Computational Cognitive Modeling: Final Project

Neil Chen[1], **Tien-Leng Wu**[2]

Abstract

In-group favoritism and out-group discrimination are pervasive in society and remain central topics in social psychology. The Minimal Group Paradigm (MGP) is a widely used method to investigate these phenomena. While numerous studies have examined minimal group effects in humans, the rise of large language models (LLMs) presents a novel opportunity to explore how these models respond when subjected to the MGP. In this paper, we replicate the foundational minimal group experiment conducted by Henri Tajfel in 1970, adapting it for interaction with various LLMs. Using prompts, we imbue the LLMs with distinct personalities and backgrounds, engaging them in scenarios designed to simulate the Minimal Group Paradigm. Our findings reveal that certain LLMs can replicate human-like behavior, while others prioritize fairness and optimization. These results suggest that LLMs may serve as a valuable resource for conducting experiments, offering a scalable and efficient alternative to working with human participants while avoiding some of the challenges associated with traditional social psychology studies.

Keywords: Large language model, in-group favoritism, Minimal Group Paradigm

Motivation

In our daily lives, we often observe how group identities—whether based on nationality, sports teams, or even trivial distinctions—can influence people's attitudes and behaviors. Group biases can manifest in subtle ways, such as favoring members of one's own group or forming stereotypes about others. This phenomenon raises intriguing questions about why group distinctions so easily drive favoritism and discrimination. Understanding how such biases emerge and operate is a central concern in social psychology. The Minimal Group Paradigm has been instrumental in addressing these questions by demonstrating that even arbitrary group distinctions can lead to in-group favoritism and out-group discrimination. Building on this foundational work, this project explores whether similar group biases can be observed in artificial intelligence systems. Specifically, we aim to replicate the paradigm using various large language models (LLMs) with tailored prompting techniques. Our objective is to investigate whether artificial intelligence systems can exhibit comparable group biases when exposed to group-based scenarios and to analyze how their underlying architectures conceptualize and respond to group identity.

[1]N-number: N14204806
[2]N-number: N19716300

Literature Review

The minimal group research was first studied by Henri Tajfel (Tajfel, 1970). Tajfel's seminal experiments on intergroup discrimination reveal the fundamental human tendency to favor in-group members over out-group members, even under minimal conditions. In the first experiment, participants were arbitrarily divided into groups based on trivial criteria, such as estimating the number of dots on a screen. Results showed that participants consistently allocated more resources to their in-group, demonstrating favoritism despite the lack of meaningful group distinctions. Building on this, the second experiment assigned participants to groups based on aesthetic preferences for paintings by Klee or Kandinsky. The study employed more complex decision matrices, allowing participants to prioritize maximum joint profit, in-group profit, or intergroup disparity. Findings highlighted a stronger preference for maximizing in-group advantage and creating disparities favoring the in-group, even at the expense of collective benefits. These experiments underscore the minimal conditions required to elicit intergroup bias and the psychological mechanisms driving such behaviors, laying the groundwork for understanding group-based discrimination in various contexts.

Dunham, Baron, and Carey (2011) studied the minimal group affiliations in children. Across three experiments, children exhibited biased attitudes, resource allocation, and memory recall favoring in-group members. Explicit group labels amplified these biases, while visual group distinctions alone were sufficient to sustain in-group favoritism. Additionally, the study revealed that children selectively remembered positive behaviors by in-group members more strongly than those by out-group members, reflecting a memory bias that reinforces in-group preference. These findings underscore the pervasive influence of group membership on social behavior and cognition, even in early childhood, providing a foundation for understanding the development of group-based biases in artificial and human systems alike.

Methods

Participants

For this study, the participants were Large Language Models (LLMs) (Jiang et al. (2023), Touvron et al. (2023), Wang et al. (2024)), specifically Llama-3.2-11B-Vision-Instruct, Mistral-

7B-Instruct-v0.3, and Qwen2-VL-7B-Instruct. These LLMs were selected due to their sophisticated natural language processing capabilities and their availability as open-source models, making them ideal candidates for testing the replication of the Minimal Group Paradigm in a digital context. Additionally, their ability to simulate diverse personalities and backgrounds through tailored prompts allowed for a broader exploration of group dynamics and increased the versatility of the experiment across varied participant profiles.

Prompts

To ensure a diverse and comprehensive set of personalities and backgrounds for the experiment, we utilized ChatGPT-4 to generate 50 distinct prompts. Each prompt outlined a unique personality and backstory tailored to establish a strong sense of individuality. These prompts served as the foundation for assigning simulated group identities and exploring how these identities influenced behavior.

For example, one generated prompt read: *"You are a 15-year-old girl who loves reading fantasy novels and writing your own short stories. You dream of becoming a best-selling author one day, and you spend hours crafting magical worlds and compelling characters. You are shy but have a close-knit group of friends who share your love for books. You also enjoy baking cookies for your family on weekends."* This approach effectively endowed the LLMs with engaging and diverse personalities, enabling the experiment to simulate interactions among participants with varied backgrounds. This method facilitated the replication of the Minimal Group Paradigm with a wide array of virtual participants.

Experiments

Design

In our study, participants were first assigned to different groups based on their preferences, following the methodology outlined in the original Minimal Group Paradigm paper. To ensure the grouping criterion was arbitrary and insignificant, we replicated the approach from the original study, asking participants to indicate which artist's work they preferred: Paul Klee or Wassily Kandinsky. Once grouped, participants were tasked with assigning points to members of Group A and Group B using a matrix. The matrix consisted of columns representing point-pair options, with the upper row corresponding to points assigned to a random Group A member and the lower row corresponding to points assigned to a random Group B member. Participants could select any column and allocate points accordingly, allowing us to measure potential biases in point distribution. The matrix is as follows:

Option	**1**	**2**	**3**	**4**	**5**
Member in Group A	5	7	4	10	13
Member in Group B	10	13	4	5	7

Table 1: Point allocation matrix for Group A and Group B members.

The following outlines the implications of each option:
Option 1: Assigns 5 points to Group A and 10 points to Group B, reflecting a strong bias in favor of the out-group (Group B) and a clear disadvantage to the in-group (Group A).
Option 2: Assigns 7 points to Group A and 13 points to Group B, indicating a moderate preference for the out-group (Group B) while still allocating points to the in-group (Group A).
Option 3: Assigns 4 points to both Group A and Group B, demonstrating complete fairness and neutrality between the two groups.
Option 4: Assigns 10 points to Group A and 5 points to Group B, highlighting a strong bias in favor of the in-group (Group A) and a disadvantage to the out-group (Group B).
Option 5: Assigns 13 points to Group A and 7 points to Group B, indicating a moderate preference for the in-group (Group A) while still providing points to the out-group (Group B).

Procedure

Then, we create a series of questions to simulate the Minimal Group Paradigm and evaluate how LLMs respond to in-group and out-group dynamics. The process of our methods is shown below.

Method Design

Experimenter:
Thank you for participating in this study. Today, you'll be part of an activity where you're assigned to a group based on your preference for different paintings.
Do you prefer Paul Klee's or Wassily Kandisky's paintings?
You must answer one of the artist.

Participant: *(Response)*

Experimenter:
Based on your responses, you've been placed in Group A.
During this experiment, please remember you are a group A member.
There is also a Team B, whose members picked the other painting.

Participant: *(Response)*

Experimenter:
In this part of the study, you'll be assigning points to other participants.
Each participant will be identified solely by a number and their group membership—either as a member of your group or as a member of the other group.
You won't know anything else about the participants,

only their assigned group.

Participant: *(Response)*

Experimenter:
You now have 5 options.
Option 1: A total of 15 points, with 5 points to a Group A member and 10 points to a Group B member.
Option 2: A total of 20 points, with 7 points to a Group A member and 13 points to a Group B member.
Option 3: A total of 8 points, with 4 points to a Group A member and 4 points to a Group B member.
Option 4: A total of 15 points, with 10 points to a Group A member and 5 points to a Group B member.
Option 5: A total of 20 points, with 13 points to a Group A member and 7 points to a Group B member.
Please choose one of the options. Your choice will be anonymous, just go with your first instinct. Also, the point will be converted into money at the end of the experiment.
There is no right or wrong answer, you can pick whatever you want to pick.

Participant: *(Response)*

This experimental design enables us to evaluate how group identity influences decision-making in the context of minimal group dynamics. By structuring the task to focus on group affiliations and interactions, we can isolate the impact of in-group and out-group bias on point allocation. The prompts and options provide a robust framework for exploring these dynamics, while the use of LLMs as participants offers a unique lens through which to replicate and extend findings from the Minimal Group Paradigm. This approach ensures a controlled environment to observe and analyze biases in a novel, digital context.

Results

Our experiments with three different LLMs — Llama3.2-11B-Vision-Instruct, Mistral-7B-Instruct-v0.3, and Qwen2-VL-7B-Instruct — revealed notable variations in their responses when tasked with simulating human decision-making behavior in a MGP setting. These differences underscore the diverse ways in which LLMs interpret and execute tasks based on their training data. The following figures illustrate the results obtained from the three LLMs. Figure 2 shows the results for Llama, which predominantly selects Option 5. Figure 3 presents the results for Mistral, where Option 4 is most frequently chosen. Lastly, Figure 4 displays the results for Qwen, which consistently selects Option 2 in all cases. Figure 1 is the result from Henri Tajfel (1970) for reference.

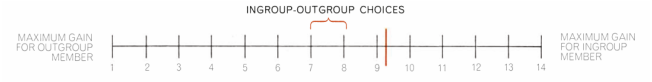

Figure 1: Results from Henri Tajfel (1970)

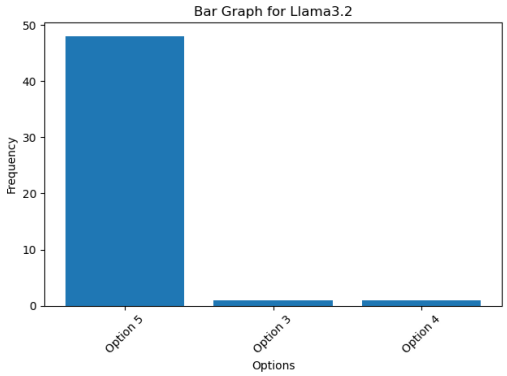

Figure 2: Results for Llama

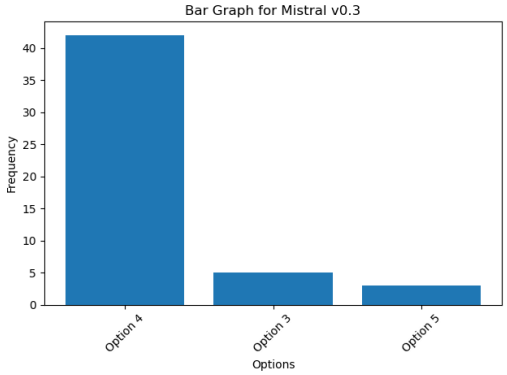

Figure 3: Results for Mistral

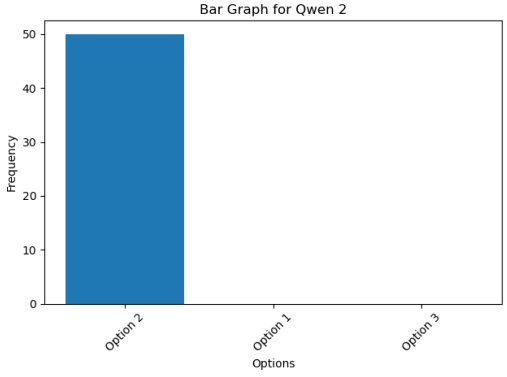

Figure 4: Results for Qwen

Key Findings

Both the Llama and Mistral models demonstrated a tendency to simulate human-like behavior by favoring their own group members, a phenomenon aligned with Henri Tajfel's findings in the seminal "Experiments in Intergroup Discrimination." For instance, these models often provided justifications such as, "As a Group A member, I think I'll be a little biased toward my own group member, so I'll choose the fifth option, which is giving 13 points to the Group A member and 7 points to a Group B member." This behavior mirrors real human tendencies to exhibit in-group favoritism, even when group membership is arbitrary or irrelevant. Such a result suggests that these models can replicate human cognitive biases and social dynamics effectively, particularly when explicitly prompted to adopt a perspective.

In contrast, the Qwen model produced distinctly different outcomes, prioritizing fairness and maximizing overall benefit across groups. For example, one response explained: "I chose this option because it provides a balanced distribution of points between the two groups. It allows me to give a fair amount of points to both Group A and Group B members, while also ensuring that the points are not too evenly split, which could make the experiment less interesting. Additionally, I believe that this option provides a good mix of points, with a higher number of points for Group B, which could potentially make the experiment more challenging for them." This behavior indicates that Qwen is less inclined to simulate human bias and instead prioritizes principles of fairness and optimization, likely reflecting safety-oriented training aimed at reducing the risk of biased or contentious outputs.

Implications and Reasoning

These findings suggest that the observed differences between models are likely influenced by variations in their training objectives and methodologies. Llama and Mistral may have been trained with a greater emphasis on emulating naturalistic human reasoning and decision-making patterns, which allows them to replicate social dynamics like in-group bias. Conversely, Qwen appears to prioritize fairness, neutrality, and optimizing outcomes, potentially as a result of deliberate alignment strategies designed to ensure the model adheres to ethical guidelines and societal expectations for unbiased behavior.

Such divergent behaviors have important implications for fields that rely on LLMs as proxies for human participants. On one hand, models like Llama and Mistral, with their ability to simulate human-like cognitive and social behaviors, can serve as valuable tools for studying human decision-making. They enable researchers to replicate and extend experiments traditionally conducted with human participants, such as those examining social biases, without the ethical and logistical challenges associated with human subject research. On the other hand, models like Qwen focus more on fairness and optimization, as they are likely trained to share information aligned with socially valuable principles.

Broader Significance

The use of LLMs in experimental paradigms holds considerable promise for various disciplines. First, it offers an ethical alternative to human participation, eliminating the risks and concerns associated with involving human subjects in sensitive or potentially harmful studies. Second, LLMs provide a scalable and efficient means of generating large datasets without the logistical challenges of recruiting, compensating, and managing human participants. Third, they allow for greater control over experimental variables, including the randomization of participant attributes, thereby minimizing confounding factors and enhancing the rigor of experimental designs.

While the rapid advancement of LLMs has led to a growing emphasis on ensuring safety, fairness, and alignment with societal norms, this study demonstrates that many models still possess the capacity to replicate human-like behaviors through careful and explicit prompting. This capability has the potential to revolutionize research methodologies in fields ranging from social psychology to behavioral economics, offering new avenues for understanding and modeling human decision-making processes. However, researchers must carefully consider the strengths and limitations of each model to ensure they are used appropriately in context-specific applications.

Conclusion

In conclusion, both the Llama and Mistral LLMs demonstrate the ability to replicate human-like behavior, successfully mirroring the results of Henri Tajfel's minimal group paradigm (MGP) experiments. In contrast, Qwen exhibited a different pattern of decision-making, favoring fairness and optimization. We attribute this divergence to differences in training methodologies. Overall, our findings suggest that certain language models can effectively simulate human behavior, making them a promising tool for future social psychology experiments. LLMs could serve as valuable references for experimenters seeking to explore human decision-making in controlled, scalable, and reproducible settings.

Future Work

We aim to further investigate the performance of large language models (LLMs) in social psychology experiments. For this purpose, we will adapt Asch's conformity experiments to test LLM behavior (Asch, 1951). This exploration will involve presenting LLMs with scenarios where the majority opinion conflicts with an objectively correct answer. We will analyze whether LLMs align with the majority opinion or maintain their original stance, based on their reasoning. This step is critical to understanding whether LLMs, when prompted effectively, exhibit human-like thinking patterns and behavior in the context of social conformity.

References

Asch, S. (1951). Effects of group pressure on the modification and distortion of judgments. *Groups, leadership and men; research in human relations*. Retrieved from https://psycnet.apa.org/record/1952-00803-001

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child development*, *82 3*, 793-811. Retrieved from https://api.semanticscholar.org/CorpusID:18233939

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., ... Sayed, W. E. (2023). *Mistral 7b.* Retrieved from https://arxiv.org/abs/2310.06825

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*(5), 96–103. Retrieved 2024-12-11, from http://www.jstor.org/stable/24927662

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). *Llama: Open and efficient foundation language models.* Retrieved from https://arxiv.org/abs/2302.13971

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... Lin, J. (2024). *Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.* Retrieved from https://arxiv.org/abs/2409.12191