# Peer-review report of

Bellina, A., De Marzo, G. & Garcia, D. (2026). Prompt-mediated perceptual conformity in multimodal language models follows Social Impact Theory. *advances.in/psychology*, 1, e680902. https://doi.org/10.56296/aip00072

## Contents

1. Round 1
2. Round 2
3. Round 3

# Round 1

## Editor

Date: 2026-07-28

Dear Authors,

Thank you for submitting your manuscript “Conformity and Social Impact on AI Agents” to our Special Issue on the Psychology of Agentic AI in *advances.in/psychology*. I was fortunate to receive detailed and constructive feedback from two reviewers with core expertise in the topic of your work, and I thank them both for their thorough reviews. In line with journal standards, I also read and evaluated your manuscript independently before turning to their comments.

The manuscript has several clear strengths. Reviewer 1 notes that “The experiments are systematic, the results are presented clearly, and the paper provides a useful characterization of how socially framed context can move otherwise accurate model responses toward an incorrect answer.” Similarly, Reviewer 2 writes “overall, I have enjoyed reading this piece” and describes the work as “well motivated and of high quality.” My own reading converges with these assessments, and the contribution fits the scope of our special issue well.

At the same time, both reviewers identify a number of valuable ways in which the paper could be strengthened. Since they lay out their concerns clearly, I will not repeat each comment here. I do ask, however, that you respond to all of their points in an itemized response letter that includes excerpts illustrating your changes. Below, I highlight the comments I consider particularly important and then add some observations of my own.

Reviewer 1 asks you to position your work more clearly within the body of studies that have tested the same phenomenon, while also acknowledging its novelty: “The main contribution is not the finding that language models can shift toward incorrect majority answers, since this has already been shown. Rather, the contribution appears to consist in the controlled extension to multimodal perceptual tasks and in the systematic investigation of variables connected to Social Impact Theory.” In line with this comment, I ask you to make the following changes.

First, please revise your title. It is currently somewhat opaque and does not clearly convey your findings or the originality of your study. I suggest something along the lines of “Perceptual conformity in multimodal language models: Prompted majorities override correct answers and scale with Social Impact Theory variables.” This should also help address the same reviewer’s comment on the scope of your claims.

Second, please position your work clearly in the introduction relative to existing research and spell out the specific open questions you answer. Your results on moderators are very rich, and their application from a Social Impact Theory perspective is clearly original, so this should be a fairly manageable task.

Finally, the discussion currently, and strikingly, contains not a single reference. Please revisit the existing literature there (including work on broader LLM misalignment) and explain how your findings confirm and extend previous research.

Reviewer 2 also makes a set of very important points. The first is that your three examples, while clearly relatable to a social science audience, are very well known and therefore likely part of the models' training data. As the reviewer puts it, "As a result, the LLM output might not be driven by conformity alone, but merely by an internal, learned association between a high number of wrong confederate answers and a high probability of providing a wrong answer." They also offer a solution, namely designing a novel additional experiment. I agree that this is important, but I suggest you stay within the perceptual domain when doing so. As Reviewer 1 points out, this is one of the novelties of your work, and when ruling out this confound you do not want to change another critical element (the perceptual nature) of the experiment.

Whether you take up Reviewer 1's suggestion to test prompt mitigations is up to you. Doing so would add value to the paper and strengthen its novelty, which Reviewer 2 questioned to some extent, and would therefore probably address concerns from both reviewers at once.

Please also address the reviewer's counterbalancing concerns. In addition, the natural language presentation of confederates comes closest to emulating the human experiments, but as the reviewer states, it may introduce additional uncertainty that simply stating the choice avoids. To be clear, I do not think the natural language approach undermines the value of your experiment, but it sets a constraint that you should discuss. Also, rather than re-running all experiments, you could run one targeted study in which you vary the confederate input style along the lines the reviewer suggests.

Finally, Reviewer 2 raises a substantive methodological concern about the task difficulty result. In the visual tasks, difficulty is defined by the distance between the correct and incorrect options, which means the same parameter also determines how large an error the model makes when it conforms. As a result, it is not possible to tell whether conformity rises because the task is harder or simply because the wrong answer is nearly right. Please address this by (i) adding the requested scatterplot of uncertainty against task difficulty to the supplement (summarizing it in the main text), and (ii) either reporting an additional experiment in which difficulty and the margin between answer options are varied independently, along the lines the reviewer suggests, or revising the interpretation of the difficulty result to acknowledge this confound explicitly In addition to these points, please address the following.

Beyond these points, please address the following.

GENERAL

In the introduction, please justify the model selection beyond the models being open source and multimodal. In the discussion, consider how this selection may have influenced the results.

RESULTS

Referring to Figure 2, you write "When the number of correct and incorrect responses is balanced, the conformity bias is almost entirely suppressed." However, if I am reading the correct graph (Figure 2, Panel C), the error rates are well above 0.

Please add 95% CIs to all figures where applicable.

The minimal group experiment results are very interesting. Please state in the caption of Figure 4d that it reflects the minimal group experiments, as this was somewhat unclear.

Please move the results on distance in space and time from the Supplementary Materials to the main text.

Although you leave internal processes to future work, I would be curious to see what explanations or justifications the models give for their responses. These could, for instance, be analyzed by topic or in a similar way.

For Panels B and C of Figure 2, please show the results across models. Focusing on a single model here does not seem justified. You can then refer to the model-specific results in the Supplementary Materials. The same applies to Figure 3, Panel A, where it is also unclear why you focus on one model. Either present lines and data points for each model in this panel, or a single line with all data points pooled across models.

ADDITIONAL TESTS

You find marked differences between chatbot confederates and human confederates. In reality, future teams may consist of both. It would therefore be interesting to explore what happens when humans and chatbots diverge (two conditions: either the humans or the chatbots are correct) versus when they converge (perhaps also two conditions, serving as baselines: all correct vs. all incorrect).

You find relatively low error rates when confederates are described as chatbots. I would be very interested to see whether confederates being (a) from the in-group model, (b) a single out-group model, or (c) various out-group models (i.e., indicating cross-model convergence) would produce different effects.

FORMATTING AND STYLISTIC ISSUES

Per journal standards, please place the methods section before the results section.

In lines 93 to 97 of the introduction, please do not foreshadow the results.

Please change the citation and reference style to APA 7. For all journal articles, provide DOIs in URL format in the reference list.

Please refer to figures with “Figures” not “Fig.”

The contrast between 0% wrong and 80% wrong in Figure 2 is too small. Please increase it or, preferably, choose other colors to improve readability. For all figures, also strongly consider using different point symbols to aid comprehension (currently you use only dots).

Please refer to “Supplementary Materials” rather than “Supplementary Information”.

In the figure captions, indicate which confidence intervals are presented (presumably 95% CIs).

SUMMARY

In sum, we all see clear value in your work, and the reviewers’ comments and my own are intended to help an already strong and rich manuscript realize its full potential. I therefore invite you to submit a revised version within six weeks. Thank you again for submitting your stimulating work to our special issue. It was a pleasure reading it.

Best,

Jonas R. Kunst

Editor-in-Chief

## Reviewer 1

The manuscript investigates conformity-like behavior in multimodal language models by means of controlled visual tasks. It examines how model responses depend on majority size, unanimity, task difficulty, source characteristics, response visibility, and social proximity. The experiments are systematic, the results are presented clearly, and the paper provides a useful characterization of how socially framed context can move otherwise accurate model responses toward an incorrect answer. In my view, the manuscript is publishable provided that the authors address the following comments.

**1. Title and scope of the claims:** An important limitation in comparison with the original Asch experiments is that the social interaction is compressed into a prompt. Human participants were not simply informed that other participants disagreed with them. Rather, they witnessed other participants giving their answers in a shared experimental setting. In the present study, by contrast, a single model receives a textual representation of previous responses. There is no actual interaction.

The manuscript acknowledges this prompt-based approximation. Nevertheless, the title "*Conformity and Social Impact on AI Agents"* seems much broader than what is directly demonstrated. The experiments establish sensitivity to socially framed prompt context, rather than social impact arising in a system of interacting agents. The title should reflect this distinction more precisely, focusing on the main contribution of this paper versus previous results, see below, rather than aiming at leaving a memorable impression. The abstract, the introduction and the conclusion should also significantly moderate the claims concerning artificial societies, multi-agent dynamics, and collective AI deployments.

**2. Positioning relative to previous work:** The most closely related studies, Refs. 59–60, are cited, but not discussed in sufficient detail. Since these works address the same central phenomenon, the manuscript should compare them more explicitly with the present study. In particular, the authors should clarify which tasks and experimental settings were used previously, how conformity was operationalized, which moderators were already investigated, and what is conceptually and empirically new here.

The main contribution is not the finding that language models can shift toward incorrect majority answers, since this has already been shown. Rather, the contribution appears to consist in the controlled extension to multimodal perceptual tasks and in the systematic investigation of variables connected to Social Impact Theory. This point should be stated clearly. The new extensive "Related work" section should also take into account relevant papers citing and building on Refs. 59–60.

**3. Interpretation of conformity-like behavior.** The present design cannot distinguish a genuinely social sensitivity from in-context learning, instruction following, or learned semantic associations between consensus, authority, and correctness. For example, stronger deference to scientists or

judges does not necessarily indicate a social-cognitive mechanism. Likewise, the public-versus-private manipulation does not establish a human-like motivation to avoid rejection.

The manuscript should therefore maintain a clear distinction between behavioral similarity and mechanistic equivalence. The strongest conclusion supported by the results is that socially framed prompt variables systematically influence model outputs in patterns resembling human conformity. Terms such as “social pressure,” “normative influence,” and “social identity” should be qualified accordingly.

If these changes in framing, positioning, and interpretation are implemented seriously and not superficially, I am happy to support acceptance.

## Reviewer 2

The paper studies conformity in AI Multi-Agent Systems (MAS), namely the behavioral tendency of AI agents to adjust their outputs to align with the observed behavior of other agents within the system. The specific angle the authors adopt is that of visual experiments from the classic social psychology literature, operationalized with large multimodal AI models as experiment participants. The experimental setup is simple: a single agent is asked to answer a factual question about an input image, while being given auxiliary information on how other agents (called "confederates") have responded to the same question. The key intuition from social psychology is that people become more likely to select the wrong answer as the number of confederates providing that wrong answer increases. The auxiliary information on the confederate responses can be fully controlled by the experimenter, allowing the authors to observe how the conformity effect varies with the three mediating factors posited by Social Impact Theory: 1) the number of confederates, 2) their perceived authoritativeness, and 3) their closeness along the social, spatial, or temporal dimensions.

The authors experiment with three visual tasks: line judgment (comparing the length of two lines), color recognition (comparing the hue of two color patches), and dot estimation (comparing the cardinality of two clouds of points). They run the experiments on multimodal AI models from four distinct families (Qwen, Gemma, Ovis, Mistral), computing the probability of error of an agent (p_error) from the model logit associated with the wrong option. They then examine how p_error varies across experimental conditions. The results reveal that, across models, conformity increases with majority size, perceived authoritativeness, and closeness, as Social Impact Theory predicts. Conformity is also stronger for more difficult tasks.

The work is positioned within the broad literature on auditing the capabilities and biases of AI models. The study of conformity, a special case of social influence, is important for refining our understanding of how MAS operate, especially given that such systems are already deployed across multiple application settings. The work therefore rests on a valuable premise. In addition, the multimodal setup (model reasoning based on visual input) is quite rare in this branch of the literature.

Overall, I have enjoyed reading this piece. While the work is well motivated and of high quality, a few aspects related to the experimental setup and the interpretation of results might benefit from improvement.

**Experimental setup**

1. Adapting well-known experiments from social psychology to a fully synthetic setting is an added value of the contribution, as it allows for direct comparison with results from human participants. However, this choice raises the potential issue of model memorization influencing the results, an issue that many contributions in this line of work are potentially subject to. LLMs are known to memorize patterns from training data (see, e.g., Carlini et al.'s seminal "Extracting Training Data from Large Language Models"). One could argue that, during training, the models were exposed to a considerable amount of information (Wikipedia pages, scientific articles, etc.) about these specific experiments and about the responses typically observed from human participants. As a result, the LLM output might not be driven by conformity alone, but merely by an internal, learned association between a high number of wrong confederate answers and a high probability of providing a wrong answer. To account for this, I would recommend running at least one additional experiment using a task that is not associated with any famous prior experiment. This task doesn't need to be multimodal in nature, and could be a math task (e.g., identifying the number closest to a reference number) or a task of counting the instances of a specific letter in a sentence, with two possible answers provided. Such tasks could also be easily tuned to cover different levels of difficulty. This additional experiment would considerably strengthen the conclusions, as it would provide reasonable evidence for rejecting the memorization hypothesis.
2. Exploring the effect of different prompt structures on the results is typically a wasteful investment of resources, as results are very often not significantly affected by prompt variations. In this case, however, I believe that some attempts to steer the observed outcome by modifying the prompt would be informative. The current prompt is quite bare, and one might wonder whether the conformity bias could be easily corrected by providing instructions on how to parse the information coming from the confederates. Conveniently, the authors could motivate these changes within the theoretical framework they present. For example, they could design a prompt that nudges the LLM toward informational conformity, suggesting that the model weight its own opinion against the confederates' opinions based on uncertainty. A prompt along the lines of "provide the response that is most objective, weighting your own assessment against the assessment of your peers" could serve this purpose. I anticipate that the authors would argue that adding complexity to the prompt would move the experimental setup away from the corresponding classic setup with humans, which is beyond their intended scope. However, given the strong practical implications of conformity for the security of AI MAS that are discussed in the manuscript, I would argue that it is quite important to understand whether unwanted conformity can be easily avoided through minor and reasonable modifications to the prompt.
3. The public vs. private setting seems quite artificial, as the prompt provides no context on why sharing the response publicly rather than keeping it private matters for the specific task. I therefore find this part of the experiment somewhat less informative. Here again, the prompt could be expanded to make that choice appear more consequential (e.g., suggesting that "the confederates could reconsider their opinion by looking at your answer"), making it clear

that the final objective of the multi agent system is to "provide a majority vote that is as close as possible to the true answer."

4. Neither the main text nor the SI makes it clear whether there is a fixed association between the letter marking each of the two options ('A', 'B') and the correctness of the option. I would expect the LLM logits to be influenced by 1) the correctness of the answer, 2) a positional bias related to how the option is presented in the image, and 3) a lexicographic bias toward selecting the letter "A" over the letter "B". Accounting for the two biases (visual and lexicographic) is necessary to ensure that the results are not substantially affected by them. The minimal intervention would be to repeat the experiments while switching the association between letter and option. In the line judgment task, for example, this would mean a new image in which A is presented on the left and associated with the longer line, and B is presented on the right and associated with the shorter line. One could of course run all four combinations of letter position (e.g., A on the left vs. right) and option position (e.g., short line on the left vs. right), but that is probably overkill if the first test shows no change in the results.I wonder why the confederate responses are presented in the prompt using a short natural language sentence. From an application perspective, a natural language sentence makes sense if it contains some *reasoning* for the judgment, but in this case the wording is at best redundant and at worst may introduce an unintended notion of uncertainty that could be non-uniform across the selected options. I would recommend removing the natural language entirely and simply reporting the letters the confederates have chosen (e.g., "reply 1: A; reply 2: B"). In line with the previous comment, the order of the confederate responses in the prompt should be randomized to account for positional biases.

**Interpretation and presentation of results**

1. Overall, the results and the authors' interpretation of them make sense to me. However, I was puzzled by the result on the relationship between conformity, uncertainty, and task difficulty. Why would conformity be higher for higher task difficulty, regardless of uncertainty? The authors do not appear to provide an answer to this question. My hypothesis is that the effect of task difficulty on conformity is confounded by the shrinking margin of error of the wrong option as the level of "difficulty" increases. In the line judgment task, for example, if the "wrong" line is very close in length to the "correct" line, either could be considered correct within a small margin of error, which is why the LLM might be more willing to adopt the group's behavior. The magnitude of the error needs to be disentangled from the difficulty of the task somehow. I am not sure how this could be achieved in the visual experiments proposed, but it could be done easily with other simple experiments. For instance, one could assign the LLM a math problem with two possible answers. The experiment could vary the difficulty of the problem (e.g., a simple addition vs. a multiplication of two large numbers) and the error of the two responses (two answers differing widely vs. two answers differing by a small delta). In addition, for the sake of clarity and transparency, I would like to see a scatterplot of uncertainty vs. task difficulty.
2. The authors claim that they observe a steady increase in probability of error (line 169), in contrast to the saturation effect seen in humans. However, the figure does appear to show

some saturation effect in several models, as the authors acknowledge in the caption. I would rephrase this part to make it consistent.

3. For a more complete overview of the results, it would be useful to see a heatmap with the number of confederates on the x-axis, the fraction of wrong answers on the y-axis, and the error probability mapped to the cell color.

---

# Round 2

## Editor

Date: 2026-10-28

Dear Authors,

Thank you for submitting your revised manuscript, “Perceptual conformity in multimodal language models follows Social Impact Theory,” to the Special Issue “Psychology of Agentic AI” in advances.in/psychology, together with your detailed response letter.

You have made an impression set of revisions, thoroughly addressing all of our concerns, with one exception outlined below. What remains are two minor issues that I am confident you can address within short time.

1. One reviewer still requests a different title. They point out that “perceptual conformity” describes the tasks but it does not communicate the way in which social influence reaches the models. Your experiments tell agents via the prompt how others responded. They do not observe or interact with them. You now clearly covey this in the abstract and introduction, but not in the title. I therefore suggest the title

“Prompt-mediated perceptual conformity in multimodal language models follows Social Impact Theory”

You are welcome to propose a different wording, provided it makes the prompt-mediated nature of the experiments and of the observed effect explicit. To avoid losing time, please email me your chosen title as soon as you have settled on it, unless you chose the one provided above.

2. Please provide all figures that will be presented in the main text as SVG files in your latex ZIP-file.

Please resubmit these changes within five days, by Monday, 2 November 2026. I will assess the revision myself, without a further round of peer review.

Thanks again for this thorough and detailed revision. I am looking forward to what will likely be the final version.

Best,

Jonas R. Kunst

### Reviewer 1

The authors have addressed my comments, and I appreciate the substantial effort they have made.

As a final point, I remain concerned about the title. The term "perceptual conformity" describes the tasks but does not resolve my previous concern. In the present study, models are told through a prompt how other participants have responded rather than directly observing or interacting with them.

This distinction should be reflected in the title, which should make explicit the prompt-mediated nature of both the experiments and the observed effect. For example, terms such as "prompt-mediated conformity," "prompt-mediated social influence," or another formulation explicitly referencing prompts would more accurately convey what the experiments establish and avoid suggesting a direct equivalence with conformity in interacting groups.

With this important final edit, I would be happy to recommend acceptance.

### Reviewer 2

I thank the Authors for their extensive response, that left none of my comments unaddressed. The supplementary sections S4-S7 provide satisfactory answers to all my points, and I believe shed new light on the mechanics of the process of conformity. I appreciated in particular the effort of rerunning the experiment for the MNIST scenario; that was really necessary for the piece to survive easy criticism about mere memorization. I don’t have further comments and I am happy to recommend the paper for acceptance.

---

## Round 3

### Editor

Date: 2026-10-29

Dear Authors,

Thanks a lot for making these last changes. I am happy to accept the manuscript for publication. Congrats on a really impressive contribution to the field! I am very glad that it will be the first paper published in the special issue.

Best,
Jonas