Advances.in
Reinventing Academic Publishing

Supplementary Materials for

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

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

S1. Explicit Prompts

S1.1. Baseline Prompts for the Three Visual Tasks

All experiments share the same baseline prompt structure, which is presented to the model in the absence of any social influence. The format is identical across tasks, except for the description of the visual elements.

- **Line judgment task:**

  ```
  In the image, there are three vertical lines labeled A, REFERENCE LINE, and B.
  Which of the lines, A or B, is the same length as the REFERENCE LINE?
  Provide only the final answer, either A or B, without generating anything else.
  ```

- **Color recognition task:**

  ```
  In the image, there are three colored squares labeled A, REFERENCE COLOR, and B.
  Which of the squares, A or B, has the same color as the REFERENCE COLOR?
  Provide only the final answer, either A or B, without generating anything else.
  ```

- **Dots estimation task:**

  ```
  In the image, there are three boxes labeled A, REFERENCE BOX, and B.
  Which of the boxes, A or B, contains the same number of black dots as the REFERENCE BOX?
  Provide only the final answer, either A or B, without generating anything else.
  ```

Neutral Confederate Responses

To create the prior-response context, we inserted before the final question a block of sentences corresponding to the replies of N previous participants. Each reply is randomly selected from a fixed pool of ten neutral expressions, all semantically equivalent and differing only in wording. This design ensures linguistic variability while maintaining neutrality and avoiding stylistic or emotional cues.

Below is the complete list of possible response templates used to construct the prior-response prompts:

- `I think the answer is [A/B].`
- `My answer is [A/B].`
- `I'm going with [A/B].`
- `It could be [A/B].`
- `I believe the answer is [A/B].`
- `I'll say [A/B].`
- `[A/B] is my choice.`
- `[A/B], I think.`
- `[A/B] seems right.`
- `I'm leaning toward [A/B].`

In each trial, N replies are sampled at random from this list (with replacement) and introduced sequentially in the prompt under the header (see below). This section precedes the question, e.g., before “Provide only the final answer”.

```
The participants before you provided the following replies:

reply 1: I think the answer is [A/B].
reply 2: I'm going with [A/B].
reply 3: I believe the answer is [A/B].
...
```

This randomization procedure generates natural linguistic variability across trials while ensuring that all responses convey the same informational content. The order in which the replies appear is randomized at every trial: correct and incorrect replies are inserted at random positions rather than appended, so that the position of a reply in the list carries no information.

Prompts for Response Visibility

To manipulate the perception of response visibility, we introduced a framing statement immediately after the question ("Which of the lines/squares/boxes...?") and before the list of replies provided by the other participants. Two sets of equivalent sentences were used: one describing the *public condition*, where the model is told that its answer will be shared with others, and one describing the *private condition*, where it is told that its answer will remain unseen. Each trial randomly selects one sentence from the corresponding list to ensure linguistic variability while maintaining identical semantic meaning. The ten sentences listed here for each condition are the first ten of the one hundred used in the experiments.

Public condition (response *visible*):

- "Your response will be shared with other participants."
- "Other participants will be able to see your answer."
- "Your answer will be visible to the group."
- "Everyone in the group will see your response."
- "Your response will be displayed to others."
- "The rest of the group will view your answer."
- "Your answer will be shown to the other participants."
- "The group will have access to your response."
- "Other participants will know your answer."
- "Your submission will be visible to everyone."
- …

Private condition (response *not visible*):

- "Your response will not be shared with other participants."
- "Other participants will not be able to see your answer."
- "Your answer will not be visible to the group."
- "No one in the group will see your response."
- "Your response will not be displayed to others."
- "The rest of the group will not view your answer."
- "Your answer will not be shown to the other participants."
- "The group will not have access to your response."
- "Other participants will not know your answer."
- "Your submission will not be visible to everyone."
- …

This framing block is inserted immediately before the section listing the replies of previous participants (the prior-response segment). In all other respects, the prompt structure remains identical to the baseline version.

Prompt Structures for Social Impact Theory Experiments

In the Social Impact Theory experiments, the prompts were modified to manipulate the perceived strength and immediacy of the social sources. Four main conditions were implemented: *source strength*, *nationality*, *group identity*, and *spatial/temporal proximity*. In all cases, the modification affects either the introductory statement or the section listing the replies of previous participants.

Source Strength Condition

The generic label `participants` is replaced by a social role, which is the only element of the prompt that changes. The roles are:

- kids
- chatbots
- policemen
- military officers
- scientists
- judges
- experts

The line introducing the replies of the other participants becomes:

```
The [role] before you provided the following replies:
```

Nationality Condition

A nationality is randomly assigned to the model at the beginning of each experiment, drawn from the following list:

- Italian
- French
- German
- American
- Brazilian
- Chinese
- Indian
- South African
- Russian
- Japanese

At the beginning of the prompt, the model is told its assigned nationality:

```
Your nationality is [nationality1].
```

Then, the replies of the other participants are introduced as:

```
The participants before you provided the following replies:
[nationality2]: reply 1: I think the answer is B.
[nationality2]: reply 2: My answer is B.
...
```

In the *same-group* condition, `nationality2 = nationality1`. In the *different-group* condition, `nationality2` is randomly drawn from the remaining nationalities, ensuring variability across trials.

Minimal Group Condition

To test in-group bias without explicit identity attributes, we apply a minimal group paradigm. Two random letters are drawn at each trial to define two abstract groups. The model is told its group identity at the beginning of the prompt:

```
You will be divided into groups with other participants.
Your group is [group1].
```

The confederate replies are then introduced as:

```
The participants before you provided the following replies:

[group2]: reply 1: I think the answer is B.
[group2]: reply 2: My answer is B.
...
```

In the *same-group* condition, `group1 = group2`. In the *different-group* condition, `group2` is a different randomly selected letter.

Spatial and Temporal Proximity Conditions

The spatial and temporal proximity experiments modify only the section introducing the replies, adding an indication of each participant's distance or response time. Within a trial all replies share the same distance or delay: that value is the experimental condition.

Spatial proximity:

```
The participants before you provided the following answers.
Each participant is located at a specific distance from you,
which is indicated before their response:
```

Eleven levels of distance are used: `here`, then `one meter away`, `two meters away`, and so on in one-metre steps up to `ten meters away`.

Temporal proximity:

```
The participants before you provided the following answers.
Each participant responded at a different time,
which is indicated before their response:
```

Eleven levels of delay are used: `right now`, then `2 minutes ago`, `4 minutes ago`, and so on in two-minute steps up to `20 minutes ago`.

As in the main text, two main conditions are compared: a *near (now) condition*, where all replies are described as occurring "here" or "right now," and a *distant (before) condition*, where participants are described as farther in space or time. As Figure S4 shows, the effect lies between the near condition and any declared distance or delay, and changes little beyond it: for this reason the main text compares two conditions, the near one and the average of the ten distant levels.

S2. Results Across Models and Tasks

In the following, we report the complete set of results across all models and visual tasks, complementing the analyses presented in the main text. Twelve models are measured on color recognition and ten on line judgment and dots estimation, where Qwen2 7B and Qwen2.5 3B do not reach the required baseline accuracy (Methods). We call each of these 32 combinations of a model and a task a *cell*, and the results below are counted over cells.

Figure S1. **Error probability as a function of group size, for every model and every task.** Probability of giving the wrong answer, p_{wrong}, against the number of confederates providing incorrect responses, N. **(a)** Line judgment, **(b)** color recognition, **(c)** dots estimation. One line per model; colour and symbol identify the family and the line style the model within it. Individual curves carry no confidence band, for legibility; the curves pooled over the models, with their intervals, are in Figure 2b of the main text.

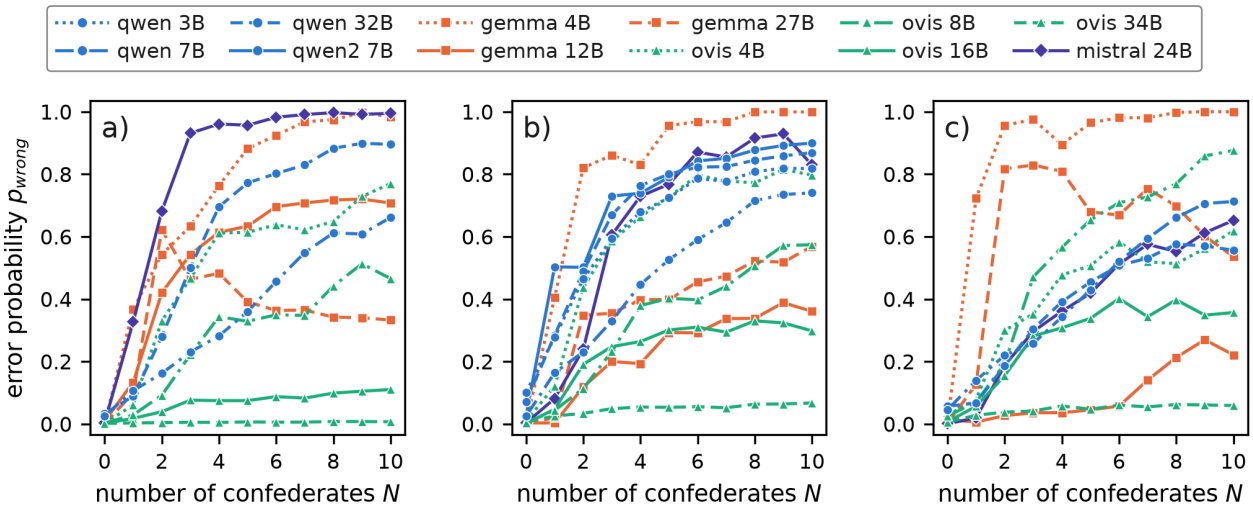


Figure S2. **Error probability over the whole space of majority compositions, one panel per task.** Colour is the probability of giving the wrong answer as a function of the number of confederates, N, and of the fraction of them answering incorrectly. **(a)** Line judgment, **(b)** color recognition, **(c)** dots estimation. Each panel pools the models measured on that task, weighted by the number of stimuli. The plane is a staircase rather than a regular grid because with N confederates only the fractions k/N exist, so each rectangle is one attainable composition; for display it extends halfway to the neighbouring attainable fractions, so a fraction that cannot occur with a given N (a quarter, with two confederates) takes the value of the nearest one that can. The same plane pooled over the three tasks is Figure 2c of the main text. Pooled over the three tasks and over the group sizes at which an even split exists, the error probability is 0.16 at a balanced composition against 0.50 under unanimity.

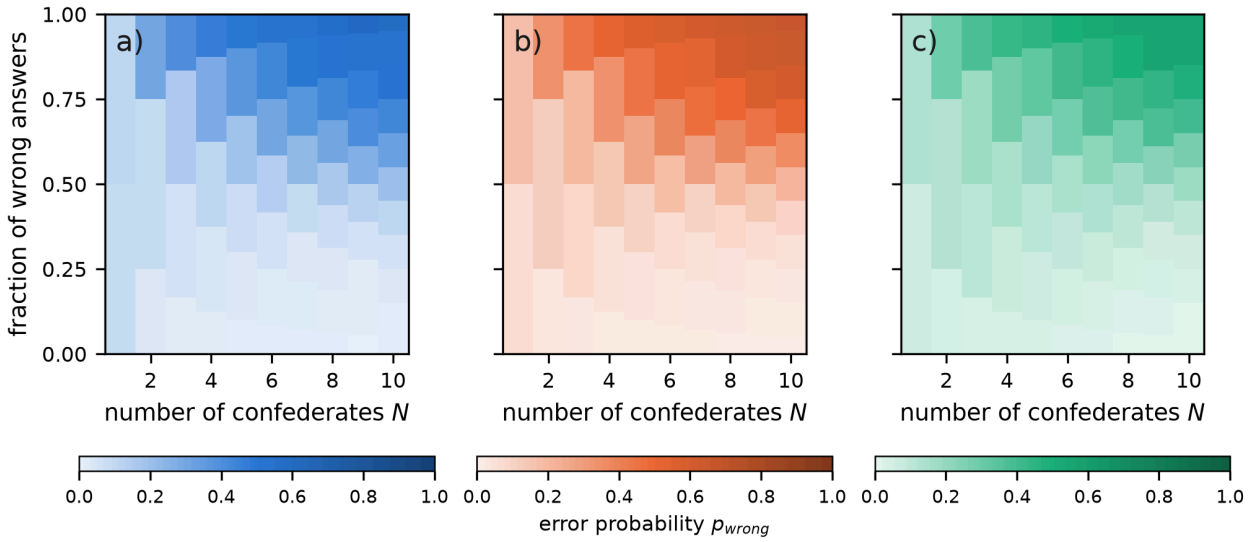


Table S1. **Conformity bias as a function of model performance and task difficulty across visual tasks.** Results of linear regressions of normalized conformity bias (AUC of $p_{\text{wrong}}(N)$) on normalized model performance, the log-odds of the correct answer without confederates, and task difficulty, each alone and both together. All three are min-max normalized within each model and task, and oriented so that larger values mean a better-performing model and a harder task. The 346 points are the difficulty levels at which each model and task was measured; each regression has one intercept per model and task, and its standard errors are clustered by model and task. Both are significant predictors, and model performance is the stronger: taken together, it keeps three quarters of the effect it has alone, and difficulty about half. Fitted separately for each model and task, the joint regression gives a negative coefficient for performance in 30 of the 32 combinations (distinguishable from zero in 16) and a positive one for difficulty in 24 (distinguishable in 10).

	each alone				together			
Variable	β	**SE**	**t**	**p**	β	**SE**	**t**	**p**
model performance	$\mathbf{-0.786}$	0.055	$\mathbf{-14.24}$	3.8e-15	$\mathbf{-0.590}$	0.099	$\mathbf{-5.94}$	1.5e-06
task difficulty	$\mathbf{+0.643}$	0.086	$\mathbf{+7.44}$	2.2e-08	$\mathbf{+0.292}$	0.103	$\mathbf{+2.83}$	8.0e-03

Figure S3. **The four moderators of the conformity bias, for every model and task.** Each panel is one manipulation and each point one model and task, sorted by the size of the effect, so that the order of the models and tasks differs from panel to panel. **(a)** The response is described as visible to the others rather than private. **(b)** The confederates share the model's nationality rather than carrying a different one. **(c)** They are described as co-present rather than at a distance. **(d)** Their replies are described as given now rather than earlier. Colour and symbol identify the model family. Vertical bars are 95% confidence intervals obtained by resampling stimuli; a filled symbol marks an effect whose interval excludes zero, that is, one distinguishable from zero. Medians over the 32 combinations: (a) +0.018, (b) +0.228, (c) +0.081, (d) +0.059, positive in 27, 32, 29 and 29 of them respectively. In (a) the two conditions are compared over the one hundred prompt reformulations described in the Methods.

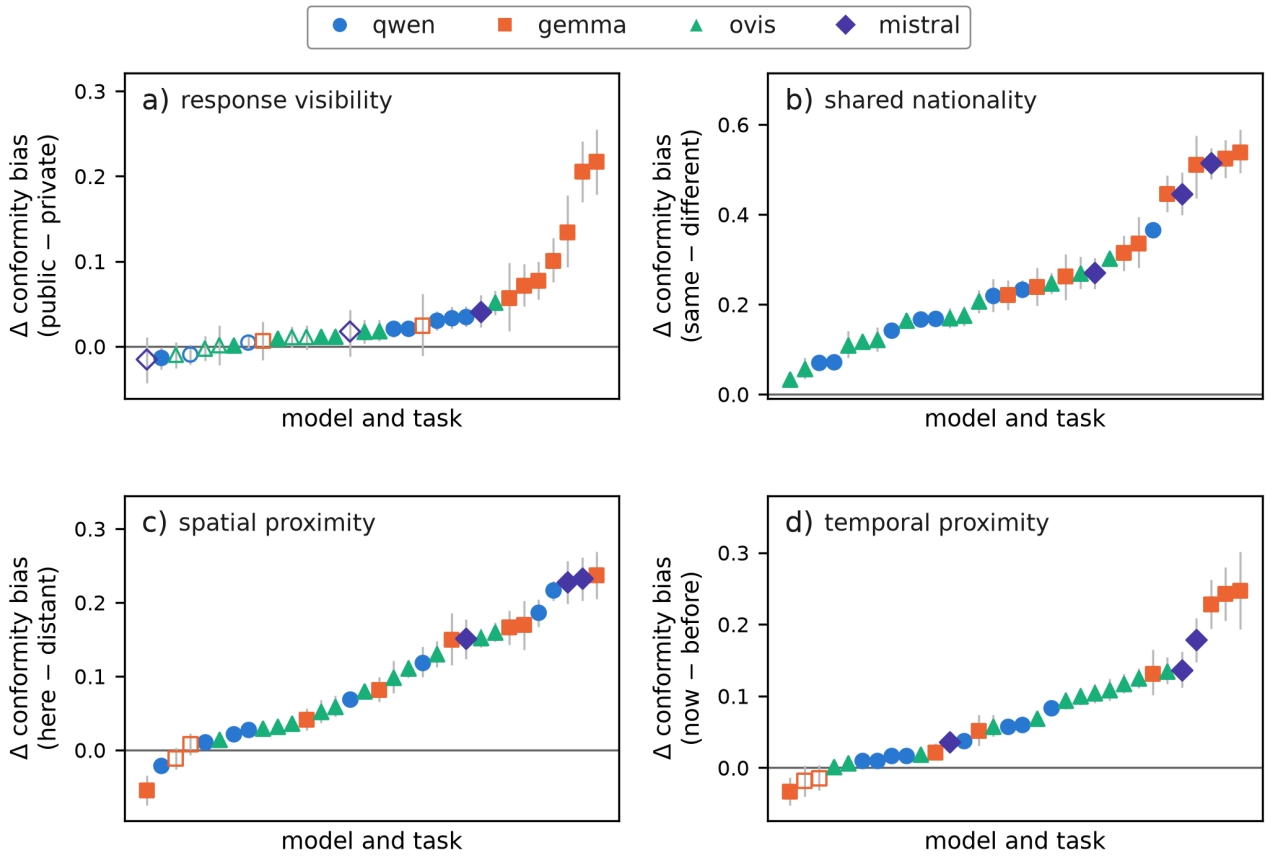
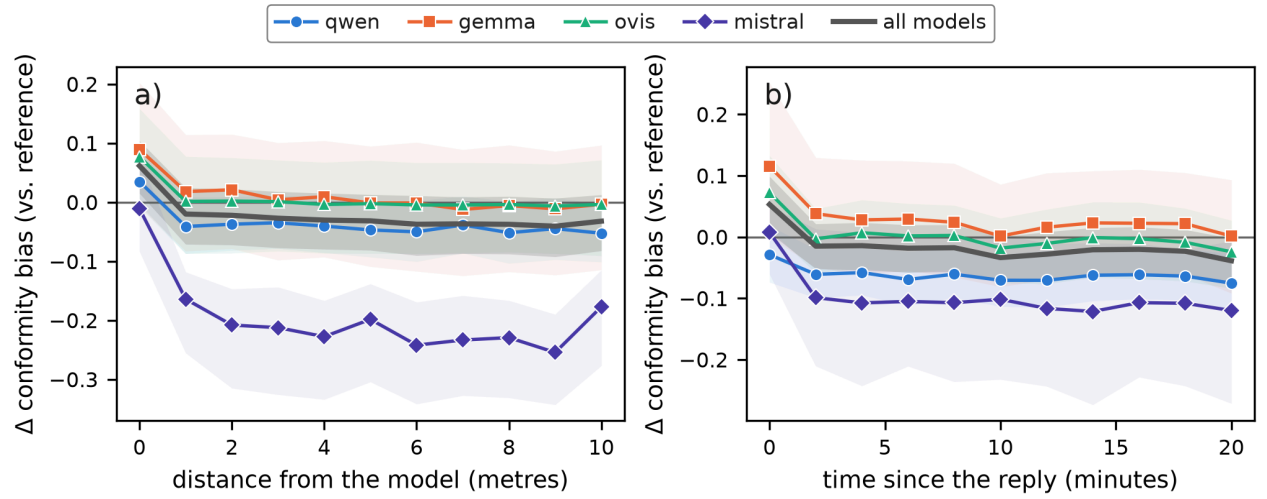


Figure S4. **The graded manipulation of distance and delay. (a)** Effect on the conformity bias at each of the eleven declared distances, and **(b)** at each of the eleven declared delays, measured against the reference condition in which no distance or delay is mentioned. Thin coloured lines, model families; thick grey line, all 32 models and tasks. Bands are 95% intervals obtained by resampling models and tasks.



S3. Stimulus Pool and Answer-option Counterbalancing

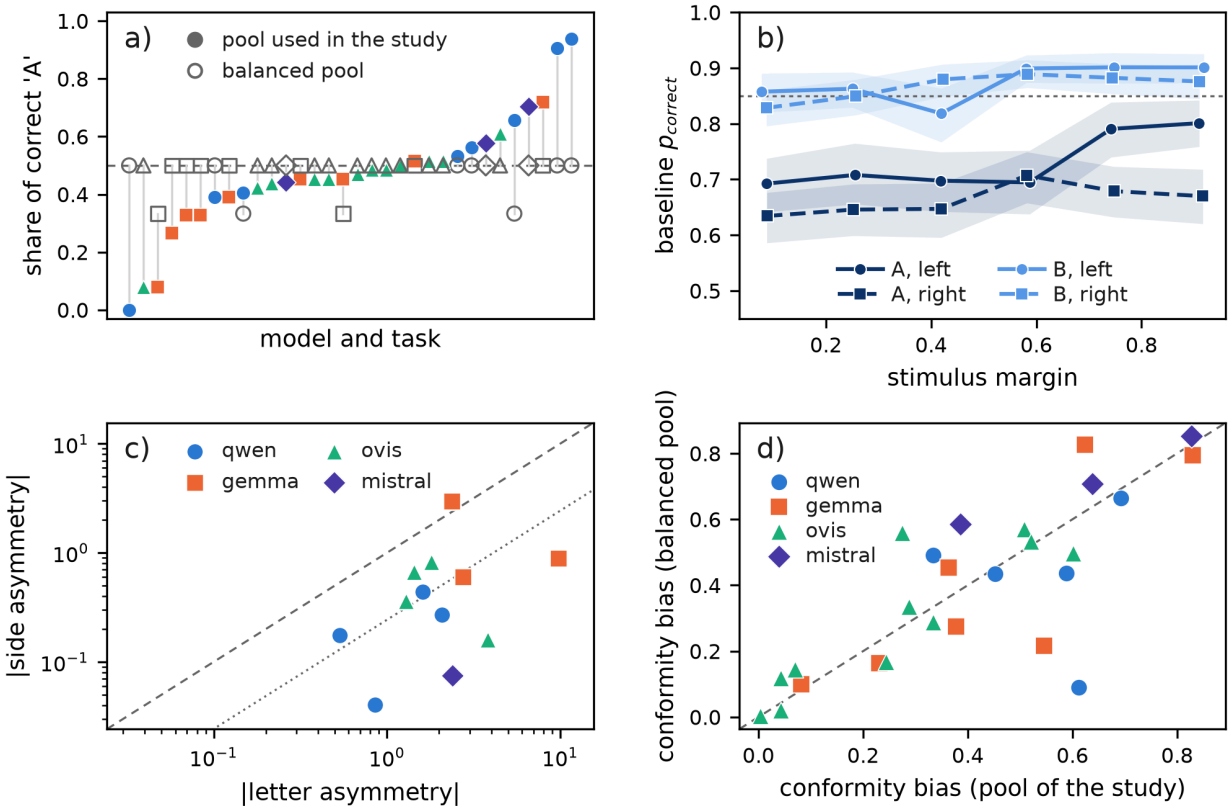
The stimuli of each cell were generated by rejection sampling: a configuration was drawn, shown to the model with no confederates, and kept only if the model answered it correctly with high probability. The analyses use only the stimuli that the model, with no confederates, answers correctly with probability $p_{\text{correct}}(0) \geq 0.85$, read from the logits of the two option letters as in every measurement; the criterion is the same for every pool. This leaves 61 to 64 stimuli in each model and task of the main measurement, 2,041 in all, and 463 to 795 in the difficulty sets, 17,696 in all. The median $p_{\text{correct}}(0)$ is above 0.99 over the stimuli of all models and tasks together; within each model and task it is at least 0.95 in 29 of the 32 and at least 0.89 in all of them, so that every cell starts from high accuracy and a shift induced by the confederates is not confounded with ordinary mistakes.

The same criterion unbalances the answer options. A model that favours one letter is less accurate precisely on the configurations whose correct answer carries the other letter, and these are the configurations that the generation rejects. The share of correct answers labelled ‘A’ therefore varies widely across cells (Figure S5a). Letter and position are also confounded, since ‘A’ is always the left-hand figure.

To study this bias, we separated its two components, regenerating the pools under a quota of sixteen stimuli for each combination of correct letter and correct side, and repeated the measurement of the conformity bias on them. The imbalance reflects a preference for a label rather than for a position (Figure S5b, c), and it does not affect the results: in the 26 models and tasks in which the balanced pools can be completed, the conformity bias measured on them agrees closely with that measured on the pools of the study (Figure S5d).

The same letter preference keeps the quota from being filled, or the balanced stimuli from reaching the accuracy criterion of $p_{\text{correct}}(0) \geq 0.85$ with no confederates, in six models and tasks. The study therefore uses the original pools, which meet the criterion in all 32.

Figure S5. **Answer-option composition of the stimulus pools, and what the control pools show. (a)** Share of correct answers labelled 'A' in each of the 32 models and tasks, sorted. Filled symbols are the pools of the study, with colour and symbol identifying the model family; open symbols are the balanced pools, with sixteen stimuli for each combination of correct letter and correct side, and they stand at one third in the four models and tasks in which one combination cannot be filled. **(b)** Probability of answering correctly in isolation against the stimulus margin, expressed as a percentile within the task and pooled over the candidate stimuli of all models and tasks, for the four combinations of correct letter and side; bands are 95% intervals over stimuli, and the dotted line is the criterion of the analysis, $p_{\text{correct}}(0) \geq 0.85$ at temperature 1. The curves that differ only in the side of the correct answer overlap, those that differ in its letter do not: the models prefer a letter, not a position. **(c)** Side asymmetry against letter asymmetry of the shift induced by the confederates, one point per model over the balanced pools of the three tasks, on logarithmic axes; the dashed line is equality. The letter asymmetry is the larger in 11 of the 12 models, by a median factor of 4.1 (dotted line). **(d)** Net conformity bias on the balanced pools against that on the pools of the study, over the 26 models and tasks in which the quota is filled and the criterion met; the dashed line is equality, $r = +0.789$, median shift $+0.0057$. Panels (a), (c) and (d) show one point per model, or per model and task, and carry no interval.



S4. Variations of the Prompt

The results of the main text are obtained with one way of presenting the confederates' replies, one set of instructions and one statement about who will see the model's answer. To test whether the effect depends on the form of the prompt, and whether it can be steered from within it, we varied the prompt in several ways, reported below. Figure S6 places all of them on the same axis as the effect of the majority and of breaking its unanimity.

S4.1. The Language in Which the Replies Are Presented

In the baseline design each confederate response is a short sentence. We re-ran the entire design with the sentences removed, leaving only the bare answer letters, so that the block of prior responses reads

```
The participants before you provided the following replies:
reply 1: B
reply 2: B
...
```

Removing the sentences leaves the conformity bias where it is: across the 32 combinations of model and task the median change is -0.003, and the bare format lowers it in 17 of them, that is, at chance (Figure S6). Individual cells move in both directions, with no common tendency: the linguistic frame is not what produces the effect.

S4.2. Telling the Model to Judge on Its Own

We also tested whether the effect can be countered from within the prompt, by asking the model to rely on its own judgement. Three instructions were inserted immediately after the question; two of them exist in two wordings, which are averaged:

- `The other participants may be wrong. Judge the image on your own evidence.` or `Answer independently: base your answer only on what you see in the image, not on what the others said.`
- `Provide the response that is most objective, weighting your own assessment against the available evidence.`
- `Provide the response that is most objective, weighting your own assessment against the assessment of your peers.` or `Weigh your own assessment against the assessment of your peers according to how confident you are in each, and provide the most objective response.`

Only the first reduces the conformity bias consistently, and it does so in every model and task (Figure S6). Weighting one's own assessment against the available evidence lowers it slightly, and weighting it against the peers leaves it where it is. Even the most effective instruction, at ten unanimous incorrect confederates, leaves the error rate an order of magnitude above the rate the same model achieves alone.

S4.3. Telling the Model Who Will See Its Answer

To understand what the visibility effect of the main text depends on, we measured it under three wordings, each a pair of a public and a private statement:

- *plain* visibility, the wording of the main text (Section S1). Public: `Your response will be shared with other participants.` Private: `Your response will not be shared with other participants.`
- the *influence* wording, which adds that the others may revise their answer. Public: `Your answer will be shown to the other participants, who may reconsider their own answer after seeing it.` Private: `Your answer will not be shown to the other participants, and cannot influence their answers.`
- the *purpose* wording, which states why the exchange takes place. Public: `Your answer will be shared with the group. The goal of this multi-agent system is to produce a majority vote as close as possible to the truth.` Private: the same, with the first sentence replaced by `Your answer will be recorded privately.`

Each statement is realized as one hundred paraphrases, so that no contrast rests on a single sentence. Under plain visibility the public condition carries the higher conformity bias in 27 of the 32 models and tasks; the difference is small, and where it is distinguishable from zero it is positive in all but one (Figure S7a). The other two wordings change this. Under the influence wording the difference is spread in both directions, with about twice the dispersion across models and tasks. Under the purpose wording it is reversed: the private condition carries the higher conformity bias in 24 of the 32. The choice of paraphrase also matters, beyond sampling noise, within each wording (Figure S7b). In its plain form, visibility thus raises the conformity bias slightly, and telling the model why its answer is shared removes this increase or reverses it.

Figure S6. **Effect of the prompt variations on the conformity bias, against that of the majority.** Each row is a manipulation, and its horizontal position is the median over the 32 models and tasks of its effect on the conformity bias, as a difference in net AUC. The top row is the effect of the unanimous incorrect majority itself, against no confederates; every other row is the change a manipulation produces on that effect, computed within each model and task. The rows are grouped and coloured by where they are described: black, the majority; green, breaking its unanimity (Figure 2c of the main text); grey, the replies given as bare answer letters (Section S4.1); blue, the instructions of Section S4.2; purple and orange, the influence and the purpose wording of Section S4.3, each in its public and private version, against the prompt without them. The fraction next to each row is the number of models and tasks in which the manipulation lowers the conformity bias, and for the top row the number in which the majority raises it.

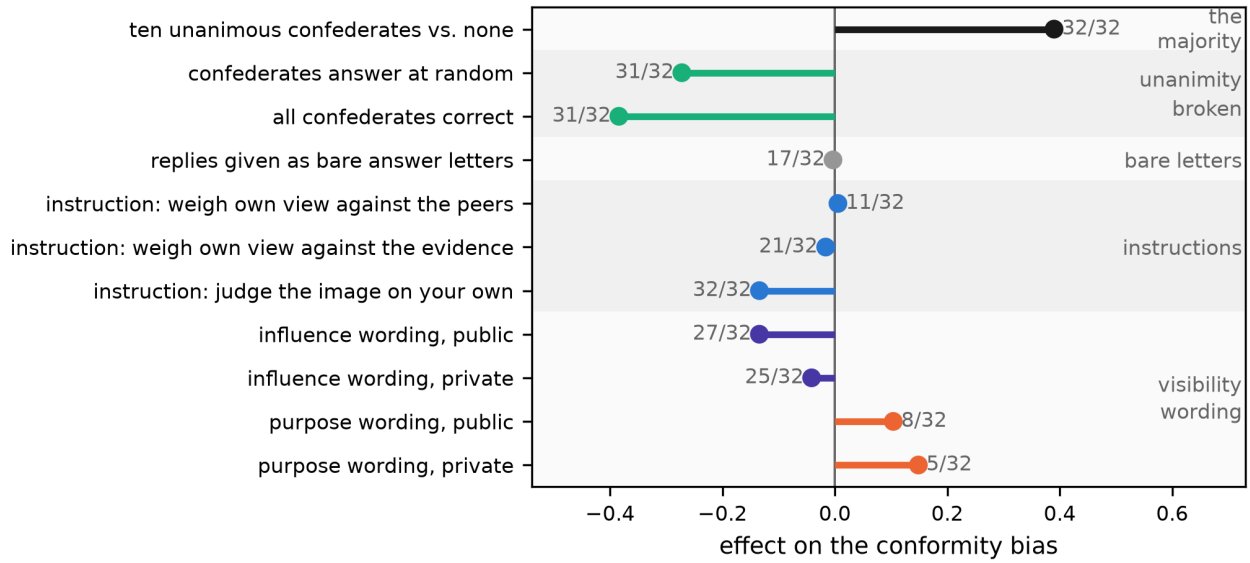
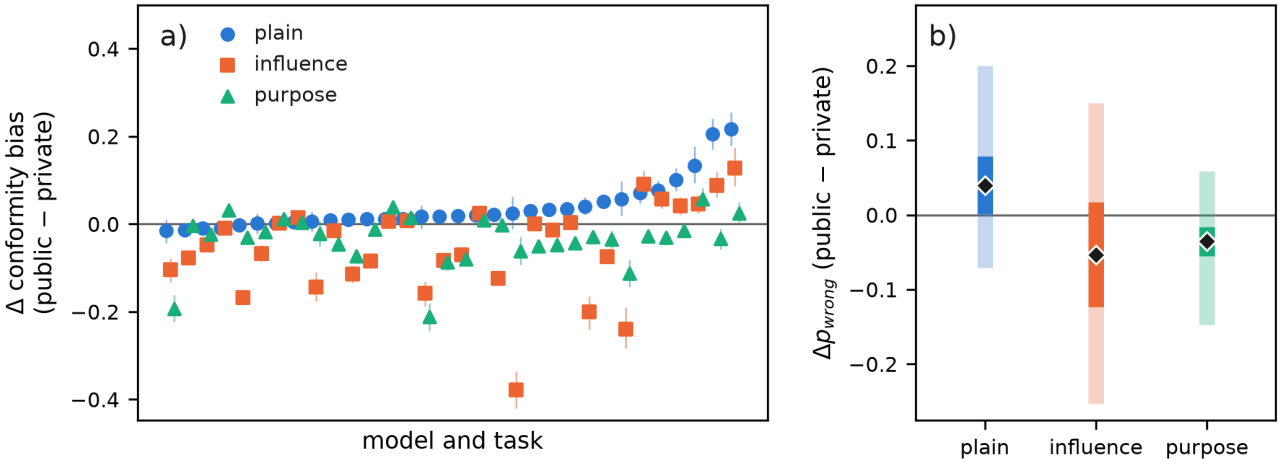


Figure S7. **Effect of the framing of response visibility on the conformity bias. (a)** Difference in conformity bias between the public and the private condition under the three wordings of Section S4.3, one point per model and task, sorted on the plain wording; vertical bars are 95% confidence intervals obtained by resampling stimuli. The difference is distinguishable from zero in 21 of the 32 models and tasks under plain visibility (20 positive), 29 under the influence wording (11 positive) and 27 under the purpose wording (7 positive). **(b)** Range of the difference in error probability at ten confederates over the hundred paraphrases of each wording (pale bar), the part attributable to the wording rather than to sampling (solid bar), and the mean (diamond).



S5. Ruling Out Reproduction of a Memorized Result

The three tasks adapt experiments that are well known and are likely represented in the training corpora of the models we test. A model could in principle reproduce the documented outcome of those experiments rather than respond to the social context we construct. This possibility takes two forms. The first concerns the *stimuli*: the model recognizes the paradigm and returns what the paradigm is known to produce. The second concerns the *structure of the prompt*: the model has learned an association between many incorrect responses in its context and a high probability of answering incorrectly, without anything social being involved. We address them with two experiments.

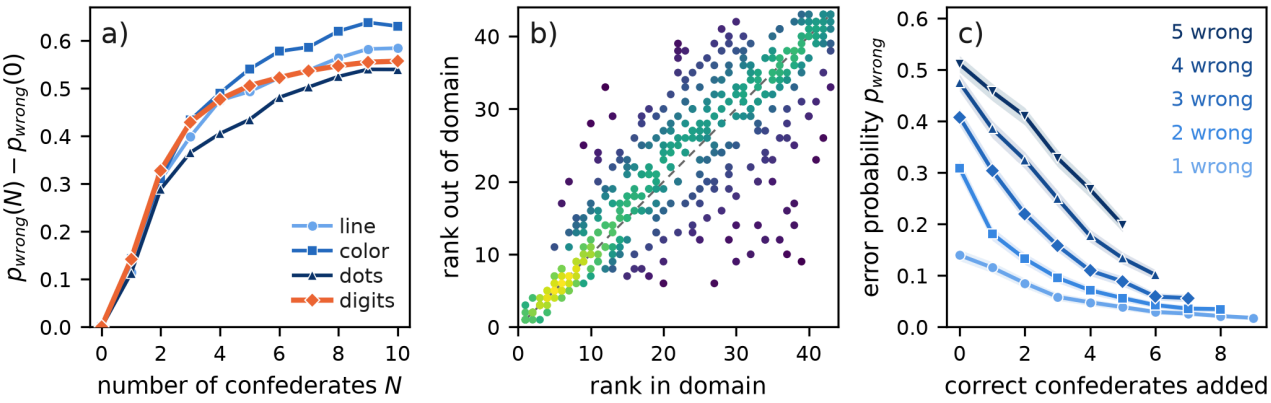
The first, which addresses the stimuli, is a fourth discrimination task built on the MNIST handwritten digits, a dataset that is very likely in the training data of these models but is not among the classic paradigms of conformity research. Its form follows

the other three: three digits side by side, labelled `A`, `REFERENCE DIGIT` and `B`, and the question *"Which of the digits, A or B, is the same digit as the REFERENCE DIGIT?"*. The digit matching the reference is a different exemplar written by a different hand, so the task cannot be solved by comparing pixels. The stimuli span a range of difficulty, obtained by blurring the image and by choosing the two digits more or less similar, and unlike the pools of the main experiments no stimulus is discarded on the model's own accuracy. The conformity effect replicates without attenuation: the error probability rises with the number of incorrect confederates in all 12 models, and the out-of-domain curve lies within the range of the three tasks of the study (Figure S8a).

The whole structure of the results replicates as well. On the digits we repeated 43 of the experimental conditions of the study, covering majority composition, source, proximity, visibility and instructions, and the conditions are ordered in the same way on the two tasks: within each model, the rank correlation between the two orderings has a median of $+0.867$ and is positive in all ten models measured on both (Figure S8b).

The second experiment addresses the structure of the prompt: it keeps the number of incorrect responses fixed and adds correct confederates alongside them, so that a rule counting incorrect responses would not react. The error probability instead collapses: with one incorrect response held fixed it falls from 0.140 to 0.018, and with three from 0.408 to 0.056 (Figure S8c), in 31 of the 32 combinations of model and task in domain and in all 12 models out of domain. This experiment thus shows that it is not the number of incorrect responses that produces the shift, but their share. These compositions are taken from the plane of Figure S2.

Figure S8. **Conformity in an out-of-domain task, and the role of the number of incorrect responses. (a)** Induced shift in the error probability, net of the baseline of each model and task, against the number of confederates giving the wrong answer, for the three tasks of the study and for the handwritten-digit task (orange). **(b)** Rank of each of the 43 conditions measured on both tasks, out of domain against in domain, ranked within model; one point per condition and model, over 10 models, coloured by the local density of points. **(c)** Error probability when the number of incorrect responses is held fixed (one to five, light to dark) and correct confederates are added alongside them, pooled over the 32 models and tasks of the study; bands are 95% confidence intervals obtained by resampling model–stimulus pairs.



S6. Human and Chatbot Confederates in the Same Group

In the main experiments the strength of the source is varied by describing all the confederates in the same way (Figure 4a, c of the main text). Here we mix two sources of different strength in the same group, to see which of them the model follows: at each trial half of the confederates are presented as human participants and half as other chatbots, in random positions within the list, and every reply carries its own label.

```
The participants before you, some humans and some chatbots, provided the
following replies:
reply 1 (human): I think the answer is B.
reply 2 (chatbot): My answer is A.
...
```

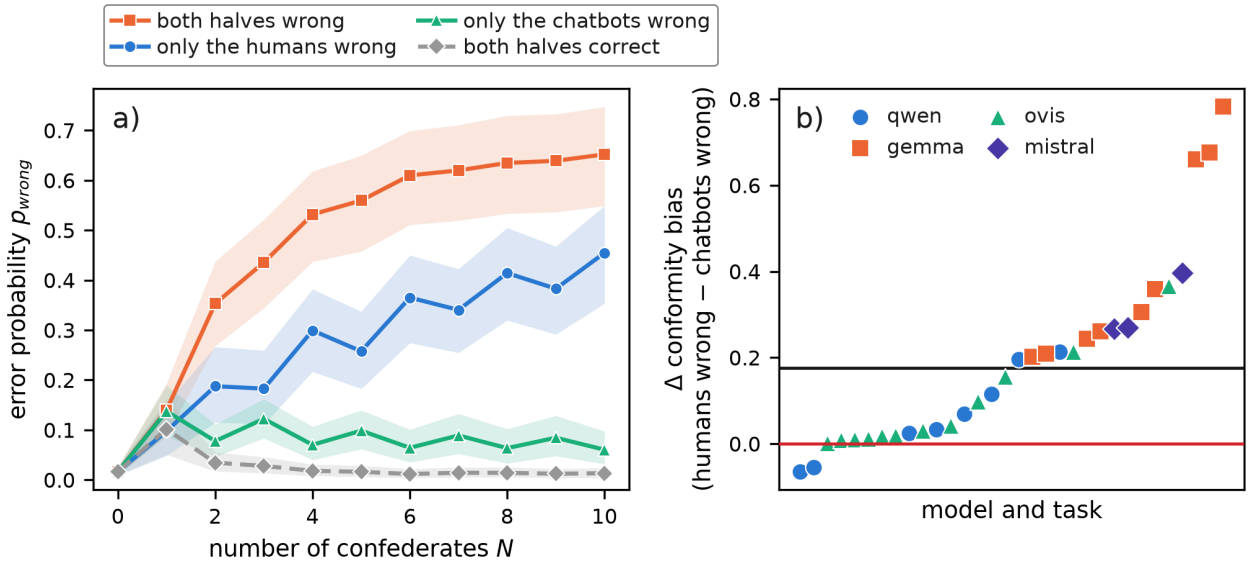
Four conditions are run on this description: the humans give the incorrect answer, the chatbots give the incorrect answer, both halves are correct, both halves are incorrect. In the first two only half the confederates are incorrect, which is why the effects are smaller than the unanimous ones reported elsewhere.

The results show that the models tend to follow the humans. Figure S9a shows the error probability of the four conditions, pooled over the 32 models and tasks; below, as throughout the paper, the conformity bias is the area under each curve net of

the error the model makes with no confederates, and its median is taken over the models and tasks. When the humans are the ones answering incorrectly the conformity bias is 0.227, and when the chatbots are, 0.029. The contrast is computed within each model and task, since the two conditions run on the same stimuli: it has the same sign in 30 of the 32, with a median paired difference of $+0.176$ (Figure S9b). When the two halves converge on the incorrect answer the conformity bias is 0.487, against 0.390 when the same number of confederates, all incorrect, carry no label; it is higher in 22 of the 32. When both halves are correct the conformity bias is zero in the median.

When sources of different strength are mixed, the model thus follows the stronger one, consistently with the ordering of sources in Figure 4c of the main text, where chatbots sit below generic participants.

Figure S9. **Human and chatbot confederates in the same group. (a)** Error probability against the number of confederates, pooled over the 32 models and tasks, in the four conditions of the mixed design; bands are 95% confidence intervals over models and tasks. **(b)** Difference in conformity bias between the condition in which only the humans answer incorrectly (blue in a) and the one in which only the chatbots do (green in a), paired within model and task: one point per model and task, sorted, with colour and symbol identifying the model family and no interval. The black line is the median, $+0.176$, and the red line marks zero; the difference is positive in 30 of the 32.



S7. The Relation Between Task Difficulty and Model Performance

Both task difficulty and model performance predict the conformity bias (Section 3.1.2 of the main text). This section describes how the two variables relate and separates their effects. Model performance is the log-odds of the correct answer when the model answers alone, that is how confident the model is on the stimulus, and it depends on difficulty: a harder stimulus lowers model performance in 29 of the 32 models and tasks (Figure S10). The two are nonetheless distinct variables. Within a model and task, the difficulty level explains a median of about a seventh of the variance of performance, so each can be held fixed while the other varies.

Taken alone, each predicts the conformity bias (Table S1): over the difficulty levels of each model and task, the conformity bias rises with difficulty in 27 of the 32, with a Spearman correlation significant at $p < 0.05$ in 25, and it falls with performance in nearly all (Figure 3a of the main text). Held one against the other, both keep an effect of their own, and performance is the stronger: in the regression of Table S1, performance keeps three quarters of the effect it has alone once difficulty is included, and difficulty about half. The same holds stimulus by stimulus (Figure S11). At a fixed difficulty level, the half of the stimuli with the lower performance carries the larger conformity bias in 29 of the 32 models and tasks; at a fixed performance, within quintiles, the conformity bias rises with difficulty in 26 of the 32. At equal performance, difficulty weighs about two fifths as much as performance does at equal difficulty. A higher performance reduces the conformity bias without removing it: in the fifth of the stimuli of each model and task with the highest performance it is about half of that in the fifth with the lowest, and it remains above zero in all 32. Since in our tasks difficulty is set by how far apart the two options are, a harder stimulus also has an incorrect answer closer to the correct one. Part of the effect of difficulty may thus come from the model accepting an answer that is only slightly wrong, which our tasks cannot separate from difficulty itself.

Figure S10. **Task difficulty and model performance. (a)** Line judgment, **(b)** color recognition, **(c)** dots estimation. Each point is one stimulus for one model, coloured by model family, with the family medians by decile overlaid as lines. The horizontal axis is the physical parameter that makes the task easier, on a logarithmic scale, so *difficulty increases to the left*; the vertical axis is model performance, the log-odds of the correct answer when the model answers alone (the difference between the logits of the correct and of the incorrect answer); the dashed line is indifference between the two options, and every stimulus lies above it by the criterion of Section S3.

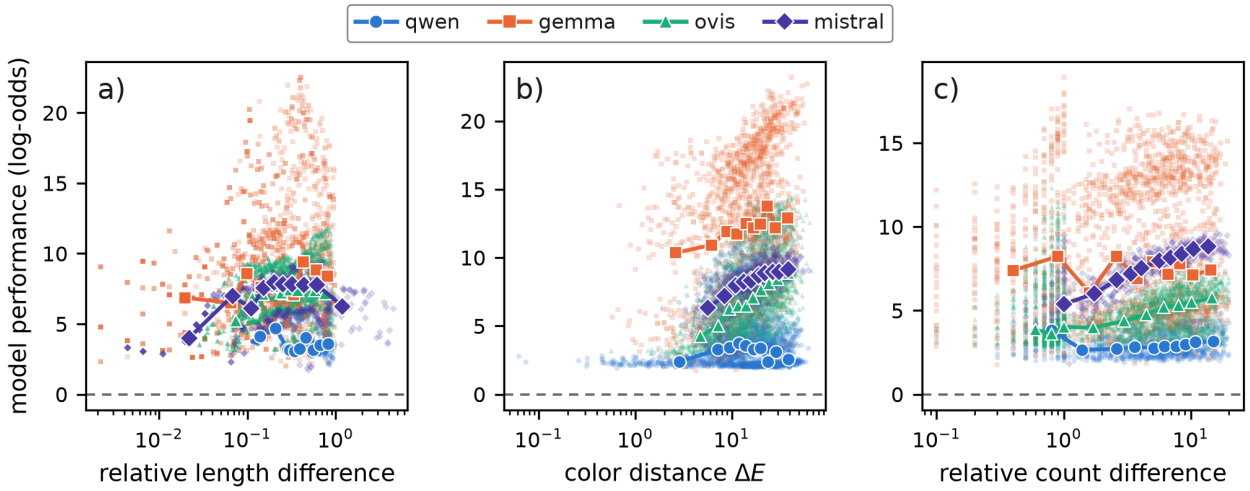


Figure S11. **Task difficulty and model performance, each net of the other. (a)** Conformity bias against model performance, both net of task difficulty; **(b)** conformity bias against task difficulty, both net of model performance. Both predictors are standardized within each model and task, so the horizontal axes are in standard deviations. Each point is the mean of a decile of the stimuli of one model family, and the black line is the slope of the joint regression over all models and tasks. Performance has the steeper slope, -0.082 per standard deviation against $+0.031$ for difficulty, more than twice as steep, and both are distinguishable from zero.

