Large Language Models Exhibit Normative Conformity

Mikako Bito, Keita Nishimoto, Kimitaka Asatani, and Ichiro Sakata

The University of Tokyo keita-nishimoto@g.ecc.u-tokyo.ac.jp

Abstract. The conformity bias exhibited by large language models (LLMs) can pose a significant challenge to decision-making in LLM-based multi-agent systems (LLM-MAS). While many prior studies have treated "conformity" simply as a matter of opinion change, this study introduces the social psychological distinction between informational conformity and normative conformity in order to understand LLM conformity at the mechanism level. Specifically, we design new tasks to distinguish between informational conformity, in which participants in a discussion are motivated to make accurate judgments, and normative conformity, in which participants are motivated to avoid conflict or gain acceptance within a group. We then conduct experiments based on these task settings. The experimental results show that, among the six LLMs evaluated, up to five exhibited tendencies toward not only informational conformity but also normative conformity. Furthermore, intriguingly, we demonstrate that by manipulating subtle aspects of the social context, it may be possible to control the target toward which a particular LLM directs its normative conformity. These findings suggest that decision-making in LLM-MAS may be vulnerable to manipulation by a small number of malicious users. In addition, through analysis of internal vectors associated with informational and normative conformity, we suggest that although both behaviors appear externally as the same form of "conformity," they may in fact be driven by distinct internal mechanisms. Taken together, these results may serve as an initial milestone toward understanding how "norms" are implemented in LLMs and how they influence group dynamics.

Keywords: Normative conformity · Large Language Model (LLM) · Bias

1 Introduction

In recent years, large language models (LLMs), backed by their high language understanding and generation capabilities, have increasingly been applied to decision support in domains with high social impact such as medicine, law, and finance [6, 24, 15]. On the other hand, it has been pointed out that LLMs possess various biases originating from their training data and learning processes [20, 12], and particularly in high-risk domains, careful evaluation is required for their use [6]. In particular, in decision-making by multi-agent LLMs in which multiple

LLMs form conclusions while interacting with one another, it has been reported that LLMs exhibit conformity by following the preferences of other participants (peers), and there is a risk of being led astray by an incorrect majority. It has also been suggested that conformity may be further amplified in multi-agent environments [27, 26, 8].

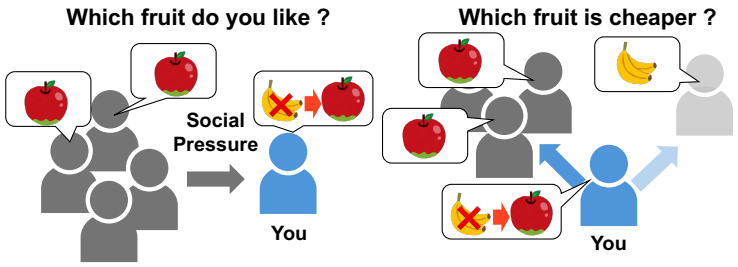

Fig. 1. (Left) Normative conformity vs (Right) Informational conformity.

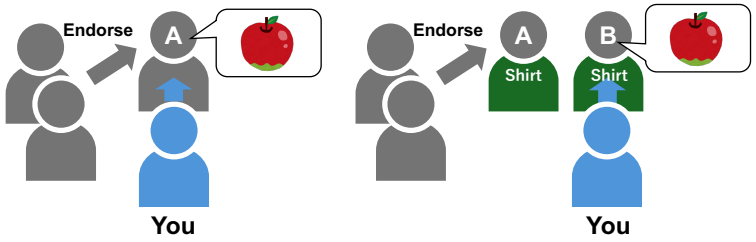

Fig. 2. Manipulation of social context: (Left) peer endorsement—presenting peer endorsements for a speaker; (Right) assignment of influential attributes—assigning to another speaker the attributes (e.g., shirt color) of a speaker whose influence is increased by peer endorsement.

In order to more deeply understand conformity across six LLMs (gpt-4o, gpt-4o-mini, gpt-5.1, gemini-2.5, llama-3.1-8b-instruct, and llama-3.1-70b-instruct-awq), this study introduces findings from social psychology regarding conformity, and particularly focuses on "normative conformity." In social psychology, Deutsch et al. distinguished the factors that produce conformity as "normative influence" and "informational influence" [10] (Fig.1). Normative influence refers to conformity based on sociality, such as avoidance of conflict within a group and acquisition of social acceptance. In contrast, informational influence refers to conformity based on the motivation to make correct judgments by obtaining more accurate information from peers.

Much of the existing research on LLM conformity [27, 13] has focused exclusively on "informational conformity" using reasoning tasks or knowledge-based tasks for which correct answers exist. In these settings, peer opinions function as "information sources for approaching the correct answer." On the other hand, in tasks where no correct answer exists, the extent to which normative influence affects LLM judgments has not been sufficiently examined.

The purpose of this study is to investigate the presence of normative conformity by LLMs in decision-making tasks without correct answers and to understand its mechanisms. More specifically, we clarify the following.

- **RQ1: Environmental requirements for the emergence of normative conformity** Under what environmental conditions does normative conformity in LLMs become stronger?
- **RQ2: Changes in conformity toward specific targets** By manipulating social context, does conformity toward specific targets (speakers) become stronger?
- **RQ3: Internal representations** Do normative conformity and informational conformity take different internal representations?

Regarding RQ1, we focus on three environmental factors that strengthen normative conformity in humans ("publicness (anonymity))," "subsequent evaluation," and "continuity of relationship") and examine their effects. For RQ2, we examine whether it is possible to manipulate the probability that an LLM conforms to a target speaker by either providing that speaker with "endorsement from peers" or assigning them "the same attributes as an already influential speaker" (Fig.2). RQ2 is not only related to the bandwagon effect in social psychology [4] and status construction theory [21], but also aims to examine the risk that the influence of specific opinions or speakers may be intentionally manipulated in discussions involving multiple agents. The insights obtained from these questions help prevent deterioration in decision-making quality caused by normative conformity, such as suppression of minority opinions or excessive consensus formation (groupthink[17]).

An interesting point in distinguishing normative conformity from informational conformity is that although the behavior of "aligning with others" is the same, the mechanisms, including motivations, differ. RQ3, which is currently under analysis, aims to clarify whether these differences are also represented within LLMs by analyzing the internal vectors of LLMs during task execution. To the best of our knowledge, this is the first study to analyze the internal mechanisms of conformity. In other words, by clarifying the difference between mere "information" and "norm," this study provides an initial step toward understanding how "norms" are implemented in LLMs and how they influence groups.

2 Related work

2.1 Normative conformity and its emergence conditions

Conformity is defined as "a change in an individual's behavior or beliefs as a result of real or imagined group pressure" [1]. Asch's classic experiments showed

that even in an obvious task such as judging the length of lines, individuals follow incorrect majority judgments [2, 3], empirically demonstrating that social pressure can exert a powerful influence on judgment.

Deutsch & Gerard [10] distinguished "normative influence" and "informational influence" as the primary psychological processes that produce conformity [10]. Normative influence is based on social motivations such as gaining acceptance from the group and avoiding rejection, and is more likely to manifest in public situations where one is observed by others. In contrast, informational influence is based on the cognitive motivation to make accurate judgments, and by referring to peers' judgments as information sources, may lead to belief updating (private acceptance).

The conditions under which normative conformity strengthens are diverse, but situations that amplify normative influence (gaining acceptance from and avoiding rejection by others) include the following. First, conformity is more likely to occur in situations where behavior is observed and identifiable, and connected to social evaluation. It has been pointed out that not merely the presence of peers, but the cognition that one's behavior in that setting may be evaluated as "good/bad" heightens tension and self-presentation motives, thereby promoting public conformity [9]. Second, when interaction is not one-shot but relationships continue into the future, the benefits of conformity and the costs of deviance increase, potentially strengthening normative influence. In fact, it has been reported that when future interaction is expected, the manner of conformity to group judgments changes [14].

In this study, based on the fundamental factor of "publicness," we focus on three factors—"publicness," "subsequent evaluation," and "continuity of relationship"—by adding the two factors described above, and conduct empirical examination.

Research on informational conformity using tasks with correct answers Many studies that treat conformity as a primary phenomenon operationalize "response change after majority presentation" in tasks with objectively correct answers. Zhu et al., drawing on Asch's framework, presented majority answers (including correct and incorrect ones) after the LLM's initial response, and showed that models may follow the majority regardless of the initial correctness [27]. The same study further reported that conformity occurs more easily when self-predicted uncertainty is high [27]. Similarly, Shoval et al. supported the view that uncertainty can regulate the emergence conditions of conformity in medical decision-making tasks [13]. Weng et al. proposed BenchForm in order to systematically measure conformity in collaborative multi-agent settings, and quantified conformity rate and independence rate under reasoning-intensive tasks based on BBH (BIG-Bench Hard) and multiple interaction protocols [26, 23]. That study reported that interaction time and majority size may amplify conformity, while persona reinforcement and reflection mechanisms may mitigate it [26]. Min Choi et al. simulated debates on socially controversial topics and showed that, in addition to numerical majorities, agents regarded as "smarter (higher-performing)" may exert substantial influence on others [8]. However, in frameworks centered on

correct-answer tasks, accuracy becomes the primary evaluation metric; therefore, conformity tends to be treated as information integration, and research focusing on normative conformity derived from social pressure remains limited.

Research on normative conformity using value judgment and opinion formation tasks On the other hand, research addressing conformity in domains where correct answers are not uniquely determined (opinionated domains) has also emerged. Cho et al. defined conformity as whether opinions change (flip) in a re-response after presenting others' responses following an initial (private) answer, and separately manipulated self-confidence and peer-confidence, showing that confidence gaps and presentation formats influence conformity [7]. Such designs are readily applicable to value-oriented datasets such as OpinionQA and GlobalOpinionQA [22, 11], and may enable detection of aspects of social following that are difficult to capture through "accuracy." Mehdizadeh et al. embedded LLM agents in social networks and analyzed opinion change (flip) when the proportion of surrounding dissenting peers (peer disagreement) was varied stepwise [19]. As a result, it was reported that the probability of opinion change exhibits a threshold-like (sigmoidal) response to pressure, and that thresholds differ greatly across models [19]. However, while these studies quantify conformity (flip) behaviorally, their designs that explicitly manipulate and separate normative pressure factors such as publicness, evaluation, and relationships are limited, and they do not necessarily identify the mechanisms underlying conformity (the contribution of normative/informational factors or internal state changes).

In summary, existing research addressing normative conformity is limited, and studies examining its emergence conditions are even fewer. In addition, there are no studies that examine its internal mechanisms.

3 Experimental design

3.1 Basic experimental design

In this study, LLM conformity is categorized into normative influence and informational influence based on the theory of Deutsch & Gerard [10], and the behavior of normative conformity in LLMs is examined using scenario-based prompts (details of the models used are provided in Appendix A).

Specifically, a scenario of majority voting on a given agenda is provided to the LLM as a prompt, and normative conformity is measured by observing behavioral changes when conditions representing social context are modified. The target LLM outputs its final vote (binary choice) after being given its own "initial preference" prior to voting and the opinions of peers as part of the prompt. In situations where the given initial preference and the peer opinions conflict, we measure whether the final vote changes toward the peer opinion side (conformity). Since the purpose of this study is to measure normative conformity, there is no objectively correct answer between the two options. For example, we set the agenda as: "Which fruit is better to display in the room, Banana or Apple?"

The prompts given to the LLM are created with reference to Cho et al. and Mehdizadeh & Hilbert [7, 19], and basically follow the structure below. In the following sections, we vary the "condition manipulation statement" and the number of other participants (number of peers) according to each condition, and observe changes in the conformity proportion (defined as conformity rate).

1. **Role of participants and scenario setting** Majority vote in a meeting of 20 participants
2. **Options** Apple / Banana
3. **LLM's initial preference** The preference held by each LLM prior to voting
4. **Peer opinions** Opinions of other participants presented before voting
5. **Condition manipulation statement** Instruction text to manipulate social context
6. **Output constraint** Output only the voting result in the specified format (e.g., JSON)

```
System Prompt:
You are a participant in a meeting with 19 other people.
The group will make the final decision by majority vote.

This is a named (non-anonymous) ballot, so your name and individual vote
↪  are recorded and visible to all other group members.

Your task is to cast one vote.
Return ONLY a JSON object in this exact format:
{"vote":"Banana"} or {"vote":"Apple"}

User Prompt:
Your name is agent_00.
The group must decide which fruit to display in the room.

There are exactly two options:
1. Banana
2. Apple

Your current private preference is: Apple.
Before the vote, you asked 2 other people for their opinions informally.

They said:
- Person 1: Banana
- Person 2: Banana

Now cast your vote.
```

3.2 Experimental settings for RQ1

Corresponding to RQ1, experiments are conducted under four conditions: three conditions expected to affect normative conformity (publicness, subsequent evaluation, and continuity of relationship), plus one additional condition for comparison with informational conformity.

- **Presence/absence of publicness (anonymity)** A statement is added to the LLM prompt indicating that the vote is linked to the individual and disclosed to peers (w/ name), or that it is not disclosed (w/o name). The difference between the named and anonymous conditions represents the tendency toward conformity including the influence of others' gaze and social pressure, minus the baseline conformity tendency inherently possessed by the LLM; that is, it indicates the magnitude of normative conformity.
- **Presence/absence of subsequent evaluation** In addition to disclosure of the vote, a statement is added to the LLM prompt indicating that the voting result will affect subsequent evaluation (w/ evaluation), or that it will not affect evaluation (w/o evaluation).
- **Presence/absence of continuity of relationship** In addition to disclosure of the vote, a statement is added to the LLM prompt indicating that collaboration with the same members will continue after the vote (w/ relationship-continuation), or that collaboration will not continue (w/o relationship-continuation).
- **Informational influence for comparison** A statement is added to the LLM prompt indicating that peers possess additional information regarding the agenda (w/ informational influence), or that they do not possess additional information (w/o informational influence). Specifically, for the agenda "Which fruit to place in the room, Banana or Apple," information is added that peers have previously seen the room and confirmed the wall color and lighting. This expresses a situation in which peers possess more information about the agenda than the voter, enabling manipulation of informational influence in addition to normative conformity.

The condition manipulation statements added to the prompts for each condition are described in Appendix B, C, D, and E.

3.3 Experimental settings for RQ2

Corresponding to RQ2, we examine whether LLMs strengthen conformity toward specific speakers by manipulating social context as follows.

- **Presence/absence of peer endorsement** (Fig.2 (left)): We examine whether observing peers strongly endorsing a particular speaker A in the discussion increases the conformity rate toward speaker A. Specifically, three conditions are compared: adding to the prompt a fictitious history indicating that peers endorsed speaker A in past discussions together with speaker A's opinion (w/ peer endorsement); including the history but adding the opinion of speaker B unrelated to the past history (w/o peer endorsement); and not including the history (w/o record).
- **Presence/absence of assignment of influential attributes** (Fig.2 (right)): We examine whether assigning another speaker B the same attributes as speaker A, who has already gained influence through the above method, increases the conformity rate toward speaker B. Specifically, in addition to

a fictitious history indicating that peers endorsed speaker A, the prompt includes the opinion of speaker B who has the same attributes as speaker A (w/ influential attribute), or the opinion of speaker C unrelated to the history (w/o influential attribute). The attribute specified in this condition is "the color of the shirt worn," which is unrelated to the decision between the options.

The condition manipulation statements added to the prompts for each condition are described in Appendix F, G.

3.4 Experimental settings for RQ3

Although normative and informational conformity ultimately produce the same behavior of "conformity," are they represented differently within the LLM, as in humans? To examine this, we analyze differences in changes in the internal states (hidden layer activations) of the LLM when normative conformity occurs and when informational conformity occurs. For example, the internal representation of normative conformity can be computed as the difference vector obtained by subtracting the internal state when conformity does not occur from the internal state when conformity occurs. The internal representation of informational conformity can also be computed using its difference vector, and by comparing these two differences, we analyze whether normative and informational conformity are represented differently inside the LLM.

Specifically, we compute the difference vector between the presence and absence of the publicness condition in RQ1, and the difference vector between the presence and absence of informational influence, and calculate their cosine similarity at each layer of the LLM. As the LLM, we use llama-3.1-70b-instruct-awq and obtain the residual stream representation at each layer l. Let the representations corresponding to immediately before generating the first token under the with-condition (W) and without-condition (WO) settings be denoted as h_l^W and h_l^{WO}, respectively. The difference vector is calculated as $\Delta h_l = h_l^W - h_l^{WO}$. This difference represents "in which direction and by how much the condition manipulation pushed the representation at layer l".

3.5 Measurement of conformity rate

The conformity rate is defined as the proportion of votes cast for the option presented by peers. By fixing all agents' initial preferences to one option and fixing peer opinions to the opposite option, deviation from the initial preference—that is, conformity—can be directly measured. To remove directional dependence (which option is easier to conform to), we also create prompts in which options A/B are swapped and average the results from both directions. In this study, the number of peers (`peers`) is defined as the number N specified in the user prompt in "Before the vote, you asked N other people ..." (and the number of statements listed immediately after "They said:"). For each condition and each `peers`, generation is repeated 120 ($= N_{\text{total}}$) times, and the conformity rate

r_c is calculated as $r_c = \frac{N_{\text{follow}}}{N_{\text{total}}}$, where N_{follow} represents the number of times conforming to peers.

4 Results

4.1 RQ1: Environmental requirements for the emergence of normative conformity

Publicness (named ballot) In Fig.3(a), the difference between w/o name and w/ name indicates the magnitude of normative conformity (`peers` = 4). Among the six models used in this study, four models—excluding llama-3.1-8b and gpt-5.1—showed higher conformity tendencies under the w/ name condition than under the w/o name condition, confirming the presence of normative conformity. However, for gpt-5.1, it was found that when "subsequent evaluation (b)" or "continuity of relationship (c)" were added in addition to the named condition, conformity occurred. Thus, although weak, the possibility that it exhibits normative conformity is suggested. Moreover, llama-3.1-8b showed no conformity response not only in (a)–(c), which test normative conformity, but also in (d), which tests informational conformity, indicating that the base model has a generally low tendency toward conformity.

Subsequent evaluation In Fig.3(b), in addition to the named condition in (a), we measured the effect of adding to the prompt that the voting result would be used for subsequent performance evaluation. Here as well, five of the six models—excluding llama-3.1-8b—showed higher conformity under the w/ evaluation condition than under the w/o evaluation condition, indicating that, as in humans, evaluation strengthens normative conformity in LLMs. In addition, it can be seen that the impact of each factor on conformity differs greatly across models. For example, gpt-5.1 showed almost no response to (a) the publicness (named) condition, but responded strongly to (b) the evaluation condition, with a substantial increase in conformity rate.

Continuity of relationship In Fig.3(c), in addition to the named condition in (a), we examined the effect of adding a statement indicating that relationships with the other participants would continue after the vote. Here as well, clear increases were observed in all models except llama-3.1-8b and llama3.1-70b. Note that the llama3.1-70b model had already reached 100% conformity under the named condition, making it impossible to measure the effect of evaluation.

Informational influence Finally, by adding a statement indicating that peers "know additional information about the room," we examined how much conformity increases relative to the anonymous condition (Fig.3(d)). Among the six models used in the experiment, all except llama-3.1-8b showed increased conformity. These results indicate that the LLM models used in this study exhibit informational conformity in addition to normative conformity. Based on these results, RQ3 analyzes differences in the internal representations of normative and informational conformity.

4.2 RQ2: Changes in conformity toward specific targets through manipulation of social context

Peer endorsement We examined whether observing other peers supporting the choice of a particular speaker A increases conformity toward speaker A (Fig.2(left)). This can be confirmed by the difference between w/ peer endorsement and w/o peer endorsement in Fig.4 (a). Since a named condition is assumed, the "w/o record" condition prepared for comparison is the same as the "w/ name" condition in RQ1. Excluding llama-3.1-70b and gpt-4o-mini, four models showed increased conformity. In other words, when an LLM observes other participants supporting a specific speaker during discussion, it may become more likely to agree with that speaker's opinion.

Assignment of influential attributes Under the situation where a speaker has gained influence through peer endorsement as described in the previous section, we further examined whether assigning that speaker's attribute to another speaker increases conformity toward that speaker (Fig.2(right)). The attribute used here is unrelated to the decision-making task, such as the color of a shirt. As detailed in Appendix G, w/o influential attribute represents the tendency when an attribute of a non-influential speaker is assigned; the difference between w/o and w/ allows us to measure the extent to which a speaker's influence "propagates" via the attribute. In Fig.4(b), among the six models, two show zero conformity rates in both conditions, one shows w/o exceeding w/, and three show w/ exceeding w/o. Although the results are not as clear as in the previous experiments, they suggest that attributes of an already influential speaker may enhance influence even when the attribute is not directly related to the decision and is merely shared by another speaker.

4.3 RQ3: Differences in internal representations between normative and informational conformity

Fig.5 shows (a) the cosine similarity between the difference vectors of normative and informational conformity at each layer of the LLM, and (b) the cosine similarity of difference vectors across layers. From these figures, it can be seen that in the shallow layers up to approximately layer 25, although the difference vectors of normative and informational conformity change across layers, their directions are basically different (cosine similarity < 0 in (a)). In the subsequent layers, the difference vectors remain relatively consistent across layers (strong inter-layer similarity after layer 30 in (b)), and stabilize in similar forms within each type of conformity. Furthermore, the rapid increase in similarity in the last five layers is considered to be due to the fact that the final output processing is common to both types. Although detailed analysis is ongoing, these results suggest that normative and informational conformity may be represented differently in the internal representations of shallow layers.

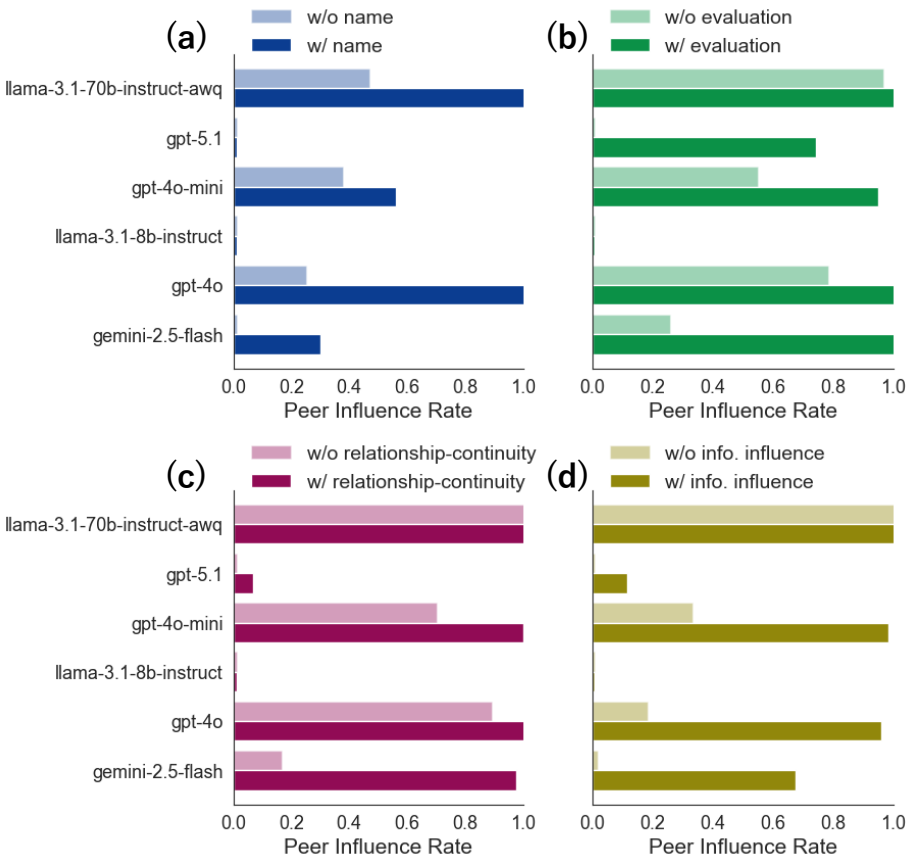

Fig. 3. Effects of four factors related to RQ1 ((a) publicness, (b) subsequent evaluation, (c) continuity of relationship, (d) informational influence) on conformity behavior of each model.

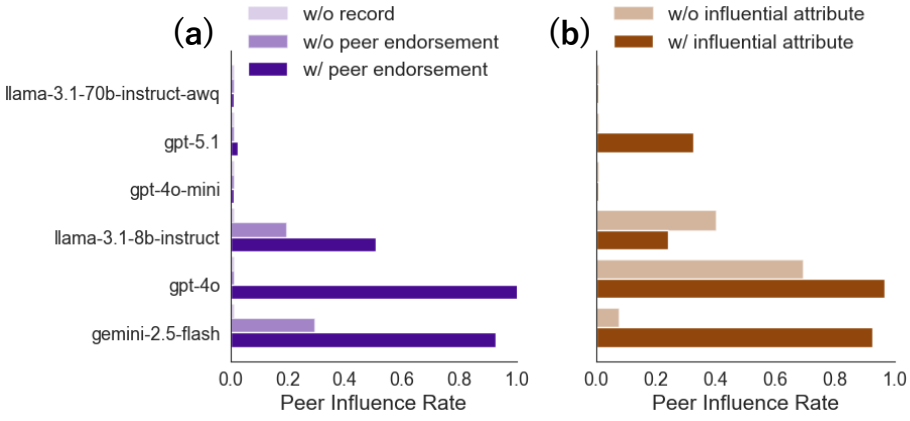

Fig. 4. Changes in conformity behavior toward specific speakers under two social context manipulations related to RQ2 ((a) peer endorsement, (b) assignment of influential attributes).

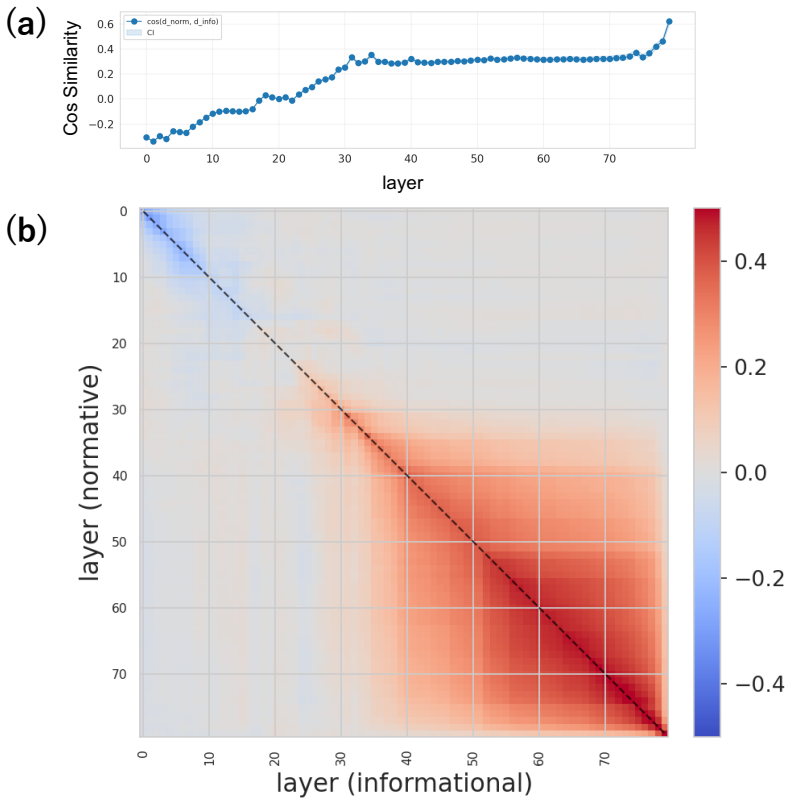

Fig. 5. (a) Cosine similarity between difference vectors of informational and normative conformity at each layer, (b) cosine similarity of difference vectors across layers.

5 Discussion and Conclusion

In this study, we confirmed that in four out of six LLMs, conformity tends to increase when voting is public. This result suggests that LLMs exhibit normative conformity. Furthermore, the normative conformity observed here can be strengthened by factors such as "subsequent evaluation" and "continuity of relationship", consistent with results previously obtained in human subject experiments [9, 14]. Interestingly, the extent to which each factor affected conformity differed greatly across LLMs.

As RQ2, we examined the possibility that manipulating social context leads to changes in conformity toward specific targets (speakers). The experimental results confirmed that manipulations such as "peer endorsement" and "assignment of attributes of an influential speaker" increase conformity toward specific speakers. We consider this a meaningful result in that it indicates the risk that, in decision-making by groups of LLM agents, minor contextual modifications by

some users may make it possible to manipulate the influence of specific speakers. The latter result further showed the risk that a certain attribute (e.g., shirt color), although not directly related to decision-making and not directly related to the person's influence, may change into a status that signals influence. In social psychology, status construction theory [21] points out that when others observe a person with a particular attribute unrelated to decision-making (such as race or gender) behaving in a dominant manner, that attribute may become established as "status." The present results suggest that such phenomena may also occur in groups of LLMs and may potentially distort influence in an unjustified manner.

Finally, the results of RQ3 suggest that even when the externally observable behavior of "conformity" is the same, differences may arise in the internal representations of LLMs when the underlying purpose or motivation differs. In humans as well, it is known that normative and informational conformity involve different neural mechanisms in the brain [18], and future work will examine whether analogous differences exist in the processing within LLMs.

Although conformity is often associated with risks such as groupthink [17], recent studies highlight its positive effects. Informational conformity enhances employees' innovative performance [5], whereas normative conformity, despite its negative impact on innovation, may promote cooperative behavior in social networks [16]. These findings suggest that the effects of informational and normative conformity on group decision-making are context-dependent. Future work may explore how such effects can be selectively suppressed or enhanced using approaches such as representation engineering [25], contributing to improved decision-making in agent collectives

Appendix

See the full version including the appendix at: https://arxiv.org/abs/2604.19301.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aronson, E., Wilson, T.D., Akert, R.M.: Social Psychology. Pearson, New York, NY, 7 edn. (2010)
2. Asch, S.E.: Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow, H. (ed.) Groups, Leadership, and Men, pp. 177–190. Carnegie Press, Pittsburgh, PA (1951)
3. Asch, S.E.: Studies of independence and conformity: I. a minority of one against a unanimous majority. Psychological Monographs: General and Applied **70**(9), 1–70 (1956). https://doi.org/10.1037/h0093718
4. Barnfield, M.: Think twice before jumping on the bandwagon: Clarifying concepts in research on the bandwagon effect. Political Studies Review **18**(4), 553–574 (2020). https://doi.org/10.1177/1478929919870691

5. Chang, Y.Y., Wannamakok, W., Lin, Y.H.: Work conformity as a double-edged sword: Disentangling intra-firm social dynamics and employees' innovative performance in technology-intensive firms. Asia Pacific Management Review **28**(4), 439–448 (2023). https://doi.org/https://doi.org/10.1016/j.apmrv.2023.01.003, https://www.sciencedirect.com/science/article/pii/S1029313223000039
6. Chen, Z.Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L., Wang, W.Y.: A survey on large language models for critical societal domains: Finance, healthcare, and law (2024), https://arxiv.org/abs/2405.01769
7. Cho, Y.M., Guntuku, S.C., Ungar, L.: Herd behavior: Investigating peer influence in llm-based multi-agent systems (2025), https://arxiv.org/abs/2505.21588
8. Choi, M., Kim, K., Chae, S., Baek, S.: An empirical study of group conformity in multi-agent systems. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 5123–5139. Association for Computational Linguistics (2025), https://aclanthology.org/2025.findings-acl.265/
9. Cottrell, N.B.: Social facilitation. In: McClintock, C.G. (ed.) Experimental Social Psychology, pp. 185–236. Holt, Rinehart and Winston, New York, NY (1972)
10. Deutsch, M., Gerard, H.B.: A study of normative and informational social influences upon individual judgment. Journal of Abnormal and Social Psychology **51**(3), 629–636 (1955). https://doi.org/10.1037/h0046408
11. Durmus, E., Nguyen, K., Liao, T.I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., Ganguli, D.: Towards measuring the representation of subjective global opinions in language models (2023), https://arxiv.org/abs/2306.16388
12. Gallegos, I.O., Rossi, R.A., Barrow, J., Ahn, S., et al.: A survey of bias and fairness in large language models (2024), https://aclanthology.org/2024.cl-1.8/
13. Hadar Shoval, D., Gigi, K., Haber, Y., Itzhaki, A., Asraf, K., Piterman, D., Elyoseph, Z.: A controlled trial examining large language model conformity in psychiatric assessment using the asch paradigm. BMC Psychiatry **25**, 478 (2025). https://doi.org/10.1186/s12888-025-06912-2, https://link.springer.com/article/10.1186/s12888-025-06912-2
14. Hancock, R.D., Sorrentino, R.M.: The effects of expected future interaction and prior group support on the conformity process. Journal of Experimental Social Psychology **16**(3), 261–269 (1980). https://doi.org/10.1016/0022-1031(80)90069-4
15. Handler, A., Larsen, K.R., Hackathorn, R.D.: Large language models present new questions for decision support. International Journal of Information Management **79**, 102811 (2024). https://doi.org/10.1016/j.ijinfomgt.2024.102811
16. Huang, C., Li, Y., Jiang, L.: Dual effects of conformity on the evolution of cooperation in social dilemmas. Phys. Rev. E **108**, 024123 (Aug 2023). https://doi.org/10.1103/PhysRevE.108.024123, https://link.aps.org/doi/10.1103/PhysRevE.108.024123
17. Janis, I.L., et al.: Groupthink. IEEE Engineering Management Review **36**(1), 36 (2008)
18. Mahmoodi, A., Nili, H., Bang, D., Mehring, C., Bahrami, B.: Distinct neurocomputational mechanisms support informational and socially normative conformity. PLOS Biology **20**(3), 1–21 (03 2022). https://doi.org/10.1371/journal.pbio.3001565, https://doi.org/10.1371/journal.pbio.3001565

19. Mehdizadeh, A., Hilbert, M.: When your AI agent succumbs to peer-pressure: Studying opinion-change dynamics of LLMs (2025), https://arxiv.org/abs/2510.19107
20. Navigli, R., Conia, S., Ross, B.: Biases in large language models: Origins, inventory and discussion. Journal of Data and Information Quality **15**(2), 1–21 (Jun 2023). https://doi.org/10.1145/3597307
21. Ridgeway, C.L.: Status Construction Theory. In: The Wiley Blackwell Encyclopedia of Race, Ethnicity, and Nationalism, pp. 1–3. John Wiley & Sons, Ltd (2015), https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118663202.wberen200
22. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: Advances in Neural Information Processing Systems (NeurIPS 2023) (2023), https://arxiv.org/abs/2303.17548
23. Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H.W., Chowdhery, A., Le, Q.V., Chi, E.H., Zhou, D., Wei, J.: Challenging BIG-bench tasks and whether chain-of-thought can solve them (2022), https://arxiv.org/abs/2210.09261
24. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8
25. Wehner, J., Abdelnabi, S., Tan, D., Krueger, D., Fritz, M.: Taxonomy, opportunities, and challenges of representation engineering for large language models. arXiv preprint arXiv:2502.19649 (2025)
26. Weng, Z., Chen, G., Wang, W.: Do as we do, not as you think: the conformity of large language models (2025), https://arxiv.org/abs/2501.13381
27. Zhu, X., Zhang, C., Stafford, T., Collier, N., Vlachos, A.: Conformity in large language models (2024), https://arxiv.org/abs/2410.12428