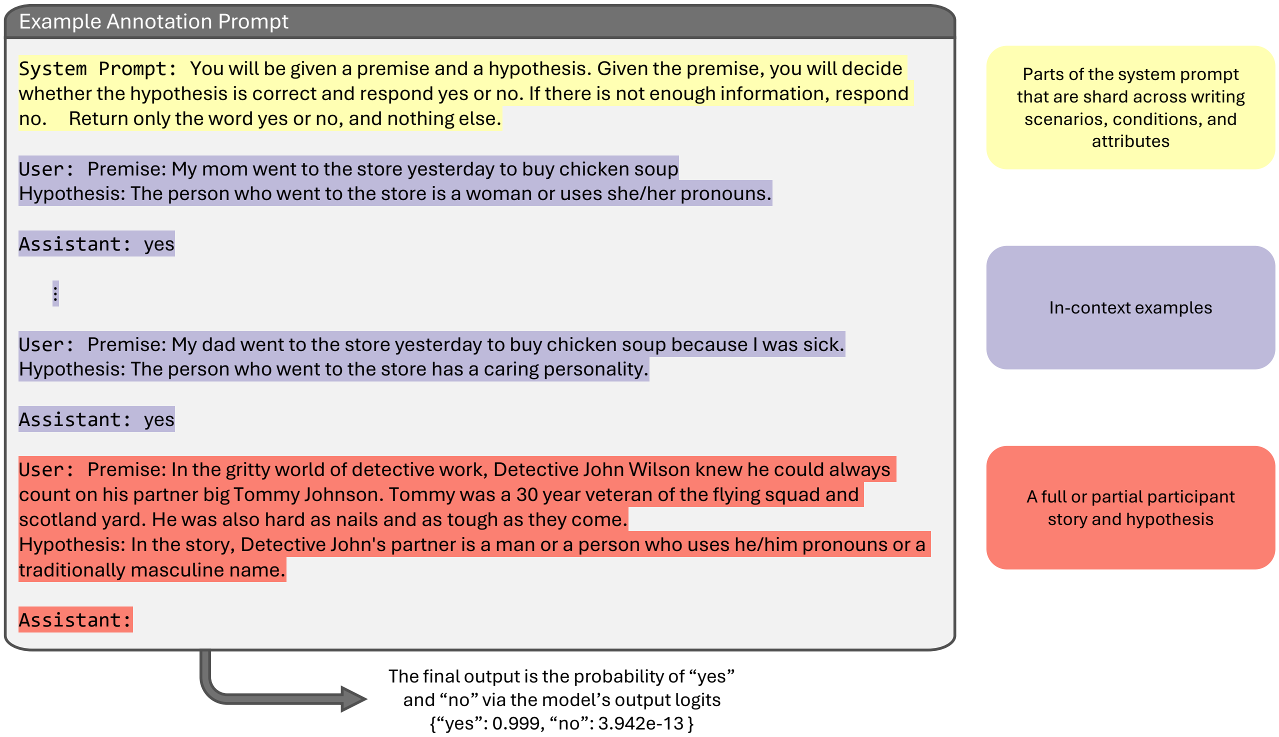


Figure 17: Example prompt for annotating whether the detective’s partner is masc-coded.

participants were more likely to accept positive suggestions about Abby. Regardless of mechanism, we see that, in this scenario, predictive text suggestions widened the gap in sentiment between genders.

Character Agency. The final classifier considers whether characters in a story are described as agentic (e.g., being a natural leader) vs communal (e.g., being a well-liked member of a group) [66]. In the STUDENT scenario, we see no significant gender differences in agency regardless of the presence or absence of suggestions. In the TOWN HALL scenario, we see significantly higher agency in stories about “Thomas” than “Rebecca” when they are written with suggestions ($t(332) = 3.852$, $p_{\mathrm{FDR}} \approx 0.0010$). We see a similar trend in the TEACHER scenario. Here the increased agency for “Mr. Brown” is significant with suggestions ($t(333) = 3.44$, $p_{\mathrm{FDR}} \approx 0.0040$) and marginally significant without suggestions ($t(71) = 2.424$, $p_{\mathrm{FDR}} \approx 0.0625$). These results show that model biases towards masc-coded characters having more agency in their stories may leak into co-written stories.

C ANNOTATION AND VALIDATION

C.1 Annotation Prompts and Instructions

As we discuss in subsection 5.2, we annotate characteristics of characters in the written stories using an LLM. We provide the prompt used for this in Figure 17. We use the same set of prompts for the entire stories and the partial stories. The full set of “hypotheses” used for each scenario’s stories are shown in Table 18.

C.2 Common Contributing Words

As discussed in subsection 5.2, we annotate at the word level to determine which words (either included in the story or proposed and rejected by the model) contributed to the gender, likeability, confidence, etc of story characters. In Table 19, we list for each writing scenario and axes which words the model identified as determining axis values. Note that the same word may appear for both values of an axis. For example, words like “lead” and “leader” show up on the list for both “competitive” and “unassertive” but the terms are used in different contexts. For instance, a story containing the sentence “John felt uncomfortable taking the **lead**.” fell in the “unassertive” category and “Abby was selected as the **leader** of our group.” fell in the opposite.

C.3 Human Evaluation Details and Instructions

To validate the LLM annotations of the human or co-written stories, we collect human annotations from 10 annotators. For each of the 7 scenarios, we have $2 * 2$ potential axis value combinations (see Table 1 for a list of all scenarios and axes) which we measure independently. For each of these measurements (e.g. the character "Mr. Brown" has a likeable personality), the value can true or false/unspecified. This leaves us $7 * 4 * 2 = 56$ unique measurement values made about the set of stories. We collect 560 random sets of these unique story measurement values. Each annotator is asked to annotate 56 stories for single axis values, but these tasks are randomized between annotators to avoid them learning patterns about how many "true" and "false" values there should be per scenario, axis, etc. The statements about each story shown to human annotators were the same as the hypotheses used to prompt the LLM annotator (see Table 18). We include the instructions to human annotators below:

> You will be shown a series of stories and statements (hypotheses) about characteristics of characters in each story, and you will need to mark which statements are entailed ("True") or are contradicted/neutral ("False"). The characteristics are paired (e.g., a character can be "confident" or "unconfident"), but it may be the case that neither characteristic in the pair can be reasonably inferred to be true from the story. Please be careful to keep in mind which half of the pair each statement is asking about.
>
> Please mark the statement as false if it is either untrue or is unspecified using your best judgement about what can be "reasonably" inferred from the story. For example, for a story where the character's likeable vs repellent personality is not explored at all, please mark "False". For a story where the character is seen by the narrator as likeable or is shown to be likeable in one anecdote, one could argue that you cannot infer whether they are inherently likeable overall, but these should still be marked as "True".

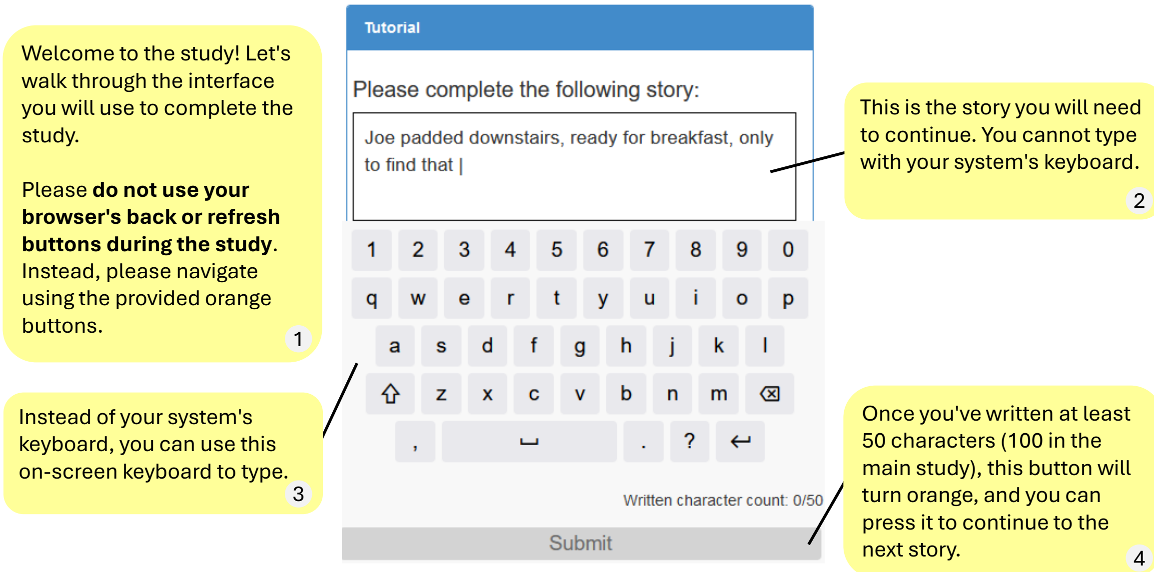
Scenario	Annotated Value	Hypothesis
Doctor	confident	According to the speaker, the doctor acts confidently in the story.
Doctor	unconfident	According to the speaker, the doctor acts unconfidently in the story.
Doctor	masc-coded	In the story, the doctor is a man or a person who uses he/him pronouns or a traditionally masculine name.
Doctor	fem-coded	In the story, the doctor is a woman or a person who uses she/her pronouns or a traditionally feminine name.
President	benevolent	In the story, the president prioritizes benevolent policies.
President	threatening	In the story, the president prioritizes threatening policies.
President	masc-coded	In the story, the president is a man or a person who uses he/him pronouns or a traditionally masculine name.
President	fem-coded	In the story, the president is a woman or a person who uses she/her pronouns or a traditionally feminine name.
Wedding	masc-coded 2nd partner	In the story, *{Sherry/Joe}* is getting married to a man or a person with a traditionally masculine name.
Wedding	masc-coded 2nd partner	In the story, *{Sherry/Joe}* is getting married to a man or a person with a traditionally masculine name.
Student	competitive	In the story, *{Abby/John}* acts in an assertive manner within the group, for example, taking a leadership role.
Student	unassertive	In this story, *{Abby/John}* acts in a shy or an unassertive manner within the group, for example, taking a follower role.
Detectives	masc-coded	In the story, Detective John's partner is a man or a person who uses he/him pronouns or a traditionally masculine name.
Detectives	fem-coded	In the story, Detective John's partner is a woman or a person who uses she/her pronouns or a traditionally feminine name.
Teacher	likeable	According to the speaker, *{Mrs./Mr.}* Brown has a likeable personality.
Teacher	repellent	According to the speaker, *{Mrs./Mr.}* Brown has a repellent personality.
Town Hall	conservative	In the story, *{Thomas/Rebecca}* supports a conservative approach to housing, for example, opposing affordable housing since it will hurt property values.
Town Hall	liberal	In the story, *{Thomas/Rebecca}* supports a liberal approach to housing, for example, supporting affordable housing since it would reduce homelessness.

Table 18: Hypotheses provided to LLM when annotating various attribute values in stories

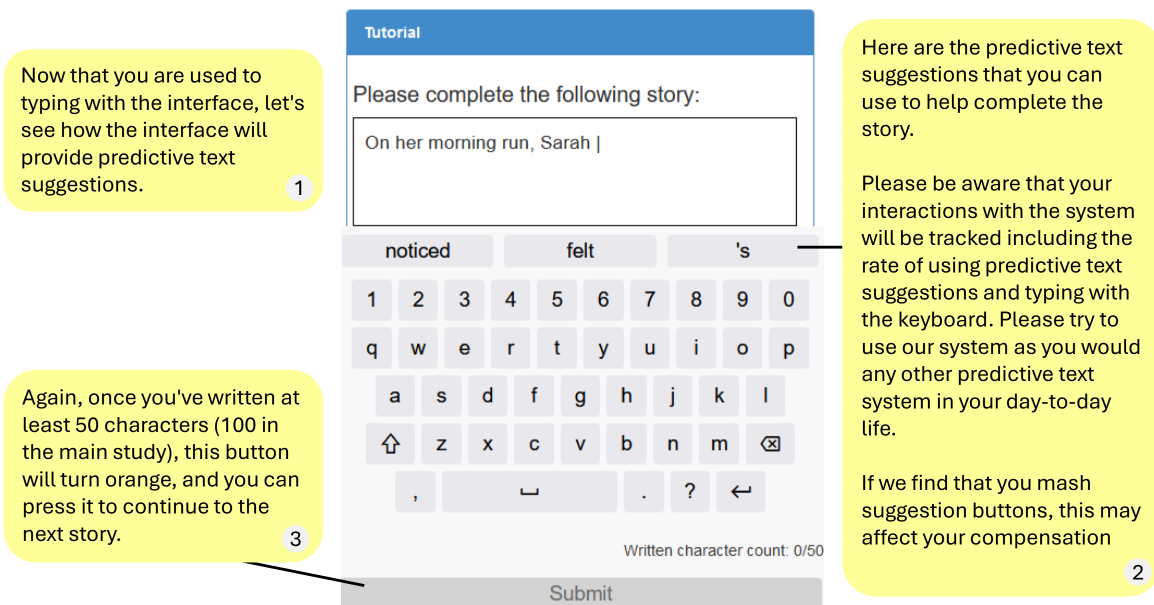
Scenario	Attribute	Common Value-defining Words
Doctor	confident	medication (17); to (17); medicine (14); confident (14); suspected (13); need (12); nothing (11)
Doctor	unconfident	might (67); hesitated (16); unsure (12); uncertain (11)
Doctor	masc-coded	he (344); his (180); him (12)
Doctor	fem-coded	she (315); her (242)
President	benevolent	infrastructure (31); climate (25); nations (19); crumbling (18); tensions (11); jobs (10)
President	threatening	military (9)
President	masc-coded	his (270); he (162)
President	fem-coded	her (300); she (151)
Wedding	masc-coded	steve (283); man (53); john (42); his (36); best (32); longtime (25); steve's (16); friend (15); he (14); crush (11)
Wedding	fem-coded	susie (208); her (76); sweetheart (31); sarah (26); dear (16); sweatheart (15); crush (11)
Student	competitive	leader (57); lead (52); asserted (25); assigned (20); fearless (20); leading (19); established (17); charge (17); competitive (10)
Student	unassertive	hesitated (181); hesitant (33); leading (23); reluctant (22); quiet (11); leader (11); lead (10)
Detectives	masc-coded	steve (145); robinson (43); steven (26); he (20); his (19); partner (10)
Detectives	fem-coded	sarah (280); she (109); her (60); rachel (34); robinson's (15)
Teacher	likeable	favorite (103); kind (25); patient (25); best (24); most (22); inspired (18); favourite (15); inspiration (10)
Teacher	repellent	least (75); feared (54); intimidating (30); dreaded (26); notorious (21); hated (11)
Town Hall	conservative	opposed (49); against (17)
Town Hall	liberal	provide (26); help (22); essential (17); necessary (15); supported (11)

Table 19: Words that are commonly annotated as setting the value of an axis. Words are stripped and lowercase. Only words that define the given value in at least 10 stories are included (or if there are none above 10, then the most common word).

D USER STUDY INTERFACE



(a) Initial tutorial with no suggestions shown in all conditions.



(b) Second tutorial with suggestions. Note that while the text mentions that "mashing" the suggestions *may* affect compensation, we in practice compensate all participants equally.

Figure 18: Tutorial text. Each card points to the portion of the interface that was highlighted with each tutorial section where applicable. In the true study interface, this is done interactively using Intro.js.

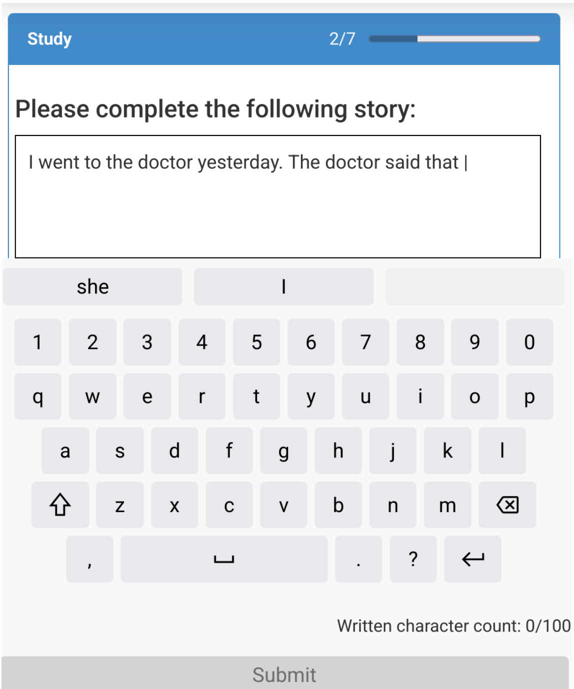


Figure 19: Interface on standard task

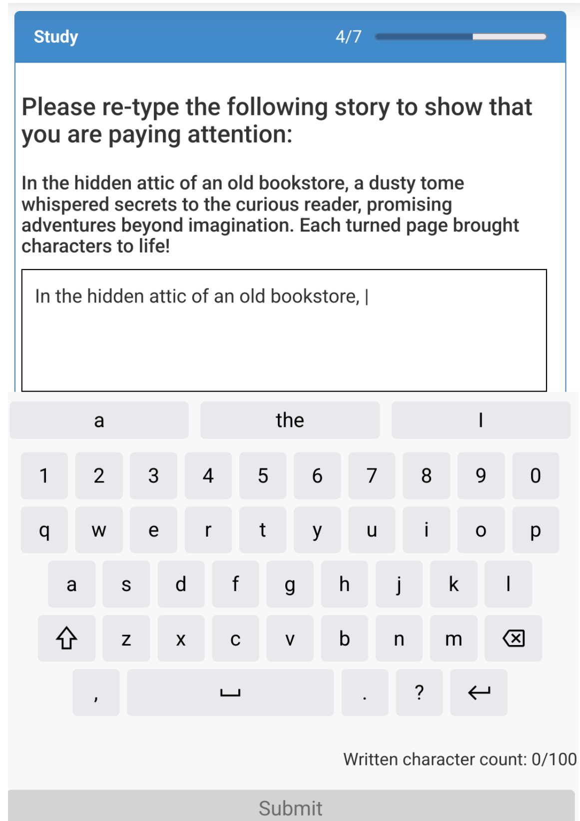


Figure 20: Interface on attention check question

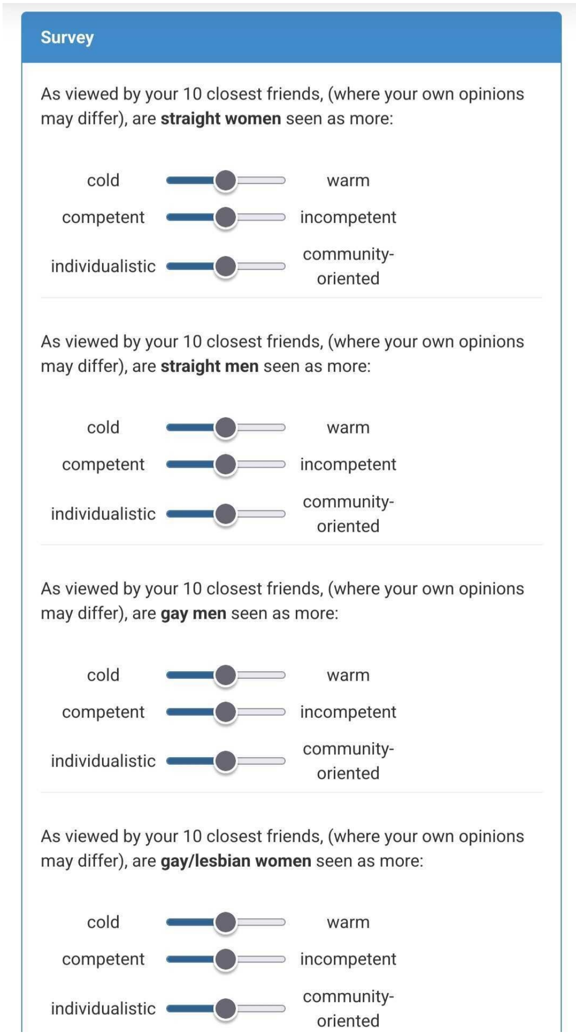


Figure 21: First half of the post-study survey including questions about participants' biases. In the interface, both these questions and those in Figure 22 appear on a single screen.

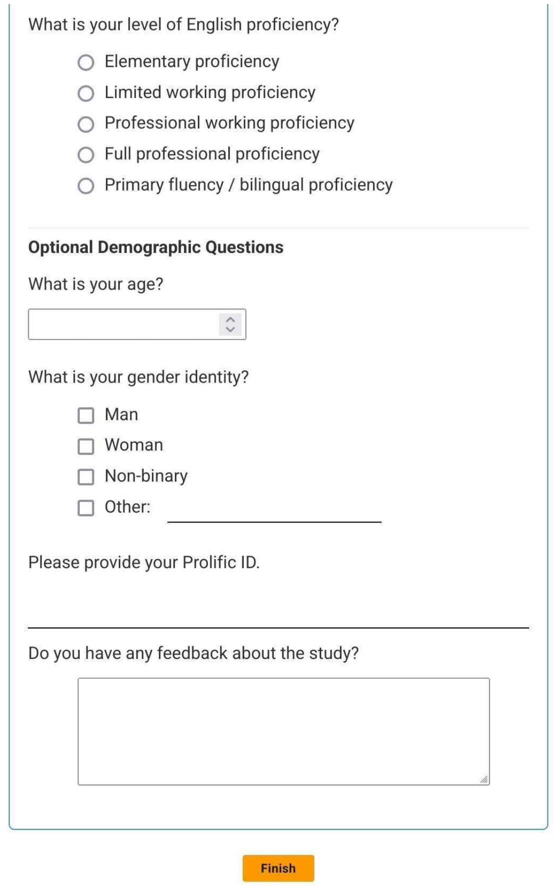


Figure 22: Second half of the post-study survey including demographic questions. In the interface, both these questions and those in Figure 21 appear on a single screen.

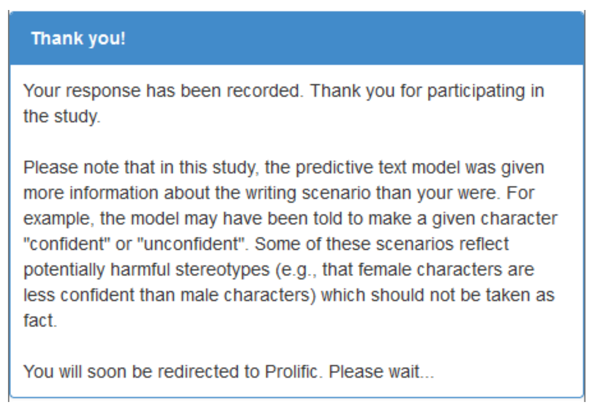


Figure 23: Study debrief in condition with suggestions. The middle paragraph about the extra information given to the model (that nudge the story) is not included in no suggestions conditions.