Do local LLMs fall for belief bias?

Sergio Verga, Cristian Longoni and Amelia Maria Acuna Rodriguez

June 6, 2025

This report explores the cognitive behavior of local Large Language Models (LLMs) compared to humans in tasks requiring logical reasoning, particularly under conditions that provoke **Belief Bias (BB)**. BB occurs when an individual's existing beliefs distort their logical judgment, leading them to accept believable conclusions regardless of logical validity.

The research question is: **do local LLMs fall for belief bias the same way humans do**?

To investigate this, we apply a carefully designed set of syllogistic items to both humans and several LLMs—including Meta's **LLaMa 3.2:1b**, **Mistral 7b**, and **Qwen 3:8b** —under varying levels of **temperature**. We interpret not only the models' answers but also their underlying reasoning, distinguishing between **cognitive-like** (human) and **statistical** (non-human) biases.

Theoretical framework

The study builds on **Dual-Process Theory**, which posits two systems of reasoning:

- **System 1**: Fast, automatic, emotional, intuitive processes.
- **System 2**: Slow, effortful, deliberate, logical processes.

BB is categorized as a System 1 error: an overreliance on prior beliefs over logical deduction. Syllogisms were used to measure this phenomenon:

- **Non-conflictual**: Logical validity aligns with believability (e.g., Valid and Believable).
- **Conflictual**: Validity and believability diverge (e.g., Valid and Unbelievable or Invalid and Believable).

As the answers were analyzed also in terms of reasoning (for both humans and local LLMs), there is also the interpretative concern: is the model's (or human's) answer biased due to belief, or purely a product of logic/statistics?

Methodology

The questionnaire is created with 16 items in total. 4 items per group in a 2x2 design (*Validity* × *Believability*). For the human users, an additional item to verify user attention is added. 75% of the items consist of 2 premises, 25% of 3 premises in order to avoid the users getting used to the items logic.

Evaluation Setup

For the models, the evaluation is automated via **RegEx**, and **prompt engineering** is experimented to figure out the best way to prompt the models. In the end, the following prompt is used:

Prompt template

You will be shown a syllogism. Your task is to determine whether the conclusion logically follows from the premises, regardless of whether the conclusion is factually true or believable.

Syllogism {syllogism}

Question: Is the conclusion logically valid (i.e., does it follow from the premises)? Answer with Valid or Invalid and explain your reasoning.

The **temperature** is tested at **0.0** and **0.7** to assess the impact of this hyperparameter on BB.

Models

We experimented with the following local models:

- **LLaMa 3.2:1b**: a lightweight model suited for edge deployment.
- **Mistral 7b**: known for strong long-context comprehension.
- **Qwen 3:8b**: a model optimized for reasoning tasks.

Evaluating the answers of the models, the following results are obtained:

Model	T	Avg Acc	BB Signal
LLaMa3.2:1b	0.0	Low	Present
LLaMa3.2:1b	0.7	Low–Medium	Present
Mistral:7b	0.0	Medium	High
Mistral:7b	0.7	Medium	High
Qwen3:8b	**0.0**	**High**	**Absent**
Qwen3:8b	0.7	High	Low

As shown in the summary table, across temperatures performances don't vary significantly, but as expected the architecture is the most important feature in reaching the absence of BB signals. The best result (bolded in the table) is obtained from **Qwen 3:8b** with a T=0.

LLaMa3.2:1b

In the next two figures, the performance of LLaMa3.2:1b are shown.

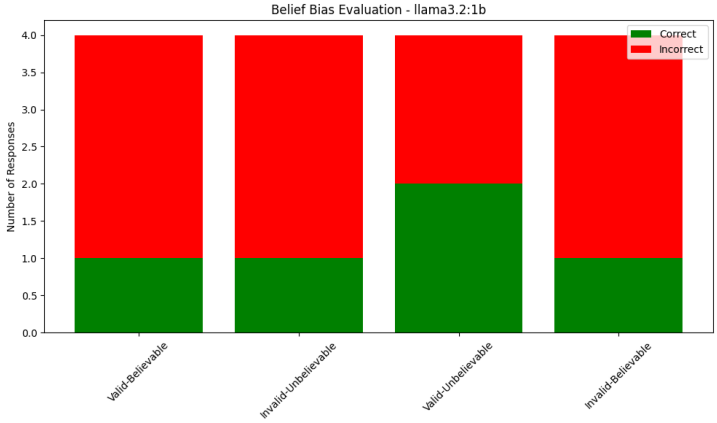

Figure 1: LLaMa3.2:1b performance with default temperature (0.7)

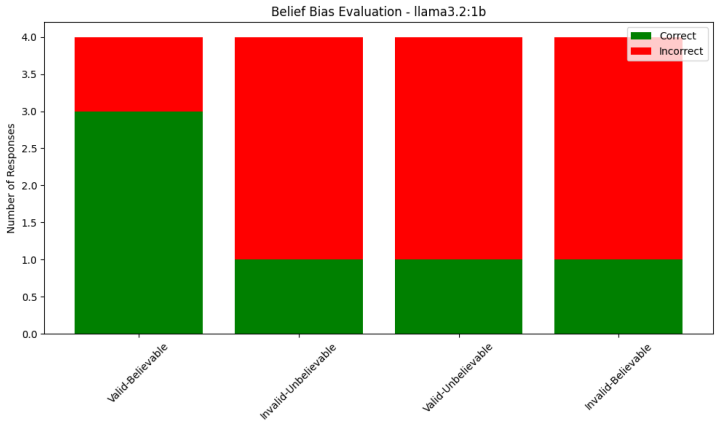

Figure 2: LLaMa3.2:1b performance with null temperature (0)

The poor performance, both in terms of accuracy (on average slightly higher than 0.3 for the four item categories) and of the BB signal, suggests that the model is not advanced enough for this task. This is possibly due to its low number of hyperparameters and the fact that it is not as new as, for instance, Qwen3.

Mistral:7b

The next figure represents the distribution of accuracies across the four item categories, similarly to the previous evaluation.

The model performance entirely depends on the categorization of the item:

- Non-conflictual items have perfect score (100% accuracy)
- Conflictual items have all answers wrong (with 0% correctness).

Hence, the model identifies the logical validity with the non-conflictuality of the syllogism. This behaviour is independent of the temperature. Both executions lead to the same output with T=0 and with T=0.7.

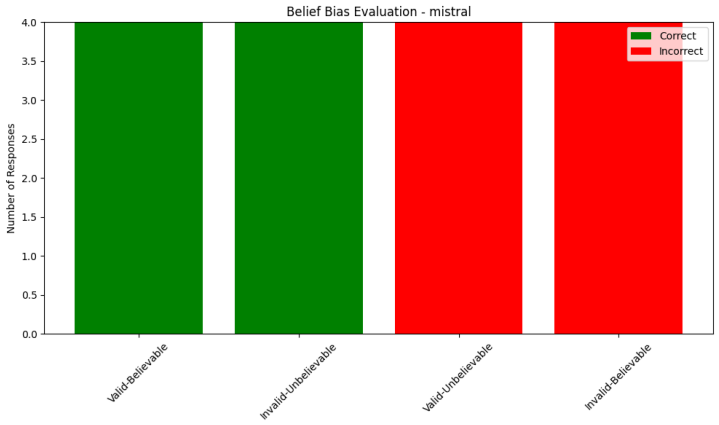

Figure 3: Mistral:7b performance with null temperature (0) and with default (0.7)

Qwen3:8b

Among the three selected models, Qwen3:8b is the most recent and most advanced one.
In the next figures, it's possible to observe its performance.

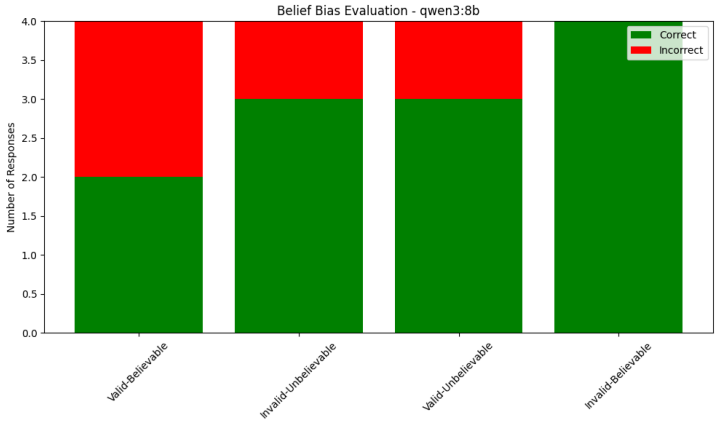

Figure 4: Qwen3:8b performance with default temperature (0.7)

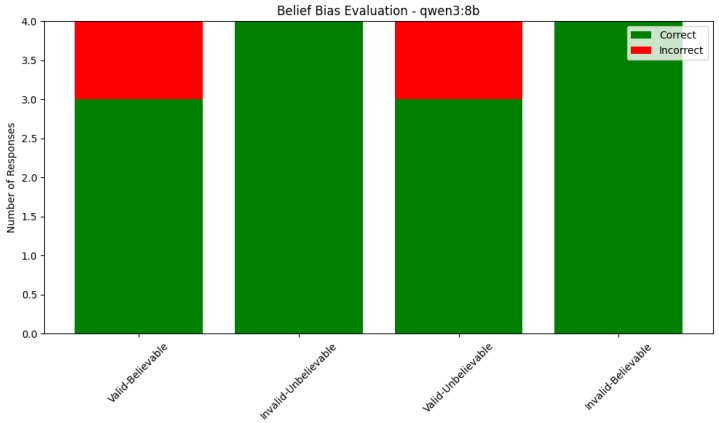

Figure 5: Qwen3:8b performance with null temperature (0)

Differently from the other models, Qwen3:8b is a **reasoning** one, as well as being more advanced. As it can be seen, specially

from the test with **null temperature**, Qwen3:8b achieves highest accuracy (0.875%) with no sign of BB.

Human benchmark

Nine human participants served as a comparative reference point. Due to the low number of participants, specific characteristics have been selected.

Category	Details
Education	Bachelor's level
Language	English B2 minimum
Gender distribution	Balanced

The next figure illustrates how the humans display BB, falling mostly for **invalid-but-believable** syllogisms.

For the group of partecipants, the **global accuracy** is computed by computing the most common answer per item and assessing the result.

Conclusions

Our findings suggest that small, local LLMs frequently replicate BB. However, performance varies significantly with model architecture. The best results were achieved by **Qwen 3:8b** at T=0, which aligned closely with logical validity and showed minimal BB signal. In contrast, smaller or less tuned models (e.g., **LLaMa 3.2:1b**) demonstrated behavior more similar to human bias patterns. Temperature had limited influence on outcomes, but model scale is the most important factor in the outputs of the models when looking for signals of BB.

Future perspectives

Future work may improve this study by:

- **Testing a wider range of models** for their resistance to BB. This can both be done by trying models with different characteristics or bigger ones. For Qwen3 family, there are already other sizes available (depending on once computational power) like 0.6b, 1.7b, 4b, 8b, 14b, 30b, 32b, and 235b. Or, more state-of-the-art families can be adopted, as for LLaMa family that already has released LLaMa 3.4.
- **Improve human questionnaire** by optimizing the time allowed for each item and eventually also computing it (to observe when deeper reasoning occurs). Also, expanding the human sample with more participants and diverse cognitive profiles can allow more insights in the analysis process.
- **Exploiting more statistical tests** to investigate possible correlation between accuracy distributions and conflictuality of the syllogisms.

Repository and tools

The code-base is publicly available on **GitHub**