Gender Bias, Recency and Recall in Large Language Models: Which Scientists and Movie Stars Does ChatGPT Forget?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used as a tool to access factual information. However, when prompted to answer factual questions LLMs frequently generate incorrect "hallucinated" responses, thus displaying imperfect recall. Given the known gender biases in LLMs, we study the prevalence of gender-based disparities in LLM responses to factual questions. Specifically, we examine the degree to which ChatGPT exhibits gender-based differences in recall for Noble Prize winners and Oscar award recipients. Our results confirm that there are gender-based differences in recall, but that the level of bias varies significantly with both subject matter factors like recency or prominence and model parameters like creativity.

1 Introduction

Large Language Models (LLMs) are increasingly used in lieu of search engines for information retrieval. Although trained on a large fraction of all available human knowledge and digital traces, it is well know that LLMs frequently "hallucinate" incorrect responses (Mittelstadt et al., 2023), posing challenges to their reliability when tasked with retrieving factual knowledge.

A less studied concern is of potential *bias* in LLM fact retrieval. As LLMs learn language, stereotypical gender-biased associations may emerge from the aspects of our language and society reflected in the digital traces they are trained on (Nadeem et al., 2020). The presence of this bias (Abid et al., 2021; Nadeem et al., 2020) raises a new trade-off between fidelity and stereotype reinforcement: should LLMs preserve the bias of the reality reflected in the their training data (Ferrara, 2023) or be tuned to generate content aligned with a society's aspirations? (Vig et al., 2020)

Recent studies of bias in LLMs add to a growing body of research about fairness in machine learning (Barocas et al., 2017; Chouldechova and Roth, 2018; Kearns et al., 2018; Mehrabi et al., 2021). Dong et al. (2023) probe LLMs for explicit and implicit bias by using conditional text generation, while Wan et al. (2023) analyze bias in LLM-generated reference letters. However, there is limited understanding about whether LLMs retreive and *recall* information in gender-biased ways, a topic we study in this paper.

Specifically, we probe the specific nature of gendered recall of LLMs when asked about notable figures. We study the following research questions:

- Does the examined LLM exhibit different recall patterns for male and female notable figures?
- What other factors (such as prominence, recency, context, creativity) affect gender differences in the recall of notable figures?

We use two prominent and publicly available sets of notable figures: Nobel Prize winners and Oscar awards recipients. Our analysis reveals a discernible gender bias within LLM recall after accounting for prominence of the figures. Individuals from further in the past are more prone to be forgotten, a phenomenon observed across both the Nobel Prize and Oscar award contexts, and consistent with the *recency effect*, a human cognitive bias in which items from the recent past are remembered more clearly. Furthermore, a higher creativity setting on the LLM (the default model) degrades performance on recall even as it produces less pronounced gender disparities.

2 Data and Methods

We utilize two distinct datasets of notable figures, the list of Nobel Prize winners and the list of Oscar winners for the Best Actor and Best Actress awards. These datasets both contain notable figures across genders and include a well-defined time component

associated with each figure. The Oscar awards (almost) always contain a single winner for each year, while the Nobel prizes often have multiple winners for one year and subject. We generate each prompt five times. We use the default DaVinci-003 engine (GPT-3, OpenAI) for our experiments, and these experiments are within the terms of use.

2.1 Data

2.1.1 Oscar Awards

We use a list of Oscar Award winners containing Year, Movie, Role and their names. For each Oscar winner, we prompted ChatGPT in the format: "Who won the Oscar for best leading role as actor awarded in the year [*Year*] for his/her movies in [*Year - 1*]? Just return the name without any text" and "Who won the Oscar for best leading role as actress awarded in the year [*Year*] for his/her movies in [*Year - 1*]?".

2.1.2 Nobel Prizes

We use a list of Nobel Prize winners containing Year, Subject, Discovery, and their names. For each Nobel Prize winner, we posed a query to ChatGPT in the format: "Who won the Nobel Prize for [*Subject*] in [*Year*]? Return the names in a list like this: Name1, Name2,.. Name n".

2.2 LLM parameters

The temperature parameter controls the diversity of generated text. We try different versions of the temperature: 0, 0.5 and 1 (0.5 is the default). Higher temperature values 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. For the other parameters, we use the default settings.

2.3 Recall

We determine whether the LLM-generated names are correct by comparing the last names of the generated names with the notable figures' last names. We define *Recall* as the percentage of instances that are correctly identified by the LLM.

2.4 Gender

To determine the gender of each winner, we calculate a gender probability based on the list of baby names by gender by the Social Security Administration (Karimi et al., 2016). [1] We assign a probability of being female based on the percentage of people with this first name that are born female.

2.5 Prominence

To understand the effect of the Nobel Prize winners' prominence, 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 and the word "winner". Search counts have been used for many years as a proxy for the current prominence of public figures (Landes and Posner, 2000).

3 Results

The results provide evidence of gender-biased recall of notable figures.

3.1 Oscar Winners

We query the LLM five times for each of the temperature values, and we report the results. Figure 1 illustrates the recall percentage for male and female notable figures. At every temperature, the recall is lower for women than for men. We also note that the recall is higher for lower values of the temperature (less creativity). Additionally, Figure 1 illustrates that a lower temperature in the Oscar award analysis correlates with improved recall. This recurrent pattern underscores the temperature sensitivity of the language model's responses, suggesting that the recall performance is influenced by the degree of randomness introduced during generation.

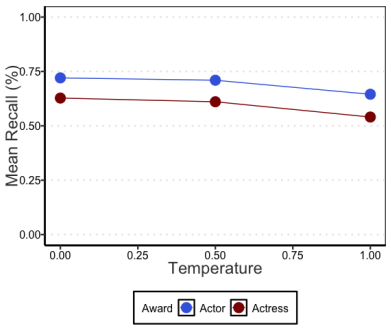


Figure 1: Gender disparity in recall for the Oscar awards

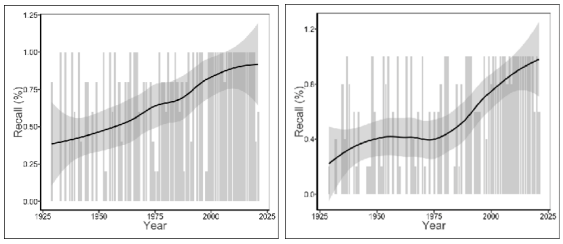
Next, we estimate a logistic regression to probe the relationship between recall, gender, recency and prominence. Our dependent variable is *Recall*,

[1] Our study employs a binary gender distribution, examining gender using gender assigned at birth. However, it is essential to acknowledge that this approach does not encompass the entirety of the gender spectrum.

	Dependent variable = Pr (recall)					
	Temp = 0		Temp =0.5		Temp =1	
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Female	**-0.385**** (0.142)	**-0.502**** (0.155)	**-0.402**** (0.141)	**-0.518***** (0.153)	**-0.388**** (0.135)	**-0.504***** (0.148)
Unknown	0.174 (0.527)	-0.460 (0.582)	0.230 (0.527)	-0.378 (0.579)	0.278 (0.498)	-0.318 (0.549)
Year		**0.036***** (0.003)		**0.034***** (0.003)		**0.034***** (0.003)
Constant	0.924*** (0.106)	-68.985*** (6.227)	0.869*** (0.104)	-66.969*** (6.103)	0.569*** (0.099)	-65.874*** (5.839)
Observations	935	935	935	935	935	935
Log Likelihood	-586.430	-510.887	-594.937	-521.631	-627.361	-551.138
AIC	1,178.860	1,029.773	1,195.874	1,051.262	1,260.721	1,110.276

Table 1: Recall of the Oscar awards

and our independent variables are *Gender* and *Year*. The results, depiced in Table 1, provide additional support for the observed trend in gender-based recall disparities, with a lower recall for Best Actress award winners compared to Best Actor award winners. Furthermore, the significant influence of the *Year* variable, indicating a decreased recall for awards won further in the past. This effect is consistent with the recency effect. Figure 2 illustrates this recency effect for both male and female actors, visually portraying the correlation between the year of the Oscar Award and the observed increase in recall.



(a) Actors (b) Actresses

Figure 2: Recency effect for Oscar winners

3.2 Nobel Prize Winners

Figure 3 illustrates the average recall across all scientists as well as the recall percentage for male, female, and scientists of unknown birth gender. In contrast with the results for movie stars, at every temperature level, the recall is higher for *female* Nobel Prize winners than the overall average recall, which is shaped by the recall percentage for male winners given that a vast majority of winners are male. We also find that the recall is higher for lower values of the temperature (less creativity), which is consistent with our findings for Oscar award winners.

There is noticeable time variation in the level of recall, as depicted in Figure 4. This phenomenon aligns with the recognized *recency effect*. These

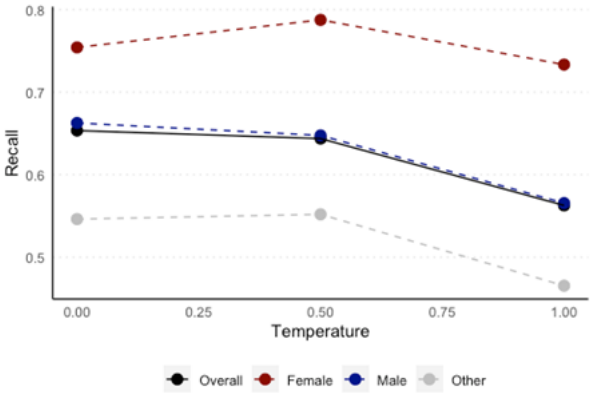


Figure 3: Gender disparity in recall for Nobel Prize winners

discussed findings remain consistent across all temperature settings. It may be caused by the dramatic increase in published content about prominent figures in recent years driven by the Internet and social media that leads to substantially greater volumes of LLM training data about figures whose fame is more recent.

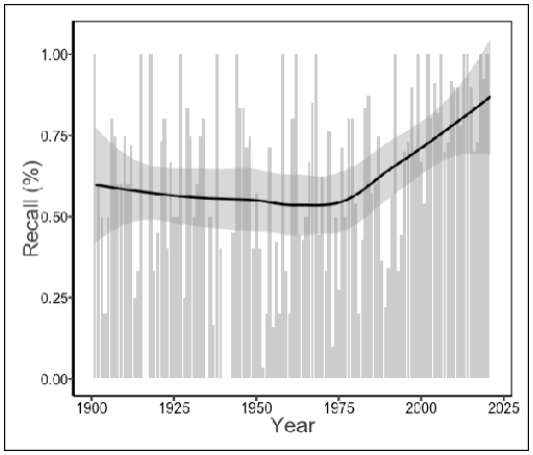


Figure 4: Recency effect in recall for Nobel Prize winners

Following our earlier sequence of analysis, we estimate a logistic regression to probe the relation-

	Dependent variable = Pr (recall)					
	Temp = 0		Temp = 0.5		Temp = 1	
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Gender – Female	**0.444**** (0.154)	**-5.894***** (1.618)	**0.699***** (0.162)	**-5.072**** (1.578)	**0.743***** (0.150)	**-2.314** (1.189)
Gender – Other	-0.492*** (0.095)	-0.268 (0.580)	-0.402*** (0.095)	-0.191 (0.556)	-0.408*** (0.094)	0.186 (0.537)
Log(Prominence)		**0.185***** (0.024)		**0.129***** (0.024)		**0.134***** (0.023)
Year		**0.006***** (0.001)		**0.007***** (0.001)		**0.007***** (0.001)
Female * Prominence		**0.477***** (0.132)		**0.437***** (0.129)		**0.216*** (0.094)
Other * Prominence		-0.023 (0.050)		-0.024 (0.048)		-0.057 (0.046)
Constant	0.677*** (0.036)	-12.449*** (2.076)	0.611*** (0.036)	-15.238*** (2.053)	0.269*** (0.034)	-15.138*** (1.993)
Subject		Yes		Yes		Yes
Observations	4,230	4,230	4,230	4,230	4,230	4,230
Log Likelihood	-2,709.223	-2,607.869	-2,732.405	-2,647.891	-2,872.474	-2,790.652
Akaike Inf. Crit.	5,424.446	5,237.738	5,470.809	5,317.782	5,750.948	5,603.304

Table 2: Recall of the Nobel Prize Winners

ship between recall, gender, recency and prominence. Our dependent variable is *Recall* and the independent variables are *Gender*, *Year*, and *Prominence*. Table 2 reports the estimates. This table confirms that the language model exhibits a higher likelihood of recalling female Nobel Prize winners without controlling for additional factors, consistent with Figure 3. However, female Nobel Prize winners are more prominent on average, perhaps because of their relative rarity, leaving more digital traces due to their rarity. As shown in Table 2, the LLM demonstrates a **lower** likelihood of recalling female winners once we control for prominence (Columns 2,4,6). One plausible explanation could be the relative scarcity of female winners, making them stand out and consequently more likely to be discussed and included in online data.

The findings from the analysis of Nobel Prize award winners corroborate and validate the conclusions drawn from the Oscars. The consistent pattern of results across fairly distinct contexts reflecting different drivers of accomplishment, different levels of fame and a very different underlying gender distribution strengthen the argument that gender-based recall differences and temporal biases are a recurring shortcoming in this language model's responses.

4 Discussion

This study shows several factors that affect bias and recall of LLMs. First, we found evidence of gender disparities in the recall of prominent figures. Female notable figures are less likely to be recalled, especially when controlling for their prominence. Limiting gender disparities should be an important consideration as LLMs are incorporated into information retrieval interfaces.

Second, more distant historical figures are more likely to be forgotten. This recency effect could lead to challenges in maintaining historical accuracy and could impact public perception. Our findings underscore the challenges of using the convenience sample of the Internet to train language models, as the Internet is naturally biased towards more recent accomplishments.

Third, model creativity influences recall. This result suggests that lower temperatures are preferable for recall-based tasks, even though the default temperature for the consumer-facing web interface may be higher. More randomness decreases the LLM's recall but does not affect the gender disparities of recall.

Finally, prominence affects recall. For rare historical figures, such as a female Nobel Prize winners, the increased celebration of their accomplishments likely increases the likelihood of being recalled. These anti-stereotypical figures are outliers, and therefore more likely to be recalled as women. This finding underscores how the relationship between gender and recall is not linear but is related to stereotypes and digital traces of historical figures' accomplishments. This could yield a superstar effect for knowledge, as more recent and more popular knowledge is more likely to be reproduced, and more obscure facts and people fade.

An interesting avenue of future research would be to analyze the "misremembered" facts or *hallucinations* produced by the LLM. These hallucinations provide insights into the underlying beliefs of the models. In future research, we will probe the gendered output of hallucinations and examine the influence of prominence, recency and creativity.

As more LLMs are incorporated into information retrieval platforms such as Internet search, these findings underscore the importance of ongoing research and improvements in LLMs to ensure they do not perpetuate biases or inaccuracies.

5 Limitations

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 is susceptible to obsolescence. However, generalized insights generated through this study can transcend the specific version of the used LLM, offering enduring value despite the rapid evolution of these language models. The pattern identified by this study is applicable to any LLM and raises concerns about integrating LLMs into information retrieval tools without full testing. These patterns of lower female recall may persist until an LLM achieves perfect recall for all notable figures across all genders. This lofty standard is likely unachievable because popular information sources on the Internet, such as Wikipedia, are known to exhibit gender disparities in their representation (Reagle and Rhue, 2011).

Acknowledgements

References

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

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

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

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

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

Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, pages 53–54.

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

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

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

Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use llms as zero-shot translators. *Nature Human Behaviour*, pages 1–3.

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

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

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

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. *arXiv preprint arXiv:2310.09219*.