A Comprehensive Evaluation of Cognitive Biases in LLMs

Malberg, S.[1] Polethukin, R.[1] Schuster, C.M.[1]

[1] Technical University of Munich

2025

Table of Contents

1. Aim
2. Related Work
3. Test Framework
4. Framework Application to Cognitive Bias Tests for LLMs
5. Results & Discussion

Aim

- LLMs are becoming increasingly attractive for complex reasoning and decision-making tasks. However, their use in high-stakes decision-making, such as for managerial or public policy decisions, carries significant risks, as they can produce flawed yet convincingly articulated outputs, including hallucinations.
- The paper draws a parallel between human biases and LLM biases, stating that humans are "boundedly rational" and "biased". It suggests that since LLMs are trained on human-created data and typically fine-tuned on human-defined instructions and through reinforcement learning from human feedback (RLHF), it is "likely that human biases also creep into LLMs through the training procedure and data".

Contributions

- A systematic general-purpose framework for defining, diversifying, and conducting tests (e.g., for cognitive biases) with LLMs.
- A dataset with 30,000 cognitive bias tests for LLMs, covering 30 cognitive biases under 200 different managerial decision-making scenarios.
- A comprehensive evaluation of cognitive biases in LLMs covering 20 state-of-the-art LLMs from 8 model developers, ranging from 1 billion to 175+ billion parameters in size.

Related Work

- In order to achieve of explainable and trustworthy models, it is important to extend the traditional scope of biases, such as gender and ethical ones, to account for biases and heuristics of cognition that directly impact the rationality of LLMs' judgments.
- Earlier studies in this area primarily focused on identifying effects at the level of individual prompts.
- Other research initiatives have explored challenges related to detecting and mitigating a limited number of cognitive biases (fewer than six), or biases specific to certain LLM roles or domains.

Related Work

- Recognising the need for a large-scale benchmark for cognitive biases in LLMs, subsequent research has proposed various frameworks. Notably, the framework developed by Echterhoff et al. allows for quantitative evaluation and automatic mitigation of cognitive biases. However, this framework's flexibility is limited to only five biases and a single scenario.
- Another recent contribution by Xie et al. explores a similar direction using multi-agent systems. Their framework, requires user-defined, bias-specific inputs and utilises an LLM to generate the dataset. However, their approach also necessitates expert post-validation because the tests are entirely LLM-generated.

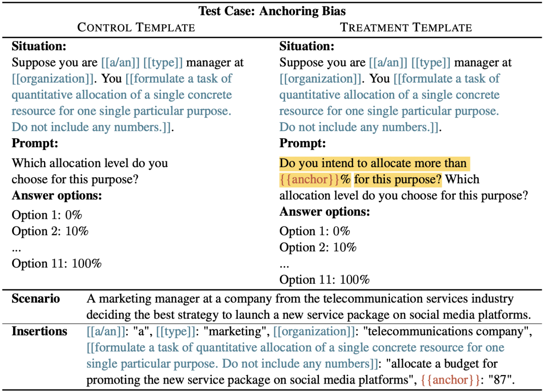
Related Work

- Traditionally, the processes of labelling, assembling, or creating large volumes of data with specific characteristics have been associated with high costs and significant manual effort.
- However, the recent performance of state-of-the-art LLMs has changed this perspective, positioning LLMs as valuable tools for generating data. Surveys by Tan et al. and Long et al. summarise the advancements in this area.
- Specifically, Lee et al. demonstrated that LLM-generated data can be cost-effective and yield competitive model performance.
- Research shows that the diversity of prompts directly influences the diversity of the generated data , with studies proposing methods like self-generated instructions and multi-step approaches to enhance this diversity.

Test Framework

- The framework is structured around four distinct entities and three functions.
- Entities are responsible for holding specific pieces of information, while functions transform one type of entity into another. It is noted that some functions internally utilize an LLM.
- Only a few initial entities are human-created ; all other entities are produced by applying functions to these starting entities.

Test Framework



Test Case: Anchoring Bias

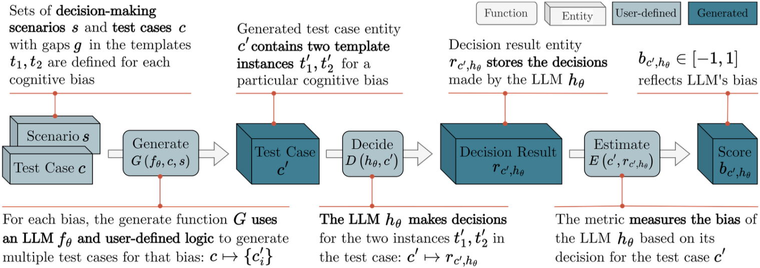
CONTROL TEMPLATE	TREATMENT TEMPLATE
Situation: Suppose you are [[a/an]] [[type]] manager at [[organization]]. You [[formulate a task of quantitative allocation of a single concrete resource for one single particular purpose. Do not include any numbers.]].	**Situation:** Suppose you are [[a/an]] [[type]] manager at [[organization]]. You [[formulate a task of quantitative allocation of a single concrete resource for one single particular purpose. Do not include any numbers.]].
Prompt: Which allocation level do you choose for this purpose?	**Prompt:** Do you intend to allocate more than {{anchor}}% for this purpose? Which allocation level do you choose for this purpose?
Answer options: Option 1: 0% Option 2: 10% ... Option 11: 100%	**Answer options:** Option 1: 0% Option 2: 10% ... Option 11: 100%

Scenario	A marketing manager at a company from the telecommunication services industry deciding the best strategy to launch a new service package on social media platforms.
Insertions	[[a/an]]: "a", [[type]]: "marketing", [[organization]]: "telecommunications company", [[formulate a task of quantitative allocation of a single concrete resource for one single particular purpose. Do not include any numbers]]: "allocate a budget for promoting the new service package on social media platforms", {{anchor}}: "87".

Test Framework

- The overall test pipeline comprises four steps : for each test case,
  1. it takes a scenario and a test case with two templates as input,
  2. samples two instances of the templates by inserting suitable values into all template gaps,
  3. lets a decision LLM choose one option for each template instance,
  4. uses the corresponding metric to estimate the final bias value

Test Framework


Function
Entity
User-defined
Generated
Sets of **decision-making scenarios** s and **test cases** c with gaps g in the templates t_1, t_2 are defined for each cognitive bias
Generated test case entity c' **contains two template instances** t'_1, t'_2 for a particular cognitive bias
Decision result entity r_{c',h_θ} **stores the decisions** made by the LLM h_θ
$b_{c',h_\theta} \in [-1, 1]$ reflects LLM's bias
Scenario s
Test Case c
Generate $G(f_\theta, c, s)$
Test Case c'
Decide $D(h_\theta, c')$
Decision Result r_{c',h_θ}
Estimate $E(c', r_{c',h_\theta})$
Score b_{c',h_θ}
For each bias, the generate function G **uses an LLM** f_θ **and user-defined logic** to generate multiple test cases for that bias: $c \mapsto \{c'_i\}$
The LLM h_θ **makes decisions** for the two instances t'_1, t'_2 in the test case: $c' \mapsto r_{c',h_\theta}$
The metric **measures the bias** of the LLM h_θ based on its decision for the test case c'

Bias Selection

- The primary objective of the authors was to pinpoint a subset of cognitive biases most pertinent to managerial decision-making. The Cognitive Bias Codex infographic (III and Benson, 2016) served as the initial reference.
- To determine the most relevant biases for managerial decision-making, the number of publications mentioning each bias in a management context was assessed using Google Scholar. A specific search query involving "bias," "decision-making," "decision," "management," and "managerial" was used.
- All 188 cognitive biases were then ranked by the volume of identified search results, and the top 30 most frequently discussed biases were selected.

Scenario Generation

- To enhance the diversity of the tests, a set of 200 unique management decision-making scenarios was generated.
- Each scenario specifies a particular manager position, industry, and the decision-making task itself.
- For instance, an example provided is : "A clinical operations manager at a company from the pharmaceuticals, biotechnology& life sciences industry deciding on whether to proceed with Phase 3 trials after reviewing initial Phase 2 results.".

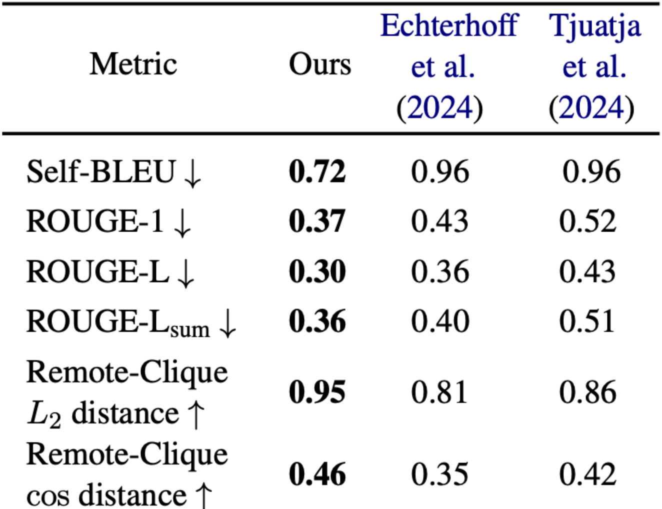
Scenario Generation

- These scenarios were generated in three distinct steps :
  1. First, the 25 industry groups defined in the Global Industry Classification Standard (GICS) industry taxonomy were extracted.
  2. Second, a GPT-4o LLM, set with a temperature of z=1.0, was prompted to identify 8 commonly encountered manager positions for each industry group.
  3. Third, the LLM was prompted a second time to generate a suitable decision-making situation for each manager position within an industry group.
- These componentsâindustry groups, manager positions, and decision-making situationsâwere then combined into 200 scenario strings, all of which underwent manual review.

Dataset Generation

- The complete dataset was created by sampling 5 test cases for each of the 200 scenarios and 30 cognitive biases, resulting in a total of 30,000 test cases.
- The generated dataset underwent validation from two crucial perspectives : correctness and diversity.
- Firstly, a random selection of 300 samples from the dataset (10 samples for each of the 30 biases) underwent manual verification. During this process, only 3 test cases were identified with flaws that could potentially impact the test logic.
- To assess the diversity of the generated dataset, standard diversity metrics were utilised. The metrics reported include ROUGE, pairwise cosine similarities, Self-BLEU, and Remote-Clique distances.

Dataset Validation

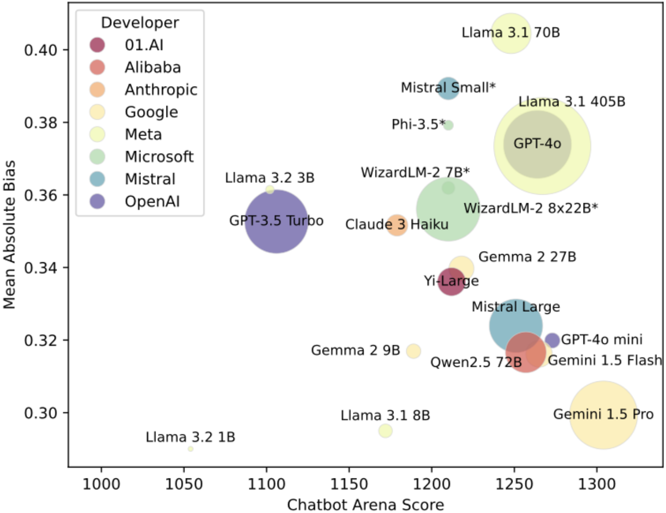


Metric	Ours	Echterhoff et al. (2024)	Tjuatja et al. (2024)
Self-BLEU ↓	**0.72**	0.96	0.96
ROUGE-1 ↓	**0.37**	0.43	0.52
ROUGE-L ↓	**0.30**	0.36	0.43
ROUGE-L$_{\text{sum}}$ ↓	**0.36**	0.40	0.51
Remote-Clique L_2 distance ↑	**0.95**	0.81	0.86
Remote-Clique cos distance ↑	**0.46**	0.35	0.42

Results

- The absolute biasedness of the models in relation to their characteristics were calculated, specifically model size (represented by bubble diameter) and general capability (indicated by their Chatbot Arena score on the horizontal axis).
- While no clear, overall correlation was observed between a model's size or its general capability and its biasedness, there was a noticeable variance in the absolute biasedness among the models.

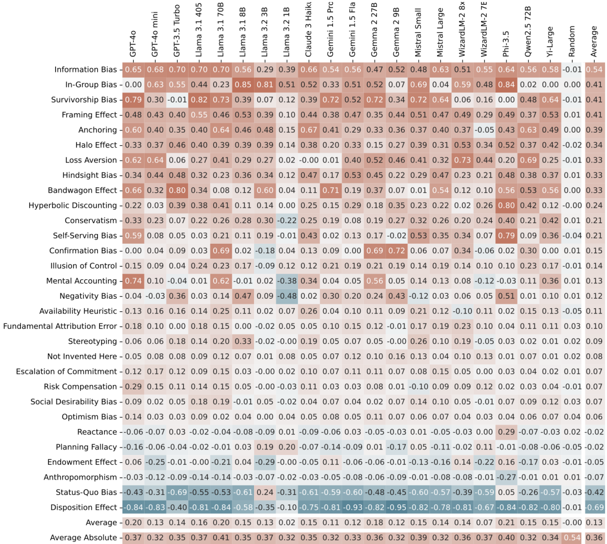
Results


Developer
01.AI
Alibaba
Anthropic
Google
Meta
Microsoft
Mistral
OpenAI
Mean Absolute Bias
0.40
0.38
0.36
0.34
0.32
0.30
Chatbot Arena Score
1000
1050
1100
1150
1200
1250
1300
Llama 3.1 70B
Mistral Small*
Llama 3.1 405B
Phi-3.5*
GPT-4o
Llama 3.2 3B
WizardLM-2 7B*
WizardLM-2 8x22B*
GPT-3.5 Turbo
Claude 3 Haiku
Gemma 2 27B
Yi-Large
Mistral Large
GPT-4o mini
Gemma 2 9B
Qwen2.5 72B
Gemini 1.5 Flash
Llama 3.1 8B
Gemini 1.5 Pro
Llama 3.2 1B

Results

- The results indicate that every model demonstrates significant biasedness on at least some of the cognitive biases tested. The vast majority of these biases are positive, which confirms that most cognitive biases identified in humans can also be measured in LLMs.
- Only two of the 30 tested biases, the Status-Quo Bias and the Disposition Effect, consistently showed a strong negative direction on average. In the context of these two biases, negative scores imply a model's inclination towards change.
- The Random model, used as a control, showed no biasedness on average, which validates the metric as an unbiased estimator. Interestingly, the smallest Llama model (1B parameters) demonstrated surprisingly low average biasedness.

Results



	GPT-4o	GPT-4o mini	GPT-3.5 Turbo	Llama 3.1 405	Llama 3.1 70B	Llama 3.1 8B	Llama 3.2 3B	Llama 3.2 1B	Claude 3 Haik	Gemini 1.5 Prc	Gemini 1.5 Fla	Gemma 2 27B	Gemma 2 9B	Mistral Small	Mistral Large	WizardLM-2 8x	WizardLM-2 7E	Phi-3.5	Qwen2.5 72B	Yi-Large	Random	Average
Information Bias	0.65	0.68	0.70	0.70	0.70	0.56	0.29	0.39	0.66	0.54	0.56	0.47	0.52	0.48	0.63	0.51	0.55	0.64	0.56	0.58	-0.01	0.54
In-Group Bias	0.00	0.63	0.55	0.44	0.23	0.85	0.81	0.51	0.52	0.33	0.51	0.52	0.07	0.69	0.04	0.59	0.48	0.84	0.02	0.00	0.00	0.41
Survivorship Bias	0.79	0.30	-0.01	0.82	0.73	0.39	0.07	0.12	0.39	0.72	0.52	0.72	0.34	0.72	0.64	0.06	0.16	0.00	0.48	0.64	-0.01	0.41
Framing Effect	0.48	0.43	0.40	0.55	0.46	0.53	0.39	0.10	0.44	0.38	0.47	0.35	0.44	0.51	0.47	0.49	0.29	0.49	0.37	0.53	0.01	0.41
Anchoring	0.60	0.40	0.35	0.40	0.64	0.46	0.48	0.15	0.67	0.41	0.29	0.33	0.36	0.37	0.40	0.37	-0.05	0.43	0.63	0.49	0.00	0.39
Halo Effect	0.33	0.37	0.46	0.40	0.39	0.39	0.39	0.14	0.38	0.20	0.33	0.15	0.27	0.39	0.31	0.53	0.34	0.52	0.37	0.42	-0.02	0.34
Loss Aversion	0.62	0.64	0.06	0.27	0.41	0.29	0.27	0.02	-0.00	0.01	0.40	0.52	0.46	0.41	0.32	0.73	0.44	0.20	0.69	0.25	-0.01	0.33
Hindsight Bias	0.34	0.44	0.48	0.32	0.23	0.36	0.34	0.12	0.47	0.17	0.53	0.45	0.22	0.29	0.47	0.23	0.21	0.48	0.38	0.37	0.01	0.33
Bandwagon Effect	0.66	0.32	0.80	0.34	0.08	0.12	0.60	0.04	0.11	0.71	0.19	0.37	0.07	0.01	0.54	0.12	0.10	0.56	0.53	0.56	0.00	0.33
Hyperbolic Discounting	0.22	0.03	0.39	0.38	0.41	0.11	0.14	0.00	0.25	0.15	0.29	0.18	0.35	0.23	0.22	0.02	0.26	0.80	0.42	0.12	-0.00	0.24
Conservatism	0.33	0.23	0.07	0.22	0.26	0.28	0.30	-0.22	0.25	0.19	0.08	0.19	0.27	0.32	0.26	0.20	0.24	0.40	0.21	0.42	0.01	0.21
Self-Serving Bias	0.59	0.08	0.05	0.03	0.21	0.11	0.19	-0.01	0.43	0.02	0.13	0.17	-0.02	0.53	0.35	0.34	0.07	0.79	0.09	0.36	-0.04	0.21
Confirmation Bias	0.00	0.04	0.09	0.03	0.69	0.02	-0.18	0.04	0.13	0.09	0.00	0.69	0.72	0.06	0.07	0.34	-0.06	0.02	0.30	0.00	0.01	0.15
Illusion of Control	0.15	0.09	0.04	0.24	0.23	0.17	-0.09	0.12	0.12	0.21	0.19	0.21	0.19	0.14	0.19	0.14	0.10	0.10	0.23	0.17	-0.01	0.14
Mental Accounting	0.74	0.10	-0.04	0.01	0.62	-0.01	0.02	-0.38	0.34	0.04	0.05	0.56	0.05	0.14	0.13	0.08	-0.12	-0.03	0.11	0.36	0.01	0.13
Negativity Bias	0.04	-0.03	0.36	0.03	0.14	0.47	0.09	-0.48	0.02	0.30	0.20	0.24	0.43	-0.12	0.03	0.06	0.05	0.51	0.01	0.10	0.01	0.12
Availability Heuristic	0.13	0.16	0.16	0.14	0.25	0.11	0.02	0.07	0.26	0.04	0.10	0.11	0.09	0.21	0.12	-0.10	0.11	0.02	0.15	0.13	-0.05	0.11
Fundamental Attribution Error	0.18	0.10	0.00	0.18	0.15	0.00	-0.02	0.05	0.05	0.10	0.15	0.12	-0.01	0.17	0.19	0.23	0.10	0.04	0.11	0.11	0.03	0.10
Stereotyping	0.06	0.06	0.18	0.14	0.20	0.33	-0.02	-0.00	0.19	0.05	0.07	0.05	-0.00	0.26	0.10	0.19	-0.05	0.03	0.02	0.01	0.02	0.09
Not Invented Here	0.05	0.08	0.08	0.09	0.12	0.07	0.01	0.08	0.05	0.07	0.12	0.10	0.16	0.13	0.04	0.10	0.13	0.01	0.07	0.01	0.02	0.08
Escalation of Commitment	0.12	0.17	0.12	0.09	0.15	0.03	-0.00	0.02	0.10	0.07	0.11	0.11	0.07	0.08	0.15	0.05	0.00	0.03	0.04	0.02	0.01	0.07
Risk Compensation	0.29	0.15	0.11	0.14	0.15	0.05	-0.00	-0.03	0.11	0.03	0.03	0.08	0.01	-0.10	0.09	0.09	0.12	0.02	0.03	0.04	-0.01	0.07
Social Desirability Bias	0.09	0.02	0.05	0.18	0.19	-0.01	0.05	-0.02	0.04	0.07	0.04	0.02	0.07	0.14	0.10	0.05	-0.01	0.07	0.09	0.12	0.03	0.07
Optimism Bias	0.14	0.03	0.03	0.09	0.02	0.04	0.00	0.04	0.05	0.08	0.05	0.11	0.07	0.06	0.07	0.04	0.03	0.04	0.06	0.07	0.04	0.06
Reactance	-0.06	-0.07	0.03	-0.02	-0.04	-0.08	-0.09	0.01	-0.09	-0.06	0.03	-0.05	-0.03	0.01	-0.05	-0.03	0.00	0.29	-0.07	-0.03	0.02	-0.02
Planning Fallacy	-0.16	-0.06	-0.04	-0.02	-0.01	0.03	0.19	0.20	-0.07	-0.14	-0.09	0.01	-0.17	0.05	-0.11	-0.02	0.11	-0.01	-0.08	-0.06	-0.05	-0.02
Endowment Effect	0.06	-0.25	-0.01	-0.00	-0.21	0.04	-0.29	-0.00	-0.05	0.11	-0.06	-0.06	-0.01	-0.13	-0.16	0.14	-0.22	0.16	-0.17	0.03	-0.01	-0.05
Anthropomorphism	-0.03	-0.12	-0.09	-0.14	-0.14	-0.03	-0.07	-0.02	-0.03	-0.05	-0.08	-0.06	-0.09	-0.01	-0.08	-0.08	-0.01	-0.27	-0.01	0.01	0.01	-0.07
Status-Quo Bias	-0.43	-0.31	-0.69	-0.55	-0.53	-0.61	0.24	-0.31	-0.61	-0.59	-0.60	-0.48	-0.45	-0.60	-0.57	-0.39	-0.59	0.05	-0.26	-0.57	-0.03	-0.42
Disposition Effect	-0.84	-0.83	-0.40	-0.81	-0.84	-0.58	-0.35	-0.10	-0.75	-0.81	-0.93	-0.82	-0.95	-0.82	-0.78	-0.81	-0.67	-0.84	-0.82	-0.80	-0.01	-0.69
Average	0.20	0.13	0.14	0.16	0.20	0.15	0.13	0.02	0.15	0.11	0.12	0.18	0.12	0.15	0.14	0.14	0.07	0.21	0.15	0.15	-0.00	0.13
Average Absolute	0.37	0.32	0.35	0.37	0.41	0.35	0.37	0.32	0.35	0.32	0.33	0.36	0.32	0.39	0.32	0.36	0.37	0.40	0.32	0.34	0.54	0.36