Evaluating LLMs for gender disparities in notable persons

Reference:
Rhue Lauren, Goethals Sofie, Sundararajan Arun.- Evaluating LLMs for gender disparities in notable persons
ArXiv, 2024, 17 p.
Full text (Publisher's DOI): https://doi.org/10.48550/arXiv.2403.09148
To cite this reference: https://hdl.handle.net/10067/2044750151162165141

EVALUATING LLMS FOR GENDER DISPARITIES IN NOTABLE PERSONS

Lauren Rhue
Robert H. Smith School of Business, University of Maryland
USA
rhue@umd.edu

Sofie Goethals
University of Antwerp
Belgium
sofie.goethals@uantwerpen.be

Arun Sundararajan
Stern School of Business, New York University
USA
arun@stern.nyu.edu

ABSTRACT

This study examines the use of Large Language Models (LLMs) for retrieving factual information, addressing concerns over their propensity to produce factually incorrect "hallucinated" responses or to altogether decline to even answer prompt at all. Specifically, it investigates the presence of gender-based biases in LLMs' responses to factual inquiries. This paper takes a multi-pronged approach to evaluating GPT models by evaluating fairness across multiple dimensions of recall, hallucinations and declinations. Our findings reveal discernible gender disparities in the responses generated by GPT-3.5. While advancements in GPT-4 have led to improvements in performance, they have not fully eradicated these gender disparities, notably in instances where responses are declined. The study further explores the origins of these disparities by examining the influence of gender associations in prompts and the homogeneity in the responses.

***K*eywords** Large Language Models, Information Retrieval, Bias, Fairness, Hallucinations

1 Introduction

Large Language Models (LLMs) are increasingly used to access the repertoire of human knowledge, including factual knowledge, due to their extensive understanding of available digital traces. This is particularly notable in the task to reproduce factual knowledge. Despite the rapid advancement and widespread deployment of these models in various applications, the development of robust and universally accepted fairness metrics has lagged behind. Current research primarily focuses on performance metrics such as accuracy, efficiency, and generative capabilities, often overlooking the nuanced aspects of fairness and bias [1].

The factuality of the LLMs outputs may vary according to the gender association of the facts; put differently, LLMs may forget and "misremember" in gendered ways. These "misremembered" responses, or factually incorrect hallucinations, reflect the models' underlying associations. When models do not correctly identify the facts, they may decline to answer or they may return hallucinations that are associated with the prompt in their learned connections. For LLMs, there is a probability over the names in the training data that reflect the associations of the names in the space. Through this project, we probe the gendered output of factual questions in LLMs to evaluate the gender disparities in the factuality of LLMs.

The evaluation of LLMs consists of a four-part process: the model, the task, the data, and the process, i.e., the what, where, and how of the model [2]. This paper evaluates the factuality of models around the task of notable persons. Current evaluations of LLMs have not examined the factuality through the lens of notable persons; however, crowd-sourced knowledge sources such as Wikipedia exhibit gender disparities in the representation of notable persons [3].

Given that models are trained and evaluated on Internet-based sources like Wikipedia, it is expected that LLMs will exhibit gender disparities in the names returned and in the factuality of the responses.

This paper evaluates GPT models' responses to prompts where the answer is a specific notable person. LLM responses are then examined through fairness metrics, and this paper evaluates responses along multiple dimensions: factuality, hallucination rate, and declination rate. First, this paper compares the inaccuracy of the responses by gender, evaluating whether female notable people are more likely to be misremembered by the LLM. Additionally, we investigate the gender distribution in the hallucinations. More recent versions of GPT can decline to answer when the LLM is uncertain about the factual response (new for GPT-4). If the LLM declines to answer, then the LLM is displaying a realization of its uncertainly about the factual answer. If an LLM produces a factually incorrect response, rather than declining, it suggests an unawareness of its incorrect response. We will also evaluate whether there are any gender disparities in the declination rate. The results suggest that the increased performance of GPT-4 is insufficient to eliminate all disparities; gender disparities can persist even with improved performance.

At the moment, there is no consensus on how to measure bias in LLM outputs, and metrics to address this gap are urgently needed. Current fairness metrics are designed to measure disparities in predictive models and are not necessarily well-suited to measure disparities in generative models. In addition to evaluating the models for gender disparities, we introduce a new fairness evaluation measure, Response Concentration Score (RCS), to evaluate how representative the responses are in comparison to the distribution in actual responses. On this measure, GPT-4 is a marked improvement in fairness over GPT-3.5.

Lastly, this paper explores patterns in hallucinated names by considering the co-occurrence of gendered names and the relationship between the gendered name and the gender distance of words in the prompts. Male names are more common as more names are hallucinated, which suggest a pattern of homogeneity in hallucinated names. Hallucinated gendered names likely emerge as a property of the words in the prompts, with more female-associated words leading more hallucinations of female names. Using the word embeddings from GloVe [4] to calculate a gender association, this study finds that names with higher connection to female pronouns are more likely to be correlated with female hallucinations for earlier models, but this correlation disappears with GPT-4.

To summarize, we make the following contributions to this domain. First, we evaluate disparities in the output of the LLMs among various dimensions (factuality, hallucinations and declination). Second, we propose a new fairness metric (Response Concentration Score), specifically for the output of generative models and compare this metric with a fairness metric, commonly used in predictive modeling (Demographic Parity Difference). Third, we investigate other factors, such as the impact of gender associations in the prompt, and how likely different genders are to co-occur in the output.

The paper is organized as follows. Section 2 describes the related work. Section 3 describes the materials and methods used in the analysis. Section 4 covers the results of the analysis. Finally, section 5 discusses the results and concludes the paper.

2 Related Work

2.1 Evaluation of Factuality in LLMs

Large language models (LLMs) are computational models that leverage probabilistic methods to understand and generate human language [5]. These models can predict a set of word sequences or generate new word text in response to specific inputs or *prompts* [2], and possess the ability for in-context learning, where they tailor their output to align with the nuances of the given prompt [2]. However, a notable challenge when using LLMs is their lack of interpretability; the internal mechanisms driving their output often remain opaque, making it critical to develop reliable evaluation metrics [2].

As LLMs expand to new contexts, such as information retrieval, there is growing concern over the factuality in the responses of LLMs and the potential for "factual hallucinations" [2]. Much of the current literature around evaluation of factuality are related to automated methods to check any statement for its truth, such as whether the response mostly accurate or contains shades of inaccuracies [6]. BERTScore, a semantic similarity measure to evaluate LLMs, does not necessarily capture the binary nature of factual recall [7].

In addition to evaluating whether a response is hallucinated, some papers examine patterns and properties around hallucinations [8]. Much of this work is related to automatically evaluating factual information across a variety of domains [8, 9, 10]. While automatic fact checking is important [11], these factual evaluation tasks do not include the distribution of gender-associated facts [8, 2].

Studies have suggested that changing the prompts and fine-tuning models will improve the factuality of the answers [12]. This paper examines multiple ways of measuring the change in factuality using the knowledge instillation. The instillation of the specific fact, e.g., mentioning that fact in the prompt, may only make sense when the fact is known a priori, which may not work for information retrieval tasks. We design our experiment to focus on the specific entity errors in the factual hallucinations [13]. We note that there are differences in the prevelance of notable figures by gender and their representation in the Internet, so we examine whether the entity errors differ by facts associated with men or women.

2.2 Evaluation of Disparities in LLMs

As LLMs learn language, stereotypical gender associations emerge from the aspects of our language and society that are reflected in digital traces [14]. Recent studies of bias in LLMs add to a growing body of research about fairness in machine learning [15, 16, 17, 18]. Several studies examine models for the presence of bias. Dong et al. [19] probe LLMs for explicit and implicit bias by using conditional text generation, while Wan et al. [20] analyze bias in LLM-generated reference letters. Cao et al. [21] probe the cross cultural context of LLMs, whereas Hartmann et al. [22] probes models' political ideology. Dhamala et al. [23] use open-ended questions to probe biases in the models. Zhuo et al. [24] explore the bias in GPT models through benchmarking on sample datasets and find evidence of bias; furthermore, the bias is difficult to eliminate from prompts.

One opportunity for further exploration is evaluating how LLMs produce factual information related to notable persons, and whether this information is generated in gendered ways. Factually incorrect hallucinations can occur for many reasons. Lin et al. [9] asks, "Why do language models generate false statements?", noting that "One possible cause is that the model has not learned the training distribution well enough." This paper wonders if the language models have learned the training distribution, including its gender disparities, and are reflecting those in their responses.

Genders emerge from a variety of personal characteristics in digital traces, such as image and names. Names can convey social identity information such as gender or race if they are sufficiently distinctive [25, 26, 27, 28, 29]. Names have also been used as the minimal information to create stereotypes [25, 27, 26], based on the association of names with social identity, and LLMs are known to have stereotypical bias [30, 14]. In understanding and evaluating LLMs, there are stereotypes and meaningful associations on different dimensions [14]. Methods to evaluate social bias in LLMs have included word-based lists (e.g., [31]), templates (e.g., complete the sentence to determine the association between concepts and words [32], crowdsourced social bias tests [32], and social media tests. When we ask generative models factual questions that should be answered with a name, its output may contain gendered implications. Given the increased use of LLMs, and the potential for negative externalities, there is an opportunity to examine the gender disparities in the responses of Large Language Model (LLMs) and its connection to factual accuracy.

Gender disparities in the generated output can emerge through multiple different processes:

1. True Representation of Underlying Distribution. Names are generated occurring to a distribution that reflects the real-world distribution of names and the associated social identity characteristics. For instance, an LLM generates names in a way that accurately reflects the underlying distribution of names in that context. If 90% of the names in the context are associated with male social identity, then the LLM would generate male names at a similar rate.

2. Association-based Representation of Distribution. Names are generated in a way that reflects stereotypical associations of the social identity characteristics, thereby leading to disparities in representation across social identities. For instance, an LLM generates names for one social identity group at a higher (lower) rate than is appropriate for the underlying context. If 60% of the names in the context are associated with male identity, yet more than 90% of the names are associated with male social identity, then the LLM is generating an association-based distribution.

3. Prejudice Representation of Distribution. The distribution of names is intentionally forced to follow a predetermined distribution that does not reflect the true underlying distribution. For instance, an LLM generates 60% male names in a context with 60% male names, but rejects the female names because the LLM has guardrails to prevent the representation of women in the output.

The uses of LLMs vary, as do the consequences; however, the gendered outputs of LLMs are neutral on their own. Some scholars argue that LLMs should be biased [33] to accurately reflect the language in the training data [34]. Gendered outputs are only problematic when they "create or reproduce structures of domination based ... identities" [35]. Gendered outputs may diminish those structures of domination if their outputs defy stereotypes; however, LLMs may learn and reinforce those structures if their outputs perpetuates stereotypes by associating stereotypical characteristics with gendered identity characteristics. Thus, there is a potential trade-off between fidelity to the real-world and stereotypical behavior.

In this project, we evaluate the gender disparities in LLMs by probing the factuality of models' responses in response to a specific question for information.

3 Materials and Methods

In our investigation, we compare the gender-specific behaviors exhibited by the GPT-3.5 and GPT-4 language models to evaluate how performance improvements shape gender disparities. The structure is as follows.

3.1 Materials

We use three lists of notable persons: a list of entrepreneurs, a list of actors, and a list of Nobel Prize winners. We prompt an LLM to answer a question in which this notable person is the correct answer. Through this prompt set-up, the study assess gender disparities in the factuality of the returned response and probes the factors associated with factual hallucinations and gender of notable figures. We run each prompt to identify the notable person five times for every parameter combination (LLM engine, temperature) for robust results. We use both the default DaVinci-003 engine (GPT-3.5, OpenAI) as well as GPT-4 for our experiments, and these experiments are within the terms of use.

3.1.1 Entrepreneurs

The Entrepreneurs list is collected data from Forbes Next 1000 via web scraping, focusing on names, company, role, and industry information. [1] The data instances with the role of 'Founder' are used for these experiments. The following prompt was run for each founder in the Forbes 1000 list.

Prompt. "Who founded the company [Company] in the industry [Industry]? Return the names in a list like this: Name1, Name2,.. Name n".

The above prompt was run five times for each notable person in the list. This entrepreneurs' list occurs outside the time frame of the training data (although the start-up companies themselves existed within the time frame of the training data). This task contains notable persons that will have fewer digital traces in the LLM training dataset, so this task is expected to generate more factual hallucinations than the other tasks.

3.1.2 Nobel Prize Winners

The list of Nobel Prize winners from 1901 through today in all fields was collected via web-scraping. The Nobel Prize winners data contains Year, Subject, Discovery, and their names. The prompt contains the Year and Subject, as shown below.

Prompt. "Who won the Nobel Prize for [Subject] in [Year]? Return the names in a list like this: Name1, Name2,.. Name n".

The above prompt was run five times for each Nobel Prize winner. The task has a narrow search space because Nobel Prize winners are well-documented and included in the training data for both GPT models. Nobel Prize winners receive global media coverage and have extensive digital traces, so this task is expected to generate fewer hallucinations. In addition, the set of Nobel Prize winners is heavily male, so this task could probe the gender disparities in the recall of notable persons.

3.1.3 Actors

The list of Oscars winners for Best Actor and Best Actress from 1929 through 2022 are collected from the Oscars website via web scraping, focusing on names only. The actors data contains Year and their names. The prompt contains the Year and the award type, as shown below.

Prompt. "Who won the Oscars for Best [Actor/Actress] in [Year]? Return the names in a list like this: Name1, Name2,.. Name n".

[1] `https://www.forbes.com/next1000/`

The above prompt was run five times for each Oscar winners. Oscar winners receive wide coverage in the United States, and the winners are often actors with world-wide fame, so there are extensive digital traces of these notable people. The list is gender-balanced as well because the Best Actor and Best Actress awards are granted every year.

3.1.4 Data Description

Each of these lists yields a dataset with attributes related to the notable figure, experimental run, and the LLM run. *Notable Person Name* is the name of the notable person. *Year* is the year of their accomplishment, ranging from 1901-2022 for the Nobel Prize Winners or 1929-2022 for the Oscars winners. *Subject* is only associated with Nobel Prize winners, and describes the subject of their Nobel Prize (Physics, Literature, Medicine, Chemistry or Economics). *Industry* is only associated with Entrepreneurs, and is the industry of the start-up. *Company* is also only associated with Entrepreneurs and is the name of the entrepreneur's start-up. *Gender of Notable Person* is the gender associated with the notable person in the list. For the Nobel Prize and Actors, gender is determined through historical records. For the Entrepreneurs, gender is predicted from the name. *GPT-generated person* is the name produced by the GPT-model in response to the prompt. *Gender of GPT-generated person* is the predicted gender associated with the name produced by the GPT-model in response to the prompt. *Correctly Identified* is a binary variable that is 1 if the GPT-model returns the correct answer for the prompt and 0 otherwise. *Incorrect* is 1 if GPT-model returns an incorrect answer. This answer can be incorrect due to a declination to answer (acknowledging the lack of knowledge) or a hallucination (factually incorrect named response). *Run* is the index for the run, ranging from 1-5. *Temperature* is the measure of creativity and takes three values: 0, 0.5, and 1. The variable definitions for these datasets are shown in Table 1.

Table 1: Variable Definitions in the datasets

Variable Name	**Definitions**
Notable Person Name	Name of the founder or award-winner
Year	Year of award (Nobel Prize and Actors)
Subject	Subject of award (Nobel Prize)
Industry	Industry of the Entrepreneur (Entrepreneur)
Company	Company name of the Entrepreneur (Entrepreneur)
Gender of notable person	Gender from historical records (Nobel Prize and Actors) or predicted from name (Entrepreneur)
GPT-generated person(s)	Name(s) produced by GPT model
Gender of GPT-generated person(s)	Gender predicted for the GPT name(s)
Correctly Identified	Binary; 1 if GPT is factually accurate
Incorrect	Binary list; 1 if factual inaccurate
Run	Index for each prompt query (1-5)
Temperature	A measure of creativity (0, 0.5, 1)

3.2 Methods

3.2.1 Prominence

To understand the effect of the prominence of both the Noble Prize winners as the Oscard Award winners, we calculated Google Search Counts for each winner. We use SerpAPI (Google Search API) and find the number of search results for each notable figure's name. Search counts have been used for many years as a proxy for the current prominence of public figures [36].

3.2.2 Creativity

The models' creativity is varied by running the same prompt with three different values for temperature – temperature = 0 (most deterministic), temperature = 0.5, and temperature = 1 (most creative). The results provide insights into the model's tendencies to generate names and into the gender patterns in the names generated. Higher values of the temperature make the output more random and creative, allowing the model to explore different possibilities and produce more varied responses. Lower values of temperature make the output more focused and deterministic, leading to more conservative and predictable responses. The results provide insights into the model's language understanding, generative capabilities, and tendencies to hallucinate incorrect responses.

3.2.3 Gender

We include additional covariates related to the names. To study gender, we need to associate each person with a gender. We determined the gender probabilities of some notable persons based on the Social Security Administration (SSA) gender file, an approach that has been used in other studies Blevins and Mullen [37], Hu et al. [38]. We collect the first names of the persons and assign a probability of the name being male based on the percentage of people with this first name that are born male.[2] Names that are not included in the SSA database are listed as "unknown". We analyzed the frequency of responses by gender.

3.2.4 Gender associations

We use GloVe to examine the gender associations with the industry and company name in order to understand the patterns in the responses. GloVe (Global Vectors for Word Representation) is an unsupervised learning algorithm for generating vector representations of words, capturing their meanings based on their co-occurrence in large text corpora [4]. Consistent with Pennington et al. [4], gender is represented as a vector of gender-associated pronouns. The female vector uses the words "she", "her", "hers", "woman", and "female" and the male vector uses the words "he", "him", "his", "man", and "male". The cosine similarity is then calculated between 1) the industry vector embedding and the gender vector embeddings, and 2) the company + industry vector embeddings and the gender vector embeddings. The gender associations of the industry and the gender association of the company names are the cosine similarity between the industry and gender vectors and the company + industry and gender vectors respectively.

3.3 Metrics

In order to evaluate the LLMs, we calculate several metrics associated with LLMs' output.

3.3.1 Recall

The first metric captures whether the LLM-generated outputs are factually correct. If the last name of the generated name is the same as the notable person's last name, then the output is categorized as correct. This automated process was checked by human annotators to determine its accuracy. *Recall* is defined as the average percentage of factually correct LLM outputs across the five runs.

3.3.2 Miss rate

The second metric is the inverse of *Recall* and captures whether the LLM-generated outputs are *not* factually correct. *Miss Rate* is defined as the percentage of instances that are incorrectly identified by the LLM over the five runs.

$$\text{Miss rate} = 1 - \text{Recall} = \text{Hallucination Rate} + \text{Declination Rate}$$

Missed or non-recalled answers could include responses with inaccurate names, i.e., hallucinations, or responses that decline to answer with a name, such "I do not know that information" or "That information is outside of my training data." Hallucination rate only refers to the percentage of named yet factually incorrect responses. Declination rates are the percentage of responses in which the LLM declines to answer.

3.3.3 Fairness metrics

There is no universal definition of fairness, but the majority of literature quantifies this by measuring the disparities in prediction outcomes. Demographic parity is among the most frequently utilized metrics and measures whether the probability of a positive outcome is the same across different demographic groups, implying that the decision process does not favor one group over another based on sensitive attributes such as race, gender, or age. Demographic parity can be quantified as either a difference or a ratio between outcomes for demographic groups.

This paper focuses on the Demographic Parity Difference (DPD). This measures the absolute difference in the rate of positive outcomes between the privileged and unprivileged groups. It is calculated as:

$$DPD = |P(\hat{Y} = 1|D = 1) - P(\hat{Y} = 1|D = 0)|$$

Lower values of DPD suggest fewer disparities between the groups. A DPD value of 0 indicates no disparities between groups.

[2] This study utilizes a binary gender classification based on gender assigned at birth, with the understanding that this approach does not capture the full diversity of the gender spectrum.

4 Results

4.1 Factuality

The first evaluation is the factuality of the GPT-models. We run the prompts five times for each temperature and report the results associated with the queries in Table 2. Each task reveals a different aspect of the GPT models, so a detailed description of the results by task are below.

Table 2: Miss rate across tasks

		GPT-3.5			GPT-4		
Task	Population	t = 0	t = 0.5	t = 1	t = 0	t = 0.5	t = 1
Entrepreneurs	Overall	0.945	0.950	0.967	0.637	0.641	0.649
	Female	0.940	0.944	0.966	0.609	0.611	0.623
	Male	0.950	0.955	0.970	0.658	0.666	0.669
	p-value	0.580	0.49	0.728	0.022	0.022	0.030
Actors	Overall	0.333	0.347	0.414	0.027	0.025	0.033
	Female	0.375	0.392	0.460	0.031	0.029	0.043
	Male	0.292	0.303	0.369	0.023	0.022	0.022
	p-value	0.007	0.005	0.005	0.423	0.538	0.073
Nobel Prize	Overall	0.347	0.357	0.438	0.044	0.044	0.046
	Female	0.241	0.208	0.290	0.020	0.029	0.037
	Male	0.352	0.364	0.446	0.043	0.043	0.045
	p-value	0.000	0.000	0.000	0.019	0.200	0.526

4.1.1 Entrepreneurs

With this task, identifying notable start-up founders from 2021, the GPT-3.5 outputs were nearly entirely hallucinated. The miss rate ranged from 94.5% (temperature = 0) to 96.7% for the most creative model (temperature = 1), and GPT-3.5 always responsed to a prompt with a name. The lower miss rate for the least creative model is driven by the decrease in the unique names and the lack of stability in the hallucinations. Therefore, the most deterministic model yields stability in the hallucinations across the five runs and a slightly better recall.

More than 95% of facts are not correctly recalled by GPT-3.5. GPT-4 is more accurate with a lower miss rate; however, this improvement is not uniformly distributed. GPT-4 has a higher miss rate with male founders, suggesting that female founders are more likely to be recalled. A t-test of the miss rates shows that the gender disparities in miss rates are significant for temperatures of 0.5 and 1 for GPT-4.

This performance difference between GPT-3.5 and GPT-4 is highlighted in Table 2. GPT-3.5 frequently produced incorrect answers for each entrepreneur, with the miss rate often reaching nearly 100% across the five runs. In contrast, GPT-4 demonstrated a greater likelihood of providing factually correct responses. However, its performance varied significantly by founder, with the miss rate fluctuating between close to 0% for some founders and around 100% for others, indicating a pattern of either consistently delivering correct answers or consistently generating factual hallucinations.

4.1.2 Actors

With this task, identifying notable acting award recipients, the GPT-3.5 outputs were somewhat hallucinated, as can be seen in Table 2. Despite the extensive award coverage and presence in the training data, GPT-3.5 exhibited a miss rate on average for between 33% to 41% of the award-winners, highlighting the challenge of factuality in earlier language models. In contrast, GPT-4 only exhibited miss rates between 2.7% (temperature = 0) and 3.3% (temperature = 1) of the award recipients. Across both models, the higher miss rate is associated with the more creative and less deterministic model.

As for gender disparities in factuality, GPT-3.5 displays a significantly higher miss rate for female award-winners as compared to male award-winners. The performance improvement brought by GPT-4 reduces disparities between groups and yields similar miss rates for men and women. To assess whether these differences are statistically significant, we compare the mean miss rates across the two groups with a t-test. The p-value row in Table 2 indicates the statistical significance of the t-test. According to this test, the GPT-3.5 outputs exhibited a notable gender disparity in the miss rates, but GPT-4 outputs exhibit statistically indistinguishable results. The Actors rows in Table 2 show the results.

4.1.3 Nobel Prize

With this task, identifying Nobel Prize winners, we only look at the winners through 2022, when GPT-3.5 could answer the question. The GPT-3.5 outputs were more factual than for the Entrepreneurs but still with a notable miss rate. The miss rates ranged from 34.7% (temperature = 0) to 43.8% (temperature = 1). Despite the extensive coverage for the awards, GPT-3.5 still faced challenges with recalling correct answers. With the improvements from GPT-4, the miss rates were reduced to only 4.4%-4.6% of the responses. As with the Entrepreneurs list, the least creative model produces fewer unique names and stability across runs, whereas the more creative model produces different names each run.

As for the gender disparities, there were significant gender differences for GPT-3.5 and GPT-4. Female Nobel Prize winners, who account for roughly 5% of winners, were significantly more likely to be recalled than male Nobel Prize winners. A t-test of the difference between the miss rates for male and female Nobel Prize winners confirms that the difference is statistically significant at the p-value < 0.001 for GPT-3.5 For GPT-4, the miss rates for female and male Nobel Prize winners are more comparable, with the differences in miss rates being non-statistically significant for GPT-4 at a temperature of 0.5 and 1. For this task, the performance improvement brought by GPT-4 reduces disparities between groups. Table 2 shows the results.

4.1.4 Summary

There are gender differences in the miss rates of notable persons, but the direction of the differences varies. For the Actors list, female names are less likely to be correctly recalled. However, for the highly skewed list (Nobel Prize winners) with only 5% women, the women are more likely to be recalled, likely due to their larger digital traces. Although this recall pattern differs across temperatures, some disparities in miss rates persist for more advanced LLMs.

The counterintuitive result, that male Noble Prize winners are less likely to be correctly recalled, is potentially due to the recent celebration of female achievements. The greater digital traces of female notable persons may explain why GPT's responses are better recalled for this group. Female Nobel prize winners have an average of 260% more search results than male Nobel Prize winners, suggesting that female Nobel Prize winners are on average more prominent than their male counterparts. This suggests that the LLM may over-represent female Nobel winners relative to their presence in the real world, explaining why the factuality is better for female Nobel Prize winners. In contrast, female Oscars award winners have 17% fewer search results than male Oscars award winners, suggesting a lower profile for female actresses relative to their male counterparts.

4.2 Fairness Metrics

The analysis of miss rates suggests evidence of gender disparities. To more deeply evaluate the gender disparities of LLMs, we compare the miss rates using the fairness measure of Demographic Parity Difference (DPD) that was introduced in Section 3.3.3.

We also introduce our measure of Response Concentration Score (RCS) that builds on previous of work to identify gender disparities in LLMs. The RCS measures the deviation of the name distribution produced by the LLM from the context-specific actual distribution. In this measure, higher scores indicate models with a greater fidelity to both accuracy and the underlying distribution of names in the hallucinations. We define the Response Concentration Score as:

$$RCS = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(1 - |\%response_k - \%actual_k|)^2}$$

Where k is the number of relevant social identity categories (gender in this instance), $\%response_k$ is the percentage of names that are LLM-produced names in identity category k, and $\%actual_k$ is the percentage of the names that are actually associated with identity category k. The RCS ranges from $1/k$ (lowest) to 1 (highest). An RCS score of 1 is consistent with outputs that reflect the true underlying distribution, whereas lower RCS scores reflects a skew that deviates from the underlying distribution and illustrates a preference for only one single social identity category in the LLM responses.

Table 3 shows the GPT-3.5 and GPT-4 evaluations using the demographic parity of the miss rates and the RCS across temperatures. According to DPD, GPT-3.5 displayed lower disparities among genders for the Entrepreneurs and Actors tasks than the more accurate GPT-4. GPT-3.5 only displayed higher gender disparities in the highly skewed Nobel Prize winners task. Because DPD only considers the relative performance rates across groups, this metric does not account for worse overall performance of GPT-3.5. In contrast, RCS increases with both the accuracy of the responses as well as fidelity to the underlying distribution in the dataset for hallucinations. According to RCS, GPT-4 generates similar or lower disparities than GPT-3.5 across all tasks. For GPT-3.5, the lower RCS for the Entrepreneur list is driven by the preference for male names across nearly all industries and across all models. The higher RCS is due to the declination of GPT-4 to answer some prompts in case of uncertainty, so the underlying distribution of names better matches the actual distribution through a combination of increased accuracy and more strategic responses.

Table 3: Fairness metrics

		GPT-3.5			GPT-4		
		Temperature			Temperature		
		t = 0	t = 0.5	t = 1	t = 0	t = 0.5	t = 1
Entrepreneurs	DPD	0.010	0.011	0.005	0.049	0.055	0.046
	RCS	0.686	0.707	0.749	0.920	0.918	0.919
Nobel Prize	DPD	0.111	0.156	0.156	0.023	0.014	0.008
	RCS	0.945	0.946	0.930	0.938	0.938	0.938
Actors	DPD	0.092	0.098	0.104	0.010	0.008	0.023
	RCS	0.989	0.988	0.978	0.984	0.984	0.984

The RCS is similar for GPT-3.5 and GPT-4 for the Nobel Prize and Actors lists. This result represents whether the gender distribution in the hallucination rate matches the gender distribution in the underlying data, which is skewed for Nobel Prize winners and balanced for Actors.

4.3 Gender Disparities in Declination

GPT-4 introduced a mechanism to decline to answer the question if the results are unknown. The improved gender performance could be driven by gender differences in declining to answer, so we verify whether the declination rate varies between male and female notable persons. If the declination rates differ between female and male names, then there is evidence that LLMs are less able to identify their own hallucinations for one gender. This effect would mainly occur in the Entrepreneurs task because the miss rate is much higher for this task as compared to the Nobel Prize and Actors datasets. Figure 1 shows that GPT-4 is more likely to decline to answer prompts about the male entrepreneurs than female entrepreneurs. In contrast, GPT-4 is more likely to produce a hallucination for female entrepreneurs than for male entrepreneurs. However, for both male and female entrepreneurs, GPT-4 is more likely to decline to answer than hallucinate.

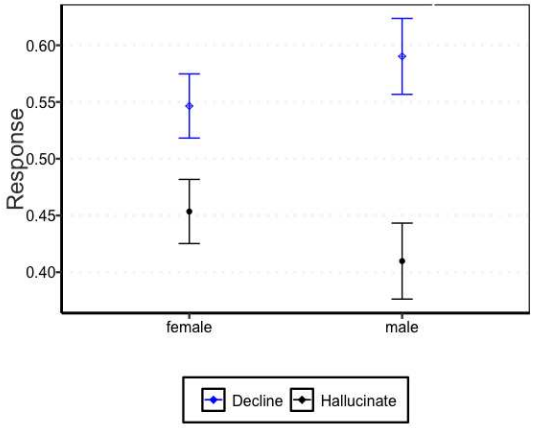


Figure 1: Decline and hallucinations by gender for GPT-4 (Entrepreneurs)

For the Nobel Prize and the Actors lists, GPT-4 only declined to answer when the prompt contained the year 2022, after the training data. GPT-4 only declined to answer for one run of Best Actor in 2022 for temperature = 1 and for most of the runs for Best Actress in 2022. However, the GPT-4 responses were accurate for both Best Actor and Best Actress in 2022 (Will Smith and Jessica Chastain respectively). GPT-4 declined to answer for Nobel Prize winners in 2022 across all subjects. However, there were inaccuracies in the GPT-4 responses for Nobel Prize winners, suggesting that GPT-4 is still unable to identify every inaccurate response and does not always decline to answer when appropriate.

4.4 Gender Distributions in Hallucinations

Next, this paper examines patterns in the hallucinated responses. According to the RCS, GPT-4 exhibits lower gender disparities than GPT-3.5. However, this result conflicts with the more established fairness metric of DPD so, to bolster this evaluation from RCS, this study further probes the gender distribution for the generated output and the degree to which it aligns with the distribution of the task. Table 4 shows the average gender of the returned names (Output column) by gender of the notable figure (Population column). In general, these results suggest that GPT-models tend to produce male names, but this trend decreases with GPT-4. The subsections below will discuss the results for each list.

Table 4: Gender responses across tasks and populations

			GPT-3.5			GPT-4		
Task	Population	Output	t = 0	t = 0.5	t = 1	t = 0	t = 0.5	t = 1
Entrepreneurs	Female	Female	0.406	0.395	0.400	0.502	0.503	0.520
	Female	Male	0.512	0.511	0.480	0.331	0.318	0.289
	Male	Female	0.247	0.250	0.288	0.048	0.047	0.056
	Male	Male	0.675	0.670	0.587	0.764	0.757	0.717
Actors	Female	Female	0.989	0.991	0.965	1	0.993	1
	Female	Male	0.011	0.008	0.035	0	0.008	0
	Male	Female	0.043	0.040	0.043	0.022	0.022	0.022
	Male	Male	0.957	0.960	0.957	0.978	0.978	0.978
Nobel Prize	Female	Female	0.431	0.455	0.412	0.659	0.660	0.655
	Female	Male	0.569	0.545	0.588	0.341	0.340	0.345
	Male	Female	0.043	0.041	0.048	0.014	0.014	0.013
	Male	Male	0.957	0.959	0.952	0.986	0.986	0.987

4.4.1 Entrepreneurs

For the entrepreneurs task, GPT models often returns multiple names rather than a single person in response to the prompt. Despite only a single name being required, GPT tends to overproduce and return multiple names. This response can be mixed with correct and incorrect information.

Table 4 reports the average percentage of female names or male names over the five runs for each notable figure. For this dataset, the GPT models return multiple names for each prompt, and this table reports the average composition of the output. Male names are the most prevalent responses both for female and male founders.[3] For the GPT-4 responses, the gender distribution in the output better reflects the gender of the actual population, partially due to increases in the accuracy.

GPT models tend not to produce multiple female names, reflecting a latent inference that women do not found companies together. As the number of names in the generated output grows, more male names are generated for the additional names, rather than a balance of male and female names. Figure 2 illustrates how the percentage of female names returned decreases as the number of founder names returned increases. As the size of the generated output grows, GPT-3.5 continues to add more male founder names, as shown in Figures 2a and 2b.

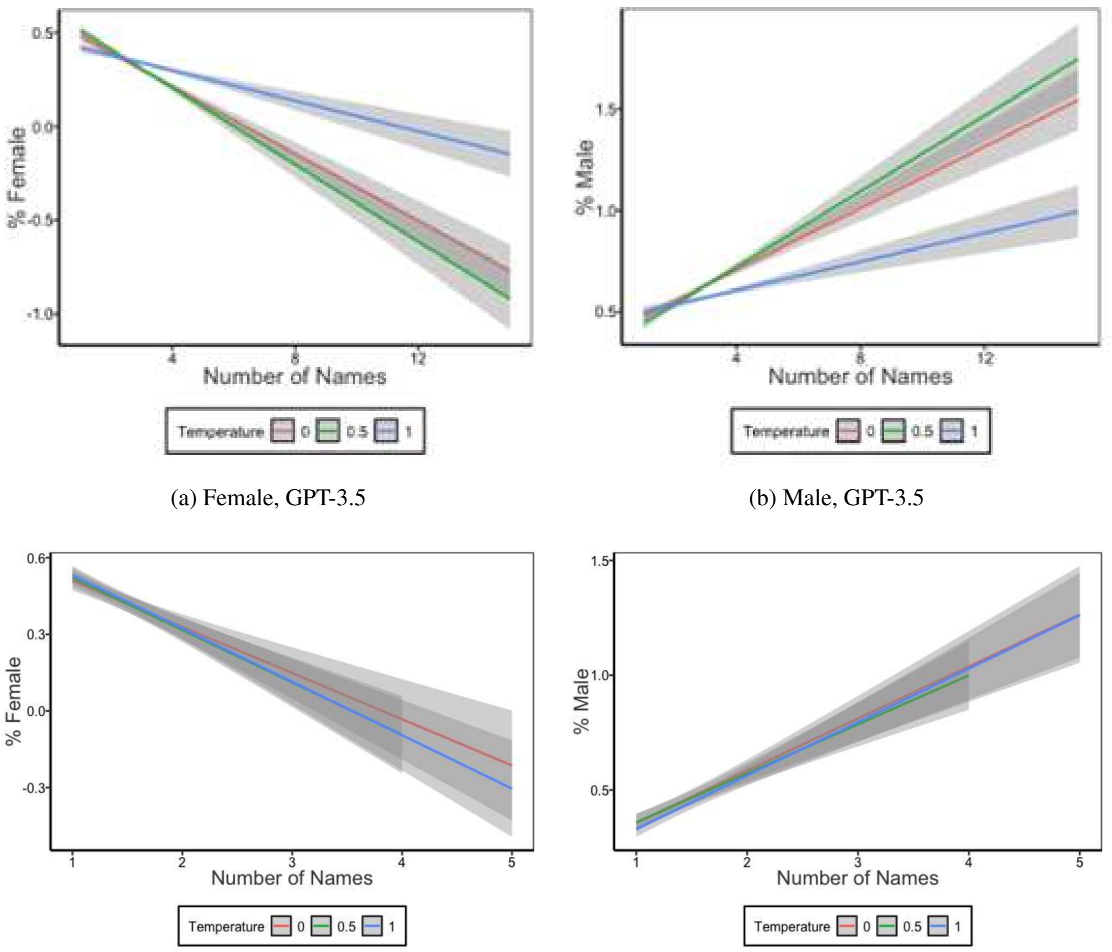
We see that the least creative models consistently put fewer female names together as co-founders. The creative model generates more completely hallucinated names, and the percentage of female founders does not go down as quickly when the set of generated names becomes larger. This result underscores an assumption of homogeneity in the founders group and which names belong together. For instance, the most common names in the Retail industry are ‘David’, ‘Karen’, and ‘John’ whereas the most common names in the Venture Capital industry are ‘Abhishek’, ‘Kiran’, and ‘Prashant’. The set of hallucinated names reflects an assumption of the similarity in teams.

This pattern of homogeneity in returned names persists for GPT-4, although GPT-4 produces fewer names in the responses. As more names are produced, the percentage of female names is still lower (see Figures 2c and 2d). However, the creativity of the model seems to have less influence on these results than in GPT-3.5.

4.4.2 Actors

In the Actors task, both GPT models tend to return a single name rather than a set of names. Table 4 shows the average percent of female or male names returned over the five runs. In general, the responses aligns closely with the gender in the underlying task for both GPT-3.5 and GPT-4. GPT-3.5 produces male names at a rate of 48.6% (temperature = 0) to

[3] Note that LLMs may also produce non-gendered responses.



(a) Female, GPT-3.5

(b) Male, GPT-3.5

(c) Female, GPT-4

(d) Male, GPT-4

Figure 2: Percent of female and male names by the number of names returned

49.5% (temperature = 1), and GPT-4 produces male names at a rate of 48.6% (temperature = 0) to 48.9% (temperature = 1) overall. The list of notable figures is nearly 50% men and women (with one dual winner for Best Actress in the 1960s.) Despite the gender indication in the prompt, in rare cases GPT-3.5 hallucinates a name of a different gender than the prompt would suggest. For instance, GPT-3.5 returned Fredric March as the winner of the 1932 Best Actress award. For GPT-4, any deviation from the factual response was a declination to answer.

4.4.3 Nobel Prize

For the Nobel Prize task, multiple people can win the award in a single year, and the models return a set of multiple names for each year and subject. Table 4 shows the average percentage of female and male names returned over the five runs for each award. The "Female" Population indicates if a woman was in the group of prize winners for that year and subject, and the "Male" population is if the winning groups was entirely male.

As displayed in Table 4, the majority of the LLM-responses (correct and hallucinated) were male names for both models, ranging from 92.8% (temperature = 0) to 92.1% (temperature = 1) for GPT-3.5 and ranging from 93.3% to 93.4% of all names for the GPT-4 model. Of only hallucinated names, GPT-3.5 returns 92.6% (temperature=0) to 92.3 (temperature = 1) male names, whereas GPT-4 returns 85% (temperature = 0) to 85.5% (temperature = 1) male names. These numbers suggest that the GPT-4 generates female suggestions at twice the rate of their prevalence in the population of Nobel Prize winners, highlighting the dual consequence of increased performance yet evidence of gender disparities. However, the improved performance leads to significantly fewer hallucinated names so the overall result is lower gender disparities in the overall factuality of GPT-4.

4.4.4 Summary

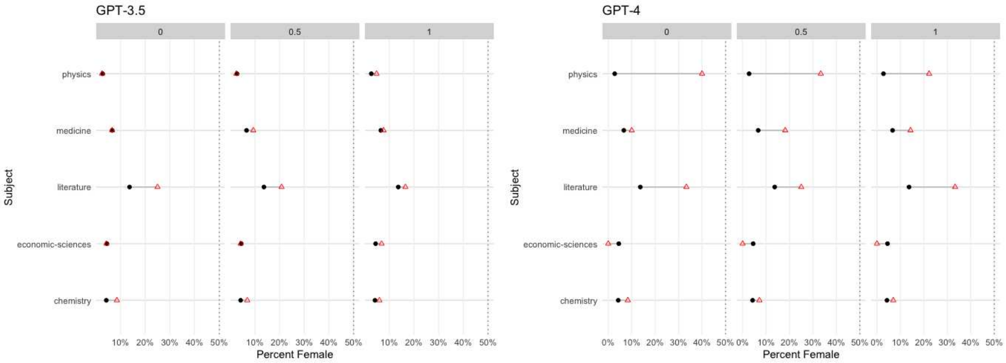
The evaluation of GPT models' outputs across these different tasks suggest that male names are more likely to be returned, particularly in the heavily hallucinatory context of the Entrepreneurs task. As the number of names returned grows, the set is more likely to be homogeneous.

4.5 Gender Associations in Prompt

Previous work has shown gendered associated of words based on the distance between that word and gendered terms in their word embeddings. In this context, some of the words in the prompt, like the names of companies and industries or the prize subject, may be closely aligned with gender. These gendered words could influence what names are generated and explain the gender disparities in the results for the entrepreneurship and Nobel Prize notable persons. Entrepreneurs in particular are less well-known so the context clues from the prompt – the company name and industry – may be more likely to shape the results. Nobel Prizes are known to be skewed by the subject, so the subject may influence the gender rates in the hallucinations.

4.5.1 Nobel Prize Subject

The gender patterns differ markedly for the heavily-skewed Nobel Prize task. Figure 3 shows the difference between the true percentage of gender (black circle) and the percentage of incorrectly generated names (red triangle) by subject. Women are generated at a higher rate than they are truly present in the winner distribution for every subject. Additionally, GPT models are clearly more likely to hallucinate female names for Literature. Figure 3a shows that the overall percent of female names produced for GPT-3.5 are similar for most subjects except Literature. Although the female percentages are greater for GPT-4 in Physics, as shown in Figure 3b, this result is driven by its hallucination of a single inaccurate female name (Marie Curie), which accounts for the higher percentage of incorrect names in the better performing GPT-4 model.



(a) GPT-3.5

(b) GPT-4

Figure 3: Gender percentage by Nobel Prize subject

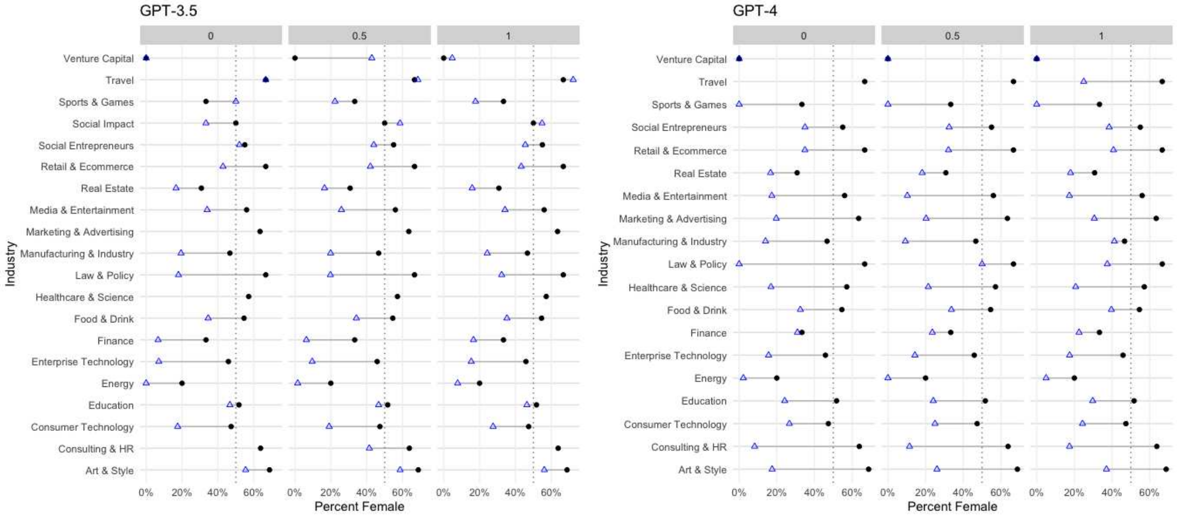
4.5.2 Industry

There are clear connections between industries and stereotypical names, but in the opposite direction for the Entrepreneurs task. Women are generated at a lower rate in generated responses compared to their actual prevalence across all industries. Figure 4 compares the average percentage of females in the dataset to the average percentage of hallucinated names. The most heavily male industries, such as Finance, Venture Capital, and Energy, generate the most male-dominated names. For Law and Policy, female names represent more than 50% of the entrepreneurs and yet only 20% of the returned names are associated with women for both GPT-models. It is clear that most hallucinations are less female than the underlying dataset would suggest.

These results underscore the potential for latent gender associations for industries and company names that could influence the LLM-responses.

4.5.3 Gender Association of Industry with Word Vectors

Another way to examine how the industry affects the gendered output is to find the gender association with the industry. We use the pre-trained GloVe model to represent the words in the industry [4]. We calculate the gender associations

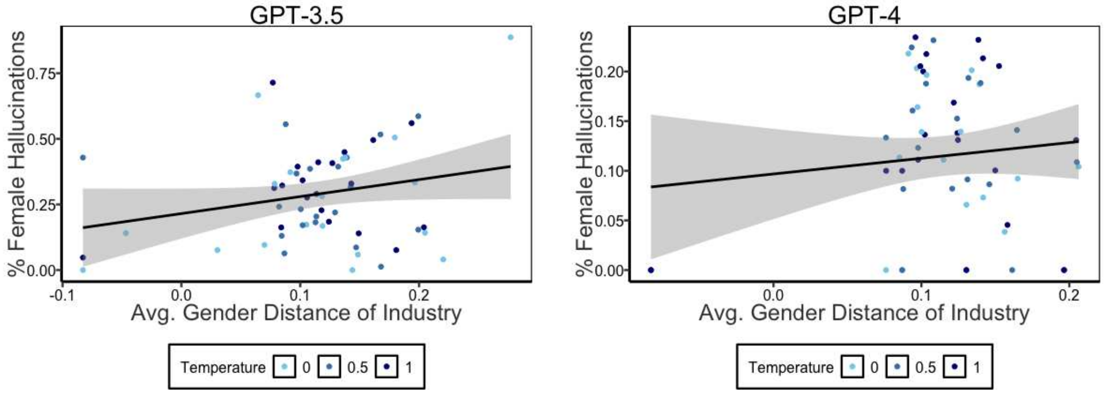


(a) GPT-3.5

(b) GPT-4

Figure 4: Gender percentage by entrepreneurs' industry

between the industry word vector and the female word vector as described in Section 3. The average percent of female hallucinations is determined for each industry and temperature. The results are displayed on Figure 5, which shows the relationship between the gender associations of the industry word embeddings and the percentage of female names hallucinated. For GPT-3.5, there is a clear positive correlation between the industry's simiarity to the female embedding and the gender of the name return, as shown in Figure 5a. The relationship between the industry word embedding and the percent female hallucinations is less strong for GPT-4, suggesting a move away from these stereotypical behaviors, as displayed in Figure 5b.



(a) GPT-3.5

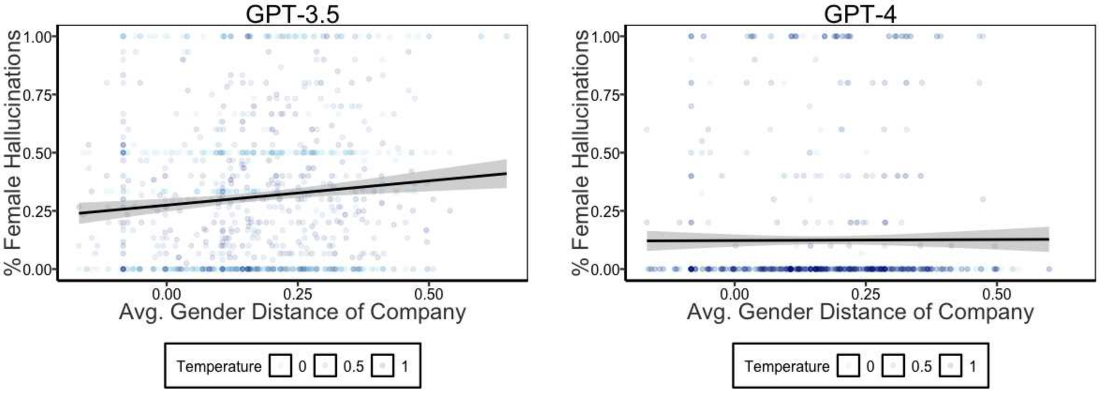
(b) GPT-4

Figure 5: Gender associations of Industry and female hallucinations

4.5.4 Gender Association of Company with Word Vectors

The company name may also contain gendered words; however, those words should be less relevant to the gender of the hallucinated names. Gender differences associated with the industry may reflect the underlying distribution of the industry, whereas the company name merely reflects the preferences of the founders. To explore the gender implications of the company name and industry, we measure the gender distance of the company name and industry and then compare that distance to the hallucination rates of female names. The methods are the same as in the preceding section, except that the average female hallucination rate is calculated for each company name and each temperature. Figure 6 shows the relationship between the gender associations and the percent of female hallucinations for GPT-3.5

and GPT-4. These results confirm our earlier analysis. The more a company or industry's name is associated with female, the more likely that GPT-3.5 will return a female hallucination. With the performance change in GPT-4, this correlation is interrupted and the gender associations of the company name and industry do not affect the likelihood of returning female hallucinations. This supports our findings from the RCS that GPT-4's responses better reflect the underlying name distribution in the dataset, and that they are trained to avoid stereotypical behavior.



(a) GPT-3.5

(b) GPT-4

Figure 6: Gender associations of company name and female hallucinations

5 Discussion and Conclusion

5.1 Summary

This paper evaluates the factuality of GPT models' responses to prompts with specific answers for notable people. Using three specific tasks, we evaluate the patterns in recall across GPT models. Gender disparities of GPT-3.5 occurred in ways that favored the more prominent group; GPT-3.5 better recalled male actors and female Nobel Prize winners. Unsurprisingly, GPT-4 outperformed GPT-3.5 in overall recall of notable persons but the increased performance of GPT-4 is insufficient to eliminate all disparities. The performance increase for GPT-4 was not evenly distributed, and the gender disparity was statistically significant according to the fairness metric DPD.

To better evaluate the fairness of the models, we introduce a new evaluation measure, Response Concentration Score (RCS). RCS calculates how representative the distribution of LLM responses is to the actual distribution of answers. According to this measure, GPT-4 exhibits little gender disparities and exhibits fairness performance better or on par with GPT-3.5 across all three tasks. A higher RCS indicates increased performance and how representative the hallucination rates are to the underlying distribution.

Because more recent versions of GPT can decline to answer when the LLM is uncertain about the response, the paper probes whether the LLM identifies its uncertainly about the answer differently for male and female persons. If an LLM produces a hallucination, rather than declining, it appears unawareness of its limitations. This paper also evaluates the gender disparities in the declination, finding that GPT-4 is more likely to decline to answer for male notable persons than female notable persons. However, declination is more common than hallucination for both male and female figures.

This paper also explores patterns in the co-occurrence of gendered names in hallucinated responses. Male names are more common as more names are hallucinated, indicating homogeneity in named hallucinations. Hallucinated gendered names likely emerge as a property of the words in the prompts, with more female-associated words leading to more hallucinations of female names. To probe the mechanism behind these results, we examine the gender associations in the prompt with the responses. The prompt contains the company name and industry for entrepreneurs and the prize subject for Nobel Prize winners. The company names most closely aligned with female gender are more likely to receive female hallucinations in GPT-3.5 but not in GPT-4, suggesting that the de-biasing procedures for GPT-4 resulted in less gendered response patterns.

Lastly, we found that there are also patterns in who will co-occur in the hallucinated responses in both GPT models. When they return larger sets of names, this set is more likely to be homogenous.

5.2 Theoretical implications

It is critical that LLMs are evaluated to ensure their alignment with human values [39]. In line with values of equality, LLMs should not exhibit gender disparities in their ability to produce factually accurate responses for notable persons. Some may believe that performance improvements may be sufficient to address gender disparities; however, improvement in the LLMs factuality may be unevenly distributed and gender disparities can still emerge. This paper supports the need for ongoing research into the disparities, biases, and ethics of LLMs even as the performance of these models increases.

To that end, this paper suggests considering the gender distribution of the responses and the actual data through the Response Concentration Score (RCS). As researchers grapple with the meaning of gender disparities in a biased world [33, 18, 15], it is critical to create metrics to compare the deviations of responses from the baseline, balancing the increased performance with the distribution in the hallucinations.

This paper also underscores how the tension between creativity and accuracy spills into gender disparities. Temperature can reduce gender disparities by encouraging "creativity" and randomness. More randomness can lead to better gender balance because the model produces more lower-probability answers; however, this approach to creativity also yields a lower identification rate.

More digital traces also produces better gender balance overall but due to the focused attention on a handful of hallucinations and not the broader representation of hallucinated female names. Representation is insufficient. Multiple suggestions to de-bias data include changing the training data, the training process, or model in order to ensure that mentions of occupations lead to an equal representation of genders [34]. Representation in the true data will not achieve parity.

5.3 Limitations

This paper focuses on GPT models, and future research could expand to examine more models. Furthermore, this paper focuses on three specific tasks with different characteristics to evaluate how the nature of the tasks influences the gender disparities in the results. Future research can expand to additional names and metrics using the RCS as the baseline. One notable limitation of research involving LLMs is the dynamic nature of their continuous updates. As newer versions of LLMs are released, the specific results obtained from experiments with a particular model are susceptible to obsolescence. Another limitation is the focus on two genders. The paper looks at the gender implications associated with male and female names, but future research could examine to including non-binary and other genders in the lists of notable persons.

5.4 Conclusion

Creating appropriate evaluation metrics for LLMs is of paramount importance. It is critical to explore the gender disparities in LLMs given the known gender disparities in the presence of notable persons on the Internet. By using fairness metrics and probing the connection between the gender associations in the prompt and the LLM response, researchers can evaluate whether LLMs exhibit gender disparities in their factuality.

Acknowledgments

This research was funded by Research Foundation—Flanders grant number 11N7723N.

References

[1] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*, 2020.

[2] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, jan 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL `https://doi.org/10.1145/3641289`. Just Accepted.

[3] Joseph Reagle and Lauren Rhue. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21, 2011.

[4] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[8] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation, 2022.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[10] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[11] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454. URL `https://aclanthology.org/2022.tacl-1.11`.

[12] Pouya Pezeshkpour. Measuring and modifying factual knowledge in large language models, 2023.

[13] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.383. URL `https://aclanthology.org/2021.naacl-main.383`.

[14] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[15] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

[16] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[19] Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. Probing explicit and implicit gender bias through llm conditional text generation. *arXiv preprint arXiv:2311.00306*, 2023.

[20] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

[21] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study, 2023.

[22] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation, 2023.

[23] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021. doi: 10.1145/3442188.3445924. URL `http://dx.doi.org/10.1145/3442188.3445924`.

[24] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*, 2023.

[25] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[26] Ruomeng Cui, Jun Li, and Dennis J Zhang. Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science*, 66(3):1071–1094, 2020.

[27] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics*, 9(2):1–22, 2017.

[28] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[29] Lauren Rhue and Jessica Clark. Who are you and what are you selling? creatorbased and product-based racial cues in crowdfunding. *MIS Quarterly*, 46(4), 2022.

[30] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[31] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[32] Debora Nozza, Federcio Bianchi, Dirk Hovy, et al. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, 2022.

[33] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

[34] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[35] Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 289–298, 2019.

[36] William M Landes and Richard A Posner. Citations, age, fame, and the web. *The Journal of Legal Studies*, 29 (S1):319–344, 2000.

[37] Cameron Blevins and Lincoln A. Mullen. Jane, john ... leslie? a historical method for algorithmic gender prediction. *Digit. Humanit. Q.*, 9, 2015. URL `https://api.semanticscholar.org/CorpusID:38649139`.

[38] Yifan Hu, Changwei Hu, Thanh Tran, Tejaswi Kasturi, Elizabeth Joseph, and Matt Gillingham. What's in a name?–gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 35(4):1537–1563, 2021.

[39] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023.