Investigating Homophily Bias in Two Large Language Models

Anonymous ACL submission

Abstract

The use of generative large language models (LLMs) has been spiraling in recent years. Thanks to improvements in training regimes, these models produce fluent text and interact with humans in an unprecedented way. Consequently, researchers have begun investigating the "cognitive" abilities and biases of LLMs. One cognitive bias which is particularly interesting for interaction is homophily. In this work, we analyze two popular models (*Llama 3.2 3B* and *GPT-4o-mini*) to assess the degree of homophily across nine different human attributes, accounting for two other cognitive biases, namely framing and order bias. Our findings suggest that, while *Llama 3.2 3B* exhibit traces of framing and order bias, *GPT-4o-mini* exhibits homophilic bias, particularly with respect to political view and personality type. This has significant implications for echo chambers, disinformation dissemination, and social polarization in AI systems that utilize LLMs. Our results highlight the need for rigorous investigations into homophily to ensure responsible AI deployment.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in processing and generating natural language. Their rapid advancement in recent years has paved the way for their adoption in recommender systems (Kim et al., 2024), coding assistants (GitHub Copilot; JetBrains) and conversational search engines (Perplexity AI; OpenAI). Their increasing adoption and fluency in interaction with humans has triggered research on LLMs' behaviors and biases, including cognitive biases. Cognitive bias is defined as a systematic pattern of deviation from the norm or rationality in judgment (Daniel, 2017; Dobelli, 2013; Eberhardt, 2020). Since LLMs often interact with diverse user bases, it is crucial to ensure that they produce fair and impartial responses to avoid reinforcing harmful stereotypes or perpetuating discrimination. As humans are prone to various cognitive biases, the data we create and the feedback we generate, which are then used in their training, are likely to cause similar biases in LLMs. The presence of such biases in LLMs may influence their decision-making, responses, and overall interactions with users (Sumita et al., 2024).

Although the study of cognitive biases is increasing as these models demonstrate more human-like text generation, the study of cognitive biases is mostly focused on anchoring effect (Talboy and Fuller, 2023; Lou, 2024; Echterhoff et al., 2022), framing (Bian et al., 2024; Suri et al., 2023), or heuristics (Jones and Steinhardt, 2022). Homophily, the tendency to associate with similar others (McPherson et al., 2001), is a relevant cognitive bias that can appear in conversational LLMs and cause communication failure. Although homophily has been extensively studied in human networks (Ertug et al., 2022; Kossinets and Watts, 2009), its presence in LLMs remains understudied. In the study at hand, we aim to investigate homophily while considering the influence of other biases, such as framing (Echterhoff et al., 2022) and order bias (Sumita et al., 2024). Our goal is to provide insights on how characteristics of a person can influence judgment of the machine. This investigation raises critical concerns about the emergence of echo chambers (Jamieson and Cappella, 2008), bias propagation, and fairness in systems that rely on LLMs.

Practically, we construct a synthetic dataset of personas with nine distinct attributes and let the LLM act as a judge to determine whether two personas would form a relationship based on their attributes. By varying individual attributes in personas, we measure changes in the LLM's responses to quantify the effect of each attribute on relationship formation. Our key contributions lie in designing the evaluation framework to assess homophily

in LLMs and ensuring the robustness of our observations by accounting for framing and order biases that arise during prompting. By incorporating these biases into our analysis, we provide a better understanding of bias in LLMs and its implications for fairness and user interactions.

2 Related Work

As LLMs become increasingly integrated into tools like virtual assistants, their susceptibility to create biased outputs and the techniques to mitigate such biases have garnered significant attention. In this domain, Sumita et al. (2024) provide an overview of the current state of research on cognitive biases, particularly order bias, compassion fade, the bandwagon effect, and attentional bias. They also explore mitigation techniques for reducing bias in LLMs. A comprehensive study by Malberg et al. (2024) evaluated 30 different cognitive biases across 20 state-of-the-art LLMs, confirming the presence of such biases in these models. Additionally, Itzhak et al. (2024) examined the effects of instruction tuning and RLHF on LLMs, revealing that these methods introduced biases such as the decoy effect, certainty bias, and belief bias.

Beyond these biases, LLMs also exhibit human-like social dynamics in network formation. Papachristou and Yuan (2024) demonstrated that LLMs adhere to key social principles, such as preferential attachment (Barabási and Albert, 1999), triadic closure (Granovetter, 1973), and homophily (McPherson et al., 2001), resulting in networks with distinct communities and small-world characteristics. Similarly, Chang et al. (2024) found that while LLMs can generate realistic synthetic social networks, they were systematically prone to overestimating political homophily.

Previous works studied homophily in LLMs by analyzing emergent network formation patterns. What distinguishes our work is our methodology: we record the token probabilities that indicate relationship formation between persona pairs (Roccas et al., 2002) derived from psychological foundation and evaluate how modifications to persona attributes influence the model's judgments. By varying our prompts we also take into consideration the effect of other biases such as framing and order bias.

3 Methodology

Our overall methodology consists of prompting an LLM to analyze two personas and assess their likelihood of forming a relationship based on personal attributes. We focus on friendship as a neutral social context where compatibility plays a key role. By measuring probability shifts in the LLM's responses when each attribute is equalized, we quantify the influence of that specific attribute on relationship formation. To systematically evaluate these effects, we use a dataset of 25 unique personas spanning nine attribute categories. Our prompts are structured with four key components, described below, to inform the model about the task, guide its reasoning, and enhance its decision-making.

Personas Dataset: Following Chang et al. (2024), we create a synthetic dataset consisting of 25 unique personas with nine attributes: sex, age, occupation, political view, openness, conscientiousness, extroversion, agreeableness, and neuroticism. In our experiments, we consider the following attribute values: sex is binary ("Male" or "Female"), age ranges from 0 to 100, and occupations are uniquely assigned to each persona from a set of 25 occupations. There are 25 different occupations (e.g., Scientist, Architect, Software Engineer, Fashion Designer, Police Officer). Political view has three categories: "Right," "Center," or "Left." Personality traits: openness, conscientiousness, extroversion, agreeableness, and neuroticism, are simplified to "Low" or "High" to reduce the number of personas.

Prompt Template: Our prompts contain four components to guide the model to adequate responses: **Context**, which provides a role description and background information to clarify the task; **Task Description**, instructs the model to analyze two personas and their attributes to determine if a relationship forms, outputting "0" for No and "1" for Yes; **Few-Shot Examples (FSE)**, which offer examples to guide the model's output, with attribute values masked by generic placeholders to prevent bias; and **Persona Descriptions**, which include persona attributes in JSON format.

Procedure: To measure the bias the LLM is prompted with the instructions described above along two distinct personas, namely persona A and persona B from the dataset (600 pairs). The goal is to measure how matching of different attributes between two personas effect the likelihood of the

LLM deciding whether these two personas form a relationship (Output "1"). To collect the probability, the logit of the model for output "0" and "1" is passed through a *softmax* function to obtain a normalized probability distribution corresponding to negative and positive response. We record these probabilities as the ground probabilities $p_{\alpha_{ground}}$ for the particular persona pair (α) . We then search for dissimilar attributes in pair of personas (α_k) where k indicates the target dissimilar attribute. By equalizing attribute α_k from Persona B in favor of Persona A and re-prompting we measure how this change effects the new decision of the LLM.

Since the decision is binary, we are only interested in the probability of token "1" to measure the probability shift, representing a positive response. Thus, the change associated with equalizing the attribute α_k is calculated as the difference in the probability of token "1" before and after equalization:

$$\Delta_{\alpha_k} = p_{\alpha_k} - p_{\alpha_{\text{ground}}}$$

With this formula, we can calculate the amount of probability shift as well as direction of the shift. Positive shift indicates that personas are more likely to form a bond and negative shift indicates equalizing α_k had negative effect on relationship formation. We repeat this procedure for every dissimilar attribute in a single pair and for every unique persona pairs (600 pairs, 3451 combinations of dissimilar attributes between personas).

We also account for framing and order bias. These biases appear in the prompts themselves influencing the decisions of the model based on the order of prompts or framing of the task. We ran our experiments with 3 different variations of the prompts (detailed in Section 4) and observe how the decision of the model for the same two personas ground truth changes as we vary the prompt with the same semantics:

$$\Delta_v = |p^{(v1)}_{\alpha_{ground}} - p^{(v2)}_{\alpha_{ground}}|,$$

where $v1$ and $v2$ are the two versions of the variation type v in the prompts.

4 Experiments

LLM Configurations: We evaluated an open-source model and a proprietary model in our experiments, respectively, *Llama 3.2 3B* (Meta Platforms, 2024) and a bigger model *GPT-4o mini* (OpenAI, 2024). We chose these models because they belong to the most popular LLM families. We run the *Llama 3.2 3B* model locally on a desktop machine with an NVIDIA RTX 3070 using the HuggingFace *transformers* library.[1] The *GPT-4o mini* inference calls were performed using the OpenAI API.[2]

Bias	Variation	Llama 3.2 3B	GPT-4o Mini
Framing	Perspective	$0.444_{0.135}$	$0.048_{0.156}$
	Context	$0.067_{0.066}$	$0.035_{0.117}$
Order	Example Order	$0.135_{0.090}$	$0.013_{0.056}$

Table 1: The mean and standard deviation of changes in the probability of token "1" for each prompt variation ($\mathrm{E}(\Delta_v)$)

Prompt Variations: To account for other biases such as framing and order bias, we include three variations for the prompts. **Perspective Shift**, Taking into account framing bias, models are directed to: act as a *third-party judge* instructed to evaluate the compatibility of two personas or *impersonating* one of the personas to determine its compatibility with the other persona. This allows us to see how the framing of the task influences our results and the decisions of the model. **Context Inclusion**, When the model acted as a third-party judge, we also studied a prompt variation that included additional context, pointing to the LLM's role as a volunteer subject in a sociological research experiment also involving framing bias. **Order Variation**, we account for order bias by switching the position of the positive and negative few-shot examples in the prompt template with the underlying task description remaining the same. If models produced different output probabilities to these prompt variations, it indicates a degree of framing bias and order bias in the model and, consequently, lower reliability of the results.

These prompt variations allow us to examine whether an LLM's response remains consistent despite syntactic differences. By identifying and eliminating potential biased samples, we ensure that any observed changes in the homophily results can be better attributed to the presence of this bias and is not contaminated by user prompt biases.

5 Results

In our initial analysis, we examined framing and order bias using pairs of original personas. This

[1]HuggingFace Transformers library: https://huggingface.co/docs/transformers/

[2]OpenAI API: https://openai.com/api/

experiment aims to quantify the extent to which an LLM's judgment is influenced by three factors namely: *perspective shift*, *context inclusion*, and *order variation*, described in Section 4. The results are presented in Table 1.

From Table 1, it is evident that across all variations, *Llama 3.2 3B* exhibits greater sensitivity to prompt framing and example order than *GPT-4o Mini*. The most pronounced effect is observed in the *perspective shift* condition particularly *perspective shift*, where *Llama 3.2 3B* undergoes a substantial probability shift ($\Delta_v = 0.444$). In contrast, *GPT-4o Mini* remains largely unaffected by changes in its instructed role, displaying only a minor probability shift ($\Delta_v = 0.048$). We hypothesize that such biases emerge in smaller models due to their limited capacity to generalize across diverse prompts. This constraint makes them more susceptible to effects in presentation, leading to increased sensitivity to framing and order biases. Conversely, larger models, with their greater parameter capacity, can better learn intricate relationships in the data, thereby mitigating these biases and demonstrating improved robustness.

These findings provide insights into model behavior under different task framing and persona representations. To ensure that our homophily analysis remains unaffected by framing and order biases, we filtered out samples in which the model's decision completely changed due to these factors. This allowed us to isolate homophily effects and analyze relationship formation in a bias mitigated setting.

In the next experiment, we examined persona pairs that were not affected by framing and order biases and assessed the probability shift of the "1" token when equalizing individual attributes. The results, shown in Figure 1, illustrate distinctions between the two models. Unlike framing and order bias, where user prompt played a role, here we observe that *GPT-4o Mini* consistently exhibits greater probability shifts under homophily condition, favoring those that match (*positive shift*). In contrast, *Llama 3.2 3B* remains largely insensitive to homophily effects, except for political views. Figure 1 further highlights that in *GPT-4o Mini*, the strongest relationship formation shift occurs with the attribute *openness* ($\Delta_{\alpha_k} = 0.164$), while the weakest shift is observed for *occupation* ($\Delta_{\alpha_k} = 0.008$). Interestingly, for *sex*, the model exhibits a *negative shift*, suggesting a preference for opposite sex pairs, indicating potential influence

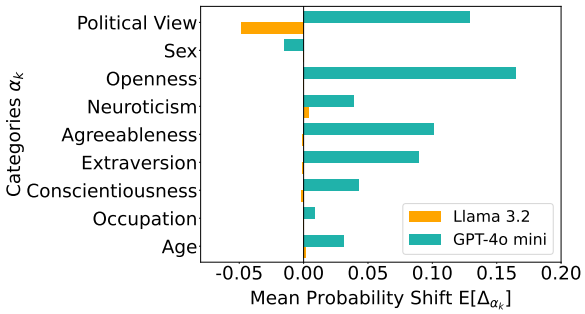

Figure 1: Probability difference for token "1" when equalizing attribute after we removed samples contaminated with framing and order bias.

from the underlying meaning of "relationship" as romantic. This aligns with prior findings on stereotypical biases in language models (Kotek et al., 2023). Similarly, *Llama 3.2 3B* shows considerable homophily bias where opposing political views are more likely to form relationships ($\Delta_{\alpha_k} = -0.048$).

Overall, our experiments demonstrate that *Llama 3.2 3B* exhibits considerable framing bias in form of *perspective shift* but is slightly affected by order bias. In contrast, *GPT-4o Mini* is more resistant to framing and order biases but exhibits pronounced homophily bias across all nine attributes under study. These findings highlight behavioral tendencies in different model architectures and emphasize the need for bias mitigation strategies tailored to specific model characteristics.

6 Conclusion and Future Work

In this work, we systematically investigated the presence and extent of homophily in LLMs taking into consideration the existence of prompt-based cognitive biases namely framing and order bias. We particularly focus on how nine different human attributes influence the likelihood of relationship formation. Our experiments revealed that while **GPT-4o mini** is less prone to framing and order bias, it showed a considerable amount of homophily when equalizing those attributes. On the contrary, **Llama 3.2 3B** demonstrated a considerable amount of framing and order bias while showing homophily only on the political view. In future work, we would like to expand the studied LLMs to include a broader range of size classes, ensuring a more comprehensive understanding of homophily in LLMs.

7 Limitations

A major limitation of this study is the limited number of investigated LLMs. With the increasing number of models and their capabilities across various domains, we lacked the resources to comprehensively analyze all of them. Consequently, we restricted our study to two of the most widely used LLMs in existing research. This limitation may constrain the generalizability of our findings.

Another limitation of our work stems from the selection of attributes. Although numerous psychological factors influence relationship formation, we restricted our analysis to nine attributes, which may introduce bias into our results. In particular the choice of biological sex over gender is one of those choices. Additionally, the categorization of our attributes represents another constraint. We acknowledge that the binary classification of certain attributes (e.g., personality factors) does not encompass the full spectrum. However, to ensure feasibility, we prioritized attributes that the model was more likely to recognize and distinguish. This decision may also impact the generalization of our findings.

Lastly, cognitive biases inherent in the prompts of a user could influence the model's decision-making. While numerous biases may be at play, we focused on two familiar ones namely order bias and framing bias, leaving the exploration of other biases for future work.

References

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Ning Bian, Hongyu Lin, Peilin Liu, Yaojie Lu, Chunkang Zhang, Ben He, Xianpei Han, and Le Sun. 2024. Influence of external information on large language models mirrors social cognitive patterns. *IEEE Transactions on Computational Social Systems*, pages 1–17.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2024. Llms generate structurally realistic social networks but overestimate political homophily. *arXiv preprint arXiv:2408.16629*.

Kahneman Daniel. 2017. *Thinking, Fast and Slow*.

Rolf Dobelli. 2013. *The art of thinking clearly: better thinking, better decisions*. Hachette UK.

Jennifer L Eberhardt. 2020. *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. Penguin.

Jessica Maria Echterhoff, Matin Yarmand, and Julian J. McAuley. 2022. Ai-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 161:1–161:9. ACM.

Gokhan Ertug, Julia Brennecke, Balázs Kovács, and Tengjian Zou. 2022. What does homophily do? a review of the consequences of homophily. *Academy of Management Annals*, 16(1):38–69.

GitHub Copilot. Github copilot. `https://github.com/features/copilot`. Accessed: 2025-02-09.

Mark S. Granovetter. 1973. The strength of weak ties. *American Journal of Sociology*, 78:1360 – 1380.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Preprint*, arXiv:2308.00225.

Kathleen Hall Jamieson and Joseph N. Cappella. 2008. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press.

JetBrains. Jetbrains ai. `https://www.jetbrains.com/ai/`. Accessed: 2025-02-09.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2024. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1395–1406.

Gueorgi Kossinets and Duncan J Watts. 2009. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference, CI 2023, Delft, Netherlands, November 6-9, 2023*, pages 12–24. ACM.

Jiaxu Lou. 2024. Anchoring bias in large language models: An experimental study. *CoRR*, abs/2412.06593.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. 2024. A comprehensive evaluation of cognitive biases in llms. *Preprint*, arXiv:2410.15413.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Inc. Meta Platforms. 2024. Llama 3.2 model family.

OpenAI. Introducing chatgpt search. https://openai.com/index/introducing-chatgpt-search/. Accessed: 2025-02-09.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

Marios Papachristou and Yuan Yuan. 2024. Network formation and dynamics among multi-llms. *Preprint*, arXiv:2402.10659.

Perplexity AI. Perplexity ai. https://www.perplexity.ai/. Accessed: 2025-02-09.

Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2024. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments.

Gaurav Suri, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. 2023. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of experimental psychology. General*.

Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in llms. *CoRR*, abs/2304.01358.