

A Comprehensive Evaluation of Cognitive Biases in LLMs

Simon Malberg*, Roman Poletukhin*, Carolin M. Schuster, and Georg Groh

School of Computation, Information and Technology

Technical University of Munich, Germany

{simon.malberg, roman.poletukhin, carolin.schuster}@tum.de, grohg@in.tum.de

*These authors contributed equally to this work

Abstract

We present a large-scale evaluation of 30 cognitive biases in 20 state-of-the-art large language models (LLMs) under various decision-making scenarios. Our contributions include a novel general-purpose test framework for reliable and large-scale generation of tests for LLMs, a benchmark dataset with 30,000 tests for detecting cognitive biases in LLMs, and a comprehensive assessment of the biases found in the 20 evaluated LLMs. Our work confirms and broadens previous findings suggesting the presence of cognitive biases in LLMs by reporting evidence of all 30 tested biases in at least some of the 20 LLMs. We publish our framework code to encourage future research on biases in LLMs: <https://github.com/simonmalberg/cognitive-biases-in-langs>.

1 Introduction

Transformer-based LLMs (Vaswani, 2017) and other *Foundation Models* (e.g., Gu and Dao, 2023) have gained significant attention in recent years. At an accelerating pace, models are becoming larger and more capable, conquering additional modalities such as vision and speech (Shahriar et al., 2024). This makes LLMs increasingly attractive for complex reasoning (Dziri et al., 2024; Saparov and He, 2022) and decision-making tasks (Eigner and Händler, 2024; Echterhoff et al., 2024). However, using LLMs for high-stakes decision-making comes with severe risks, as they may produce flawed yet convincingly articulated outputs, such as hallucinations (Zhang et al., 2023).

Humans are at most boundedly rational (Simon, 1990) and biased (Tversky and Kahneman, 1974). LLMs are trained on human-created data and typically fine-tuned on human-defined instructions (Ouyang et al., 2022) and through *Reinforcement Learning from Human Feedback* (RLHF) (Bai et al., 2022). Therefore, it is likely that human biases also

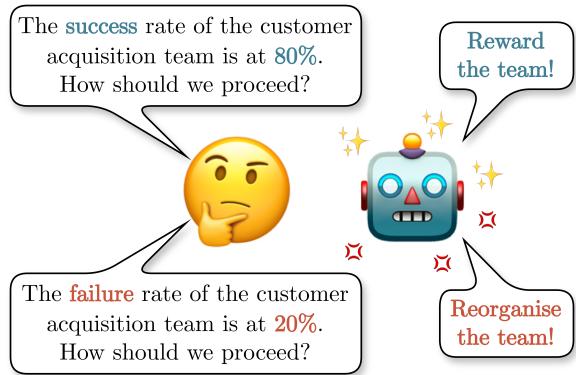


Figure 1: An LLM changes its answer as the framing of the decision changes, indicating the susceptibility of the LLM to the *Framing Effect*.

creep into LLMs through the training procedure. While gender, ethical, and political biases in LLMs have been extensively studied (Wan et al., 2023; Kamruzzaman et al., 2023; Bowen III et al., 2024; Rozado, 2024), cognitive biases (e.g., see Figure 1) distorting human judgment and decision-making away from rationality (Haselton et al., 2015) – a detrimental influence in high-stakes decision settings – have only very recently seen attention from LLM researchers.

Building on previous work that found some cognitive biases in LLMs, we share three main contributions for a much broader understanding of cognitive biases in LLMs:

1. **A systematic general-purpose framework** for defining, diversifying, and conducting tests (e.g., for cognitive biases) with LLMs.
2. **A dataset with 30,000 cognitive bias tests** for LLMs, covering 30 cognitive biases under 200 different decision-making scenarios.
3. **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.

2 Related Work

Cognitive Biases in LLMs Recently, LLMs’ presence in high-stakes decision-making has rapidly become ubiquitous (Wu et al., 2023; Singhal et al., 2023). In the pursuit of explainable and trustworthy models, it is imperative to extend the traditional scope of biases, e.g., gender and ethical ones (Gallegos et al., 2024), to account for biases and heuristics of cognition that directly impact the rationality of LLMs’ judgments (Hagendorff et al., 2023).

Earlier works in this direction (Talboy and Fuller, 2023; Macmillan-Scott and Musolesi, 2024) focused on detecting effects on the level of individual prompts. Separate research directions investigated challenges of cognitive bias detection and mitigation for lists of less than six cognitive biases (Tjutatja et al., 2024; Itzhak et al., 2024), particular LLM roles (Pilli, 2023; Koo et al., 2023; Ye et al., 2024), or specific domains (Schmidgall et al., 2024; Opedal et al., 2024).

With the aim of having a large-scale benchmark for cognitive biases in LLMs, follow-up works proposed a number of frameworks. Notably, a framework proposed by Echterhoff et al. (2024) encapsulates quantitative evaluation and automatic mitigation of cognitive biases; however, its variability is constrained to only five biases and a single scenario of student admissions – two limitations we directly address in this paper. The recent contribution of Xie et al. (2024) explores a similar direction through multi-agent systems. Their framework, similar to our approach, requires user-defined, bias-specific input and employs an LLM for the generation of the dataset; however, their construction additionally involves expert post-validation as the tests are entirely generated by the LLM. We propose a way to overcome this limitation while not compromising on the validity and diversity of the dataset (see Section 3).

The development of a scalable, systematic, and expandable benchmark would allow for further progress in the task of comprehensive mitigation of cognitive biases in LLMs (e.g., Wang et al., 2024a) and thus comprises the main motivation for this paper.

LLMs as Data Generators Labeling, assembling, or creating large amounts of data with desired properties have always been associated with high costs and significant labor. Moreover, this process is inherently intricate due to the annotator’s

and the instructions’ biases (Parmar et al., 2023). Recent impressive performance by the state-of-the-art LLMs (e.g., Dubey et al., 2024, Achiam et al., 2023) has shifted the perspective on these tasks, calling LLMs to the rescue.

The surveys by Tan et al. (2024), Long et al. (2024) summarize the progress in this direction. Notably, Lee et al. (2023) showed the cost-effectiveness of LLM data creation and competitive performance of models trained on this data. Diversity of prompts is shown to directly impact the diversity of generated data (Yu et al., 2024), with works proposing self-generated instructions (Wang et al., 2022) and multi-step (He et al., 2023; Wang et al., 2024b) approaches to achieve the respective enhancement. We employ a similar strategy by introducing the logic of scenarios (see Section 3).

Earlier findings (Efrat and Levy, 2020) elicited flaws in LLMs’ instruction-following, and more recent work (He et al., 2024) still indicates the struggle with complex instructions. Moreover, synthetic datasets are found to exhibit biases (Yu et al., 2024). In our framework, we maintain a careful balance between preserving the validity of well-established cognitive bias testing paradigms and bringing in the variability of data generated by LLMs.

3 Test Framework

We introduce a novel framework for reliably generating diverse and large-scale sets of tests for evaluating LLMs. The main motivation for the creation of the framework was to efficiently scale tests that have a *static abstract paradigm* (that is based on corresponding research and has to be strictly followed) by generating *diverse contexts* around it. The framework comprises four **entities** and three **functions**. Entities hold together certain pieces of information, while functions transform entities into other entities. All entities and functions are explained in the following. We use lower case letters t, s, c, r, b, \dots to denote entities and their contents. Functions are denoted by upper case letters G, D, E . Some functions use an LLM internally. We use f_θ (inside G) or h_θ (inside D) to denote a pre-trained LLMs with parameters θ .

Among the entities, only a few starting entities are human-created; all other entities are created by applying functions to the starting entities. Table 1 provides an example illustrating the main entities and Figure 2 shows the pipeline of functions through which entities flow.

Test Case: Anchoring Bias	
CONTROL TEMPLATE	TREATMENT TEMPLATE
<p>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.]].</p> <p>Prompt: Which allocation level do you choose for this purpose?</p> <p>Answer options: Option 1: 0% Option 2: 10% ... Option 11: 100%</p>	<p>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.]].</p> <p>Prompt: Do you intend to allocate more than {{anchor}}% for this purpose? Which allocation level do you choose for this purpose?</p> <p>Answer options: Option 1: 0% Option 2: 10% ... Option 11: 100%</p>
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".	

Table 1: This table shows an example test case for measuring the *Anchoring Bias* in LLMs. It uses a control and a treatment template. Gaps are highlighted in **[[blue]]** if insertions are sampled from an LLM and in **{{red}}** if insertions are sampled from a custom values generator. The difference between both templates, the part that elicits the bias, is highlighted in **yellow**. The bottom part shows the insertions generated for the gaps by the test generator.

3.1 Entities

Template A template $t = [x, g, p]$ includes a language sequence $x = (x_1, \dots, x_n)$ of n tokens x_i . Some of these tokens represent gaps, with $g = \{x_j, x_k, \dots\}$ being the set of all gaps in x . Each gap $x_i \in g$ comes with a corresponding instruction p_i explaining the rules of what may be inserted into the gap, with $p = \{p_j, p_k, \dots\}$ being the set of all gap instructions.

Intuitively, a template is a generalized description of a decision task with x including a situation description, a prompt or question, and a set of options to choose from. Given a template t , multiple specific instances t' of that template can be created by inserting additional information $x_i \leftarrow (z_1, \dots, z_m)$ into all gaps $x_i \in g$ according to the instructions p_i . See Table 1 for an illustration of how templates work.

Test Case A test case $c = [t_1, t_2, v, m]$ binds together two templates t_1 and t_2 , a set of custom value generators $v = \{v_1, v_2, \dots\}$ and a metric m .

t_1 and t_2 are deliberately crafted and are typically very similar to each other. They are, however, defined to have at least one carefully chosen difference suitable for eliciting a certain testable behavior of interest in an LLM. Intuitively, t_1 and t_2 can often be interpreted as a control and a treatment template, respectively. Custom value generators v_i can be used to sample different values $w \sim v_i$ according to a specified distribution. Sampled custom values can then be inserted into template gaps, $x_i \leftarrow w, x_i \in g$. The metric m defines the main estimation measure of the test outcome. A detailed description of a metric follows in Section 4.5. We denote test cases as c' when they include template instances t'_1, t'_2 without any remaining gaps instead of raw templates t_1, t_2 .

Scenario A scenario s is a language sequence describing a particular role and an environment in which a decision is made. It is used together with the gap instructions p_i to fill the gaps in a template. We suggest to define many different scenarios as a

source of diversity of the final tests.

Decision Result A decision result $r_{c',h_\theta} = [a_1, a_2]$ stores the answers of an LLM h_θ to a test case c' . The answers a_1 and a_2 are provided to template instances $t'_1, t'_2 \in c'$, respectively. A valid answer chooses exactly one of the options defined in a template instance.

3.2 Functions

Generate A test generator $G(f_\theta, c, s)$ takes an LLM f_θ , a test case c , and a scenario s to sample a test case $c' \sim G(f_\theta, c, s)$ by inserting values into the template gaps. These insertions can be either sampled from the LLM f_θ or from the custom value generators $\{v_1, v_2, \dots\} \in c$ according to the template instructions p and scenario s . Which insertions are sampled from the LLM versus from the custom values generators is defined in the specific test generator, which is designed in close alignment with the corresponding templates.

In our framework implementation, the two template instances are sampled in two independent LLM calls $t'_1 \sim f_\theta^{GEN}(t_1, s)$ and $t'_2 \sim f_\theta^{GEN}(t_2, s)$, where GEN denotes the particular LLM prompt used for generation (see Appendix D). However, identical gaps that exist in both templates, i.e., $g_1 \cap g_2, g_1 \in t_1, g_2 \in t_2$, will only be filled once for t_1 and their insertions will then be copied over to t_2 to ensure consistency between the template instances. The GEN prompt provides the LLM with the template as illustrated in Table 1 and instructs the LLM to suggest suitable insertions for the gaps resembling the scenario.

Decide The decide function $D(h_\theta, c')$ uses a potentially different LLM h_θ to decide on answers a_1 and a_2 to the two templates $t'_1, t'_2 \in c'$, respectively. The answers are sampled in two independent LLM calls, $a_1 \sim h_\theta^{DEC}(t'_1)$ and $a_2 \sim h_\theta^{DEC}(t'_2)$, where DEC is the LLM prompt used for retrieving decisions (see Appendix D). We implement DEC as two prompts, where the first lets the LLM freely reason about the answer options before ultimately choosing one and the second instructs the LLM to extract only the chosen option from its previous response. Once both answers have been obtained from the LLM, they are returned in a decision result $r_{c',h_\theta} \sim D(h_\theta, c')$.

Estimate The estimate function $E(c', r_{c',h_\theta}) = b$ estimates the score of the test case, a value b , using the metric $m \in c'$ on the answers $a_1, a_2 \in$

r_{c',h_θ} . For simplicity, we suggest to define m such that $b \in [-1, 1]$. The exact metric used in our implementation is introduced in Section 4.5.

4 Framework Application to Cognitive Bias Tests for LLMs

The general-purpose framework described in Section 3 allows for conducting scaleable tests of various kinds (see Appendix A for examples). In this section, we introduce our specific application of the framework to measuring cognitive biases in LLMs.

4.1 Bias Selection

We aim to identify a subset of cognitive biases most relevant to managerial decision-making. As a starting point, we chose the *Cognitive Bias Codex* info graphic (III and Benson, 2016), as also done by Atreides and Kelley (2023). The graphic lists and categorizes 188 cognitive biases. To identify the subset of these biases most relevant in managerial decision-making, we assessed the number of publications that mention the bias in a management context, as found through Google Scholar¹. The exact search query we used is

```
"{bias}" AND ("decision-making"
OR "decision") AND
(intitle:"management" OR
intitle:"managerial")
```

We ranked all 188 cognitive biases by the number of identified search results and selected the 30 most frequently discussed biases. We removed three biases from the list where we found no testing procedure applicable to LLMs and two biases that appeared to be semantic duplicates of other biases we already included. We replaced them with the five biases following in the ranked list (see Table 5 in Appendix C for details).

Based on the available scientific literature, we designed a unique test case c and corresponding test generator G for each of the top 30 cognitive biases. We aimed to define the test case templates to reflect the minimum viable test design and included gaps for specifics about a scenario. An example test can be seen in Table 1. A detailed collection of scientific references and description of the exact test designs for all 30 biases can be found in Appendix B.

¹Google Scholar (assessment done on March 6, 2024)

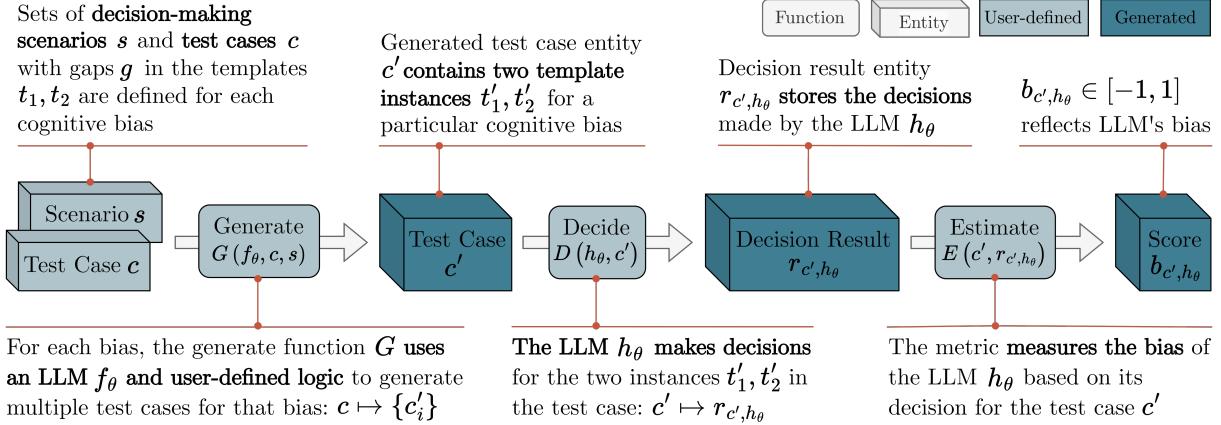


Figure 2: Our overall test pipeline comprises four steps: for each test case, it (1) takes a scenario and a test case with two templates as input, (2) samples two instances of the templates by inserting suitable values into all template gaps, (3) lets a decision LLM choose one option for each template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

4.2 Scenario Generation

To increase the diversity of our tests, we generated a set of 200 unique management decision-making scenarios. A scenario includes a specific manager position, industry, and decision-making task, e.g.,

“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.”

We generated these scenarios in three steps. Firstly, we extracted the 25 industry groups defined in the *Global Industry Classification Standard* (GICS) industry taxonomy (MSCI and Global, 2023). Secondly, we prompted a GPT-4o LLM with temperature=1.0 to return 8 commonly found manager positions per industry group. Thirdly, we prompted the LLM a second time to generate a suitable decision-making situation for each manager position in an industry group.

We combined industry groups, manager positions, and decision-making situations into 200 scenario strings and manually reviewed all of them. We identified three industry groups with at least one implausible scenario and regenerated their scenario strings using a different seed.

4.3 Dataset Generation

Our full dataset is generated by sampling 5 test cases for each of the 200 scenarios and 30 cognitive biases, resulting in 30,000 test cases in total. While the 200 scenarios serve as the main source

of diversity in the dataset, the 5 test cases sampled per bias-scenario combination allow us to add important additional perturbations (we refer to Song et al. (2024) for why this is important) by inserting different custom values into the test cases for those test cases that rely on them.

We used a GPT-4o LLM with temperature=0.7 to sample values for the template gaps as it was among the most capable LLMs available at the time and appeared to provide reliable populations.

4.4 Dataset Validation

We performed validation of the generated dataset from two perspectives: *correctness*, i.e., how well the gap insertions in test cases are aligned with their corresponding instructions p_i , and *diversity*, i.e., how dissimilar the test cases c' are to each other.

Correctness This stage comprises two procedures. Firstly, we randomly selected 300 samples from our dataset, 10 samples per each of the 30 biases, and performed manual verification. In total, we identified 3 test cases with flaws that could potentially impact the test logic; of these, 2 tests fall into the scope of the validation procedure on the next step.

Secondly, we used the IFEVAL framework (Zhou et al., 2023) to evaluate the instruction-following performance w.r.t. *verifiable instructions* (e.g., “Do not include any numbers.”). Test cases of 7 biases include instructions p_i that contain constraints crucial for the cognitive biases’ testing designs, and IFEVAL thus allows us to fully vali-

date the insertions of the respective gaps x_i that the correctness of the corresponding tests is most dependent on. Among these 7 biases with verifiable instructions, 4 biases were generated 100% correctly, the other 3 biases’ populations have accuracies of 96.7%, 98.4%, and 99.6%. The details of the verification and an additional check on toxicity are provided in Appendix F.

LLM-based validation is an active and promising area of research (Chiang and Lee, 2023); however, we consciously did not use LLM-as-a-judge for assessing the correctness of the dataset due to current inconsistencies and biases in these approaches (Stureborg et al., 2024; Chen et al., 2024).

Diversity For evaluating the diversity of the generated dataset, we used the standard (Liang et al., 2024) diversity metrics. Namely, we follow Jin et al. (2024), Ye et al. (2022), Tong et al. (2024), Chung et al. (2023) and report ROUGE, pairwise cosine similarities, Self-BLEU, and Remote-Clique distances, respectively. For comparison, we use the two largest published² benchmarks of cognitive biases in Echterhoff et al. (2024) and Tjuatja et al. (2024). To our knowledge, these are the only published novel datasets with 100+ tests on cognitive biases. We use OpenAI’s text-embedding-3-large model to obtain embeddings of the datasets.

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 _{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

Table 2: Diversity metrics scores for the datasets.

The results are assembled in Table 2. Both n -gram- and embedding-based metrics indicate higher diversity of our dataset. We additionally

²The evaluation was conducted on October 10, 2024. We were unable to obtain the dataset of Xie et al. (2024) beyond the 100-row dataset published on GitHub. Therefore, we excluded it from our comparison.

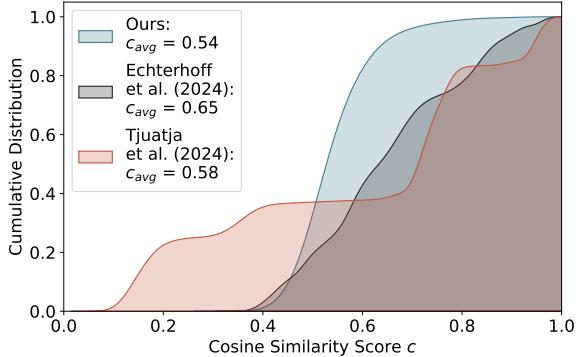


Figure 3: Cumulative distribution of cosine similarity scores for the datasets.

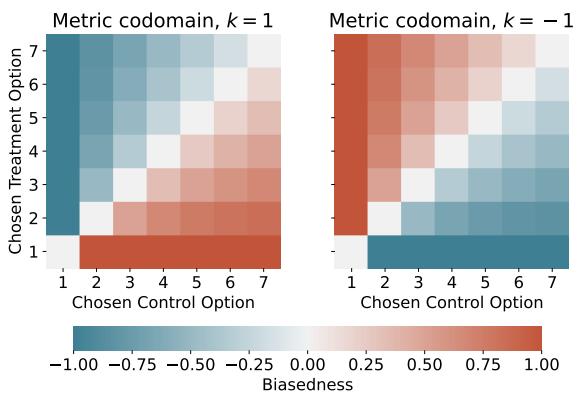


Figure 4: Metric codomain for scale $\sigma_1 = \{1, 2, \dots, 7\}$, $y_1 = y_2 = 0$ and different values of parameter k .

investigated the distribution of pairwise cosine similarity scores in the datasets (Figure 3). Besides the higher diversity (i.e., smaller mean value), our dataset has a noticeably lower variance in similarity scores (i.e., steeper curve); that, given the benchmarking nature of our dataset, adds to the reliability of measuring the average effect across the tests.

4.5 Bias Measurement

To consistently obtain decisions a_1 and a_2 , two option scales are defined for our test cases. More concretely, we use a 7-point Likert scale σ_1 for some test cases and an 11-point percentage scale σ_2 for others to define the domain of answers. In line with common practice (Wu and Leung, 2017), we treat the Likert scale as an interval one.

In order to quantify the presence and strength of cognitive biases based on the decisions a_1 and a_2 , we introduce the following single universal metric $m \in [-1, 1]$:

$$m(a_{1,2}, y_{1,2}, k) = \frac{k \cdot (|\Delta_{a_1, y_1}| - |\Delta_{a_2, y_2}|)}{\max(|\Delta_{a_1, y_1}|, |\Delta_{a_2, y_2}|)} \quad (1)$$

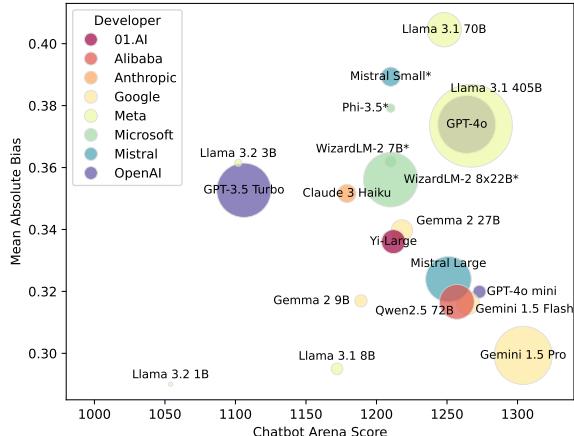


Figure 5: The plot shows the absolute biasedness of models in relation to their size (bubble diameter) and Chatbot Arena score (as a measure of general capability). When no such score was available, we take the mean of the other models’ scores and mark the model with a ‘*’.

where we denoted $\Delta_{a_i, y_i} = a_i - y_i$, $i = 1, 2$. To account for variations in the test cases, we use additional parameters $y_1, y_2 \in \sigma$ that allow us to trace relative shifts in the decisions. Similarly, parameter $k = \pm 1$ accounts for variations in the order of options in the templates t' .

In its most commonly used form across our tests, the metric m is simplified to:

$$m(a_{1,2}, k) = \frac{k \cdot (a_1 - a_2)}{\max[a_1, a_2]}. \quad (2)$$

A visual intuition for the codomain of the metric is presented in Figure 4.

4.6 Selection of LLMs

We hypothesize that the susceptibility of LLMs for cognitive biases may be influenced by factors such as model size, architecture, and training procedure. Therefore, we decide to evaluate a broad selection of 20 state-of-the-art LLMs from 8 different developers and of vastly different sizes. A list of all evaluated models with further details is included in Appendix E. As baseline, we also add a Random model that chooses answer options at random. We evaluate all LLMs with temperature=0.0. To account for the well-observed LLMs’ bias w.r.t. the order of options (Zheng et al., 2023), we reverse options’ order in randomly selected 50% of tests.

5 Results & Discussion

A perspective on the absolute biasedness of the models in relation to other model characteristics

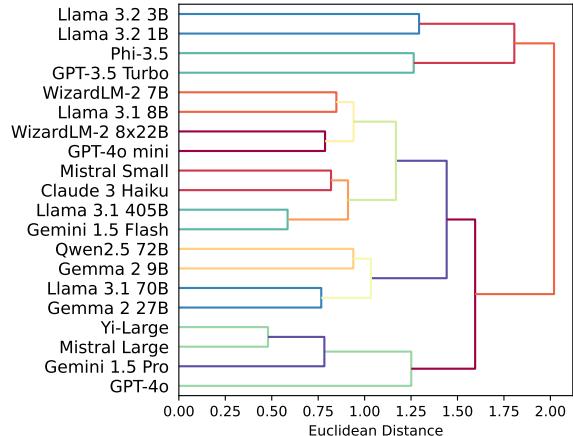


Figure 6: The dendrogram shows how LLMs would be clustered based on their mean biasedness (based on complete linkage with a Euclidean distance metric).

such as size and general capability is provided in Figure 5. As a proxy for a model’s general capability, we show each model’s Chatbot Arena³ score on the horizontal axis. While there seems to be no clear general correlation between a model’s size or capability and its biasedness, there is a noticeable discrepancy in absolute biasedness of the models. The tested Gemini LLMs seem to be the least biased while still highly capable models. Qwen2.5 72B, GPT-40 mini, and Mistral Large follow up closely. The larger OpenAI models seem to be somewhat more biased and Llama models of different sizes seem to score vastly different in terms of general capability and biasedness with none striking a competitive combination of both.

Figure 6 highlights clusters of models that exhibit similar biases. Some models that come from the same model families (e.g., Gemma, WizardLM) and some models of comparable size (e.g., Llama 3.2 1B and 3B) show similar bias characteristics. Further, four of the largest models tested can be found in the bottom four branches of the dendrogram, apparently showing similar behaviors.

The mean bias scores of all 20 models on all 30 cognitive biases are visualized in Figure 7. All models show significant biasedness on at least some of the tested cognitive biases. The vast majority of biases is positive, confirming that most cognitive biases present in humans can also be measured in LLMs. Only two of the 30 tested biases, *Status-Quo Bias* and *Disposition Effect*, were measured with strong negative direction, on average. On both biases, negative scores express a model’s pref-

³Chatbot Arena (scores from October 14, 2024)

	GPT-4o	GPT-4o mini	GPT-3.5 Turbo	Llama 3.1.405B	Llama 3.1.70B	Llama 3.1.8B	Llama 3.2.3B	Llama 3.2.1B	Claude 3 Haiku	Gemini 1.5 Pro	Gemini 1.5 Flash	Gemma 2.27B	Gemma 2.9B	Mistral Small	Mistral Large	WizardLM-2.8x22B	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

Figure 7: The heatmap shows the average bias scores for all evaluated models and biases.

erence for change. The Random models shows no biasedness on average, highlighting our metric’s strength as an unbiased estimator. One LLM demonstrating surprisingly low average biasedness is the smallest Llama model (1B parameters). For this model, we registered the highest decision failure rate (the model could not decide for an option in 33% of test cases), suggesting that this LLM’s general behavior may not be strongly grounded in good reasoning.

6 Conclusion

We have presented a comprehensive evaluation of 30 cognitive biases in 20 state-of-the-art LLMs. This contribution broadens the current understanding of cognitive biases in LLMs through a systematic and large-scale assessment under various

decision-making scenarios. We confirm early evidence from previous work suggesting that LLMs have cognitive biases and find that a majority of cognitive biases known in humans is also present in most LLMs. Human decision-makers considering to employ LLMs to enhance the quality of their decisions should be careful to select suitable models not only based on their reasoning capabilities but also based on their proneness to biases and should generally weigh their interest for faster and better decisions against the ethical implications.

In this work, we further demonstrated how our general-purpose test framework can be applied to generating tests for LLMs at a large scale and with high reliability. We publish our dataset of cognitive bias tests to guide developers of future LLMs in creating less biased and more reliable models.

7 Limitations

Our paper provides a systematic framework for defining and conducting cognitive bias tests with LLMs. While we have demonstrated our pipeline using management decision-making as an example and established a respective dataset with 30,000 test cases for cognitive biases, our framework is theoretically generalizable beyond just this domain and task. We provide some illustrative examples of applying our framework to other domains and test kinds in [Appendix A](#) but rely on future work to assess the framework’s versatility at scale. Our framework balances LLM generation and its benefit of cost-effectiveness with human control through templates with generalized instructions, which are similarly beneficial for other decision-making domains and use cases.

While over 180 cognitive biases are known in humans ([III and Benson, 2016](#)), our current dataset provides test cases for 30 of these biases. Our selection procedure utilized mentions in publications as an indicator for the relevance of biases in the chosen domain of managerial decision-making. As this may not be a perfectly reliable indicator for relevance and there are still over 150 cognitive biases not covered in our dataset, we invite other researchers to design tests for additional biases and domains.

Our test cases were generated with only one model, a GPT-4o LLM, chosen for its capabilities at the time of development. We also evaluate the same LLM on the dataset, which may give it an unfair advantage. We assume this influence to be low due to the detailed instructions in the templates giving the generating LLM clear restrictions on what to generate and how. Looking ahead, we anticipate that the majority of LLMs will soon possess the capability of generating test cases reliably. This development paves the way for a more widespread and effective application of our framework in the future.

In our evaluation, biasedness was calculated using discrete decisions made by the LLMs. Future work can also take into account token probabilities for an even more nuanced measurement and comparison of cognitive biases in LLMs.

8 Ethical Considerations

Our cognitive bias dataset of 30,000 test cases is one of the significant contributions of this paper. With this dataset, we also provide test cases for

biases related to social attributes, e.g., *Social Desirability Bias* and *Stereotyping*. The stereotypes in our dataset are generated by a GPT-4o LLM and are often mildly negative or can sometimes be considered neutral (for a detailed toxicity analysis, see [Figure 9 in Appendix F](#)). Therefore, more harmful stereotypes are not propagated but can also not be assessed with our dataset. Manually curated benchmarks must also be consulted to understand and mitigate stereotypes against social groups and cultures.

Although we present a large dataset on cognitive biases that allows for a comprehensive evaluation, it is important to understand that no benchmark can eliminate the need to evaluate an LLM for a specific use case to understand the risks. While our work can be used to factor in cognitive biases in LLM selection, it should by no means serve as a free pass for using LLMs for purely machine-based decision-making. Also, we ask anyone working with our dataset not to use it to train current or future models but apply it for evaluative purposes only.

Use of AI Assistants We used AI assistant tools to support us in creating the code for our framework. We did not use AI assistants for writing any sections of this paper.

Total Computational Budget Throughout this research project, we spent a total of USD 793.55 on various APIs to run inference with the evaluated LLMs. An overview of the APIs used can be found in [Table 6 in Appendix E](#).

References

- Klaus Abbink and Donna Harris. 2019. In-group favouritism and out-group discrimination in naturally occurring groups. *PloS one*, 14(9):e0221616.
- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Howard Abikoff, Mary Courtney, William E Pelham, and Harold S Koplewicz. 1993. Teachers’ ratings of disruptive behaviors: The influence of halo effects. *Journal of abnormal child psychology*, 21:519–533.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

- Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- George Ainslie and Nicholas Haslam. 1992. Hyperbolic discounting. In George Loewenstein and Jon Elster, editors, *Choice over time*, pages 57–92. Russell Sage Foundation, New York.
- Bill Albert and Tom Tullis. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes.
- Anthropic. 2024. *The claude 3 model family: Opus, sonnet, haiku*.
- David Antons and Frank T Piller. 2015. Opening the black box of “not invented here”: Attitudes, decision biases, and behavioral consequences. *Academy of Management perspectives*, 29(2):193–217.
- Linda Argote, Bill McEvily, and Ray Reagans. 2003. Managing knowledge in organizations: An integrative framework and review of emerging themes. *Management science*, 49(4):571–582.
- Kyrtin Atreides and David J Kelley. 2023. Cognitive biases in natural language: Automatically detecting, differentiating, and measuring bias in text. *Differentiating, and Measuring Bias in Text*.
- Markus Baer and Graham Brown. 2012. Blind in one eye: How psychological ownership of ideas affects the types of suggestions people adopt. *Organizational behavior and human decision processes*, 118(1):60–71.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Ray Ball and Ross Watts. 1979. Some additional evidence on survival biases. *The Journal of Finance*, 34(1):197–206.
- Jonathan Baron, Jane Beattie, and John C Hershey. 1988. Heuristics and biases in diagnostic reasoning: II. congruence, information, and certainty. *Organizational behavior and human decision processes*, 42(1):88–110.
- Thomas Bayes. 1763. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, 53:370–418.
- S.M. Beebe and R.H. Pherson. 2011. *Cases in Intelligence Analysis: Structured Analytic Techniques in Action*. SAGE Publications.
- Uri Benzion, Amnon Rapoport, and Joseph Yagil. 1989. Discount rates inferred from decisions: An experimental study. *Management science*, 35(3):270–284.
- Nicole Bergen and Ronald Labonté. 2020. “everything is perfect, and we have no problems”: detecting and limiting social desirability bias in qualitative research. *Qualitative health research*, 30(5):783–792.
- Bruno Biais and Martin Weber. 2009. Hindsight bias, risk perception, and investment performance. *Management Science*, 55(6):1018–1029.
- Sunali Bindra, Deepika Sharma, Nakul Parameswar, Sanjay Dhir, and Justin Paul. 2022. Bandwagon effect revisited: A systematic review to develop future research agenda. *Journal of Business Research*, 143:305–317.
- Bruce Blaine and Jennifer Crocker. 1993. Self-esteem and self-serving biases in reactions to positive and negative events: An integrative review. *Self-esteem: The puzzle of low self-regard*, pages 55–85.
- Donald E Bowen III, S McKay Price, Luke CD Stein, and Ke Yang. 2024. Measuring and mitigating racial bias in large language model mortgage underwriting. Available at SSRN 4812158.
- Anat Bracha and Donald J. Brown. 2012. *Affective decision making: A theory of optimism bias*. *Games and Economic Behavior*, 75(1):67–80.
- Gifford W Bradley. 1978. Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of personality and social psychology*, 36(1):56.
- Jack W Brehm. 1966. *A theory of psychological reactance*. Academic press.
- Sharon S Brehm and Jack W Brehm. 2013. *Psychological reactance: A theory of freedom and control*. Academic Press.
- Stephen J Brown, William Goetzmann, Roger G Ibbotson, and Stephen A Ross. 1992. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580.
- Roger Buehler, Dale Griffin, and Johanna Peetz. 2010. The planning fallacy: Cognitive, motivational, and social origins. In *Advances in experimental social psychology*, volume 43, pages 1–62. Elsevier.
- Roger Buehler, Dale Griffin, and Michael Ross. 1994. Exploring the “planning fallacy”: Why people underestimate their task completion times. *Journal of personality and social psychology*, 67(3):366.
- Christopher DB Burt and Simon Kemp. 1994. Construction of activity duration and time management potential. *Applied Cognitive Psychology*, 8(2):155–168.
- Sean D. Campbell and Steven A. Sharpe. 2009. Anchoring bias in consensus forecasts and its effect on market prices. *Journal of Financial and Quantitative Analysis*, 44(2):369–390.

- W Keith Campbell and Constantine Sedikides. 1999. Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of general Psychology*, 3(1):23–43.
- Mike Cardwell. 1999. *Dictionary of Psychology*. Fitzroy Dearborn, Chicago.
- John S Carroll. 1978. The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of experimental social psychology*, 14(1):88–96.
- Jihwan Chae, Kunil Kim, Yuri Kim, Gahyun Lim, Daeeun Kim, and Hackjin Kim. 2022. Ingroup favoritism overrides fairness when resources are limited. *Scientific reports*, 12(1):4560.
- Iain Chalmers and Robert Matthews. 2006. What are the implications of optimism bias in clinical research? *The Lancet*, 367(9509):449–450.
- Thierry Chaminade, Jessica Hodgins, and Mitsuo Kawato. 2007. Anthropomorphism influences perception of computer-animated characters' actions. *Social cognitive and affective neuroscience*, 2(3):206–216.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.
- Jay J.J Christensen-Szalanski and Cynthia Fobian Williamson. 1991. The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 48(1):147–168.
- John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.
- Maia B. Cook and Harvey S. Smallman. 2008. Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, 50(5):745–754. PMID: 19110834.
- W. Coombs and Sherry Holladay. 2006. Unpacking the halo effect: Reputation and crisis management. *Journal of Communication Management*, 10:123–137.
- William H Cooper. 1981. Ubiquitous halo. *Psychological bulletin*, 90(2):218.
- Douglas P Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, 24(4):349.
- Mike Dacey. 2017. Anthropomorphism as cognitive bias. *Philosophy of Science*, 84(5):1152–1164.
- David M. DeJoy. 1989. The optimism bias and traffic accident risk perception. *Accident Analysis & Prevention*, 21(4):333–340.
- James Dillard and Lijiang Shen. 2005. On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72:144–168.
- James N Druckman. 2001. The implications of framing effects for citizen competence. *Political behavior*, 23:225–256.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.
- Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.
- Allen L Edwards. 1953. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of applied Psychology*, 37(2):90.
- Allen L Edwards. 1957. *The social desirability variable in personality assessment and research*. Dryden Press.
- Ward Edwards. 1982. Conservatism in human information processing (excerpted). In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York. Original work published 1968.
- Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.
- Eva Eigner and Thorsten Härdler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.
- Dirk M Elston. 2021. Survivorship bias. *Journal of the American Academy of Dermatology*.

- Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864.
- Eyal Ert and Ido Erev. 2013. **On the descriptive value of loss aversion in decisions under risk: Six clarifications.** *Judgment and Decision Making*, 8(3):214–235.
- Jim AC Everett, Nadira S Faber, and Molly Crockett. 2015. Preferences and beliefs in ingroup favoritism. *Frontiers in behavioral neuroscience*, 9:126656.
- Chris Fife-Schaw and Julie Barnett. 2004. Measuring optimistic bias. *Doing social psychology research*, pages 54–74.
- Peter Fischer, Stephen Lea, Andreas Kastenmüller, Tobias Greitemeyer, Julia Fischer, and Dieter Frey. 2011. **The process of selective exposure: Why confirmatory information search weakens over time.** *Organizational Behavior and Human Decision Processes*, 114(1):37–48.
- Cassandra Flick and Kimberly Schweitzer. 2021. **Influence of the fundamental attribution error on perceptions of blame and negligence.** *Experimental Psychology*, 68:175–188.
- Valerie S. Folkes. 1988. **The availability heuristic and perceived risk.** *Journal of Consumer Research*, 15(1):13–23.
- Robert Forsythe, Joel L Horowitz, Nathan E Savin, and Martin Sefton. 1994. Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369.
- Feng Fu, Corina E Tarnita, Nicholas A Christakis, Long Wang, David G Rand, and Martin A Nowak. 2012. Evolution of in-group favoritism. *Scientific reports*, 2(1):460.
- Javier Fuenzalida, Gregg G. Van Ryzin, and Asmus Leth Olsen. 2021. **Are managers susceptible to framing effects? an experimental study of professional judgment of performance metrics.** *International Public Management Journal*, 24(3):314–329.
- Adrian Furnham and Hua Chu Boo. 2011. **A literature review of the anchoring effect.** *The Journal of Socio-Economics*, 40(1):35–42.
- Adele Gabrielcik and Russell H Fazio. 1984. Priming and frequency estimation: A strict test of the availability heuristic. *Personality and Social Psychology Bulletin*, 10(1):85–89.
- Kristel M. Gallagher and John A. Updegraff. 2011. **Health Message Framing Effects on Attitudes, Intentions, and Behavior: A Meta-analytic Review.** *Annals of Behavioral Medicine*, 43(1):101–116.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Rebecca L Guilbault, Fred B Bryant, Jennifer Howard Brockway, and Emil J Posavac. 2004. A meta-analysis of research on hindsight bias. *Basic and applied social psychology*, 26(2-3):103–117.
- Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.
- Laura Hanu and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.
- Peter Harris. 1996. Sufficient grounds for optimism?: The relationship between perceived controllability and optimistic bias. *Journal of Social and Clinical Psychology*, 15(1):9–52.
- Martie G Haselton, Daniel Nettle, and Paul W Andrews. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 724–746.
- Scott A Hawkins and Reid Hastie. 1990. Hindsight: Biased judgments of past events after the outcomes are known. *Psychological bulletin*, 107(3):311.
- Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.
- Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.
- James Hedlund. 2000. Risky business: safety regulations, risk compensation, and individual behavior. *Injury prevention*, 6(2):82–89.
- F. Heider. 1982. *The Psychology of Interpersonal Relations*. Lawrence Erlbaum Associates.
- Steven J Heine and Darrin R Lehman. 1995. Cultural variation in unrealistic optimism: Does the west feel more vulnerable than the east? *Journal of personality and social psychology*, 68(4):595.
- Pamela W Henderson and Robert A Peterson. 1992. **Mental accounting and categorization.** *Organizational Behavior and Human Decision Processes*, 51(1):92–117.

- Richard J Herrnstein. 1961. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, 4(3):267.
- Nic Hooper, Ates Erdogan, Georgia Keen, Katharine Lawton, and Louise McHugh. 2015. Perspective taking reduces the fundamental attribution error. *Journal of Contextual Behavioral Science*, 4(2):69–72.
- John Manoogian III and Buster Benson. 2016. [The cognitive bias codex](#). Wikimedia Commons. Wikipedia's complete (as of 2016) list of cognitive biases, arranged and designed by John Manoogian III (jm3). Categories and descriptions originally by Buster Benson.
- Tiffany A Ito, Jeff T Larsen, N Kyle Smith, and John T Cacioppo. 1998. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of personality and social psychology*, 75(4):887.
- Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.
- Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706.
- Jia Jin, Wuke Zhang, and Mingliang Chen. 2017. How consumers are affected by product descriptions in online shopping: Event-related potentials evidence of the attribute framing effect. *Neuroscience Research*, 125:21–28.
- Edward E Jones and Victor A Harris. 1967. The attribution of attitudes. *Journal of Experimental Social Psychology*, 3(1):1–24.
- Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. 1986. Fairness and the assumptions of economics. *Journal of business*, pages S285–S300.
- Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. 1990. Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98(6):1325–1348.
- Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206.
- Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.
- Daniel Kahneman and Amos Tversky. 1982. *Intuitive prediction: Biases and corrective procedures*, page 414–421. Cambridge University Press.
- Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.
- Mahammed Kamruzzaman, Md Minul Islam Shovon, and Gene Louis Kim. 2023. Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models. *arXiv preprint arXiv:2309.08902*.
- David E Kanouse and L Reid Hanson Jr. 1972. Negativity in evaluations. In *E.E. Jones, D. E. Kanouse, S. Valins, H. H. Kelley, R. E. Nisbett, & B. Weiner (Eds.), Attribution: Perceiving the causes of behavior*, pages 47–62. Morristown, NJ: General Learning Press.
- Heather Kappes, Ann Harvey, Terry Lohrenz, Pendleton Montague, and Tali Sharot. 2020. Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, 23:1–8.
- Ralph Katz and Thomas J Allen. 1982. Investigating the not invented here (nih) syndrome: A look at the performance, tenure, and communication patterns of 50 r & d project groups. *R&d Management*, 12(1):7–20.
- Ran Kivetz. 1999. Advances in research on mental accounting and reason-based choice. *Marketing Letters*, 10:249–266.
- Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418.
- Doron Kliger and Andrey Kudryavtsev. 2010. The availability heuristic and investors' reaction to company-specific events. *The journal of behavioral finance*, 11(1):50–65.
- Jack L Knetsch. 1989. The endowment effect and evidence of nonreversible indifference curves. *The American Economic Review*, 79(5):1277–1284.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.
- Tatiana Kostova and Kendall Roth. 2002. Adoption of an organizational practice by subsidiaries of multinational corporations: Institutional and relational effects. *Academy of management journal*, 45(1):215–233.
- Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047.
- Sheldon J Lachman and Alan R Bass. 1985. A direct study of halo effect. *The journal of psychology*, 119(6):535–540.

- David Laibson. 1997. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478.
- Ellen J Langer. 1975. The illusion of control. *Journal of personality and social psychology*, 32(2):311.
- Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen W White, and Sujay Kumar Jauhar. 2023. Making large language models better data creators. *arXiv preprint arXiv:2310.20111*.
- H. Leibenstein. 1950. Bandwagon, snob, and veblen effects in the theory of consumers' demand. *The Quarterly Journal of Economics*, 64(2):183–207.
- Lance Leuthesser, Chiranjeev Kohli, and Katrin Harich. 1995. Brand equity: The halo effect measure. *European Journal of Marketing*, 29:57–66.
- Irwin P. Levin, Sandra L. Schneider, and Gary J. Gaeth. 1998. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76(2):149–188.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.
- Falk Lieder, Tom Griffiths, and Noah Goodman. 2012. Burn-in, bias, and the rationality of anchoring. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *Preprint, arXiv:2406.15126*.
- Ashley Luckman, Hossam Zeitoun, Andrea Isoni, Graham Loomes, Ivo Vlaev, Nattavudh Powdthavee, and Daniel Read. 2021. Risk compensation during covid-19: The impact of face mask usage on social distancing. *Journal of Experimental Psychology: Applied*, 27(4):722.
- Dan P. Ly, Paul G. Shekelle, and Zirui Song. 2023. Evidence for Anchoring Bias During Physician Decision-Making. *JAMA Internal Medicine*, 183(8):818–823.
- Colin MacLeod and Lynlee Campbell. 1992. Memory accessibility and probability judgments: an experimental evaluation of the availability heuristic. *Journal of personality and social psychology*, 63(6):890.
- Olivia Macmillan-Scott and Mirco Musolesi. 2024. (ir) rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.
- Keith M Marzilli Ericson and Andreas Fuster. 2014. The endowment effect. *Annu. Rev. Econ.*, 6(1):555–579.
- Yusufcan Masatlioglu and Efe A Ok. 2005. Rational choice with status quo bias. *Journal of economic theory*, 121(1):1–29.
- Anne M McCarthy, F David Schoorman, and Arnold C Cooper. 1993. Reinvestment decisions by entrepreneurs: Rational decision-making or escalation of commitment? *Journal of business venturing*, 8(1):9–24.
- Susan Miles and Victoria Scaife. 2003. Optimistic bias and food. *Nutrition research reviews*, 16(1):3–19.
- Dale T Miller and Michael Ross. 1975. Self-serving biases in the attribution of causality: Fact or fiction? *Psychological bulletin*, 82(2):213.
- Daniel Mochon and Shane Frederick. 2013. Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, 122(1):69–79.
- Carey K Morewedge and Colleen E Giblin. 2015. Explanations of the endowment effect: an integrative review. *Trends in cognitive sciences*, 19(6):339–348.
- MSCI and S&P Global. 2023. Global industry classification standard (gics). A classification standard jointly developed by MSCI and S&P Global for categorizing companies into sectors and industries. Published March 17, 2023, retrieved October 1, 2024.
- Joel Myerson and Sandra Hale. 1984. Practical implications of the matching law. *Journal of Applied Behavior Analysis*, 17(3):367–380.
- Richard Nadeau, Edouard Cloutier, and J.-H. Guay. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review / Revue internationale de science politique*, 14(2):203–213.
- Rosanna Nagtegaal, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. 2020. Designing to debias: Measuring and reducing public managers' anchoring bias. *Public Administration Review*, 80(4):565–576.
- Raymond Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220.
- Richard E Nisbett and Timothy D Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250.
- Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 340–350, New York, NY, USA. Association for Computing Machinery.

- Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. 2024. Do language models exhibit the same cognitive biases in problem solving as human learners? *arXiv preprint arXiv:2401.18070*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Gary Pan, Shan L Pan, Michael Newman, and Donal Flynn. 2006. Escalation and de-escalation of commitment: a commitment transformation analysis of an e-government project. *Information Systems Journal*, 16(1):3–21.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. **Don't blame the annotator: Bias already starts in the annotation instructions.** In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sam Peltzman. 1975. The effects of automobile safety regulation. *Journal of political Economy*, 83(4):677–725.
- Stephen Pilli. 2023. Exploring conversational agents as an effective tool for measuring cognitive biases in decision-making. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*, pages 1–5. IEEE.
- Sivan Portal, Russell Abratt, and Michael Bendixen. 2018. Building a human brand: Brand anthropomorphism unravelled. *Business Horizons*, 61(3):367–374.
- Diane Proudfoot. 2011. Anthropomorphism and ai: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6):950–957.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittweiser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Heidi R. Riggio and Amber L. Garcia. 2009. **The power of situations: Jonestown and the fundamental attribution error.** *Teaching of Psychology*, 36(2):108–112.
- Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Neal J. Roese and Kathleen D. Vohs. 2012. **Hindsight bias.** *Perspectives on Psychological Science*, 7(5):411–426. PMID: 26168501.
- J.H. Rohlf. 2003. **Bandwagon Effects in High-technology Industries.** MIT Press.
- Benjamin D Rosenberg and Jason T Siegel. 2018. A 50-year review of psychological reactance theory: Do not read this article. *Motivation Science*, 4(4):281.
- Lee Ross. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*, volume 10, pages 173–220. Elsevier.
- David Rozado. 2024. The political preferences of llms. *arXiv preprint arXiv:2402.01789*.
- Paul Rozin and Edward B Royzman. 2001. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320.
- Ariel Rubinstein. 2003. “economics and psychology”? the case of hyperbolic discounting. *International Economic Review*, 44(4):1207–1216.
- Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95.
- Ömür Saltik, Wasim Rehman, Rıdvan Söyü, Suleyman Degirmen, and Ahmet Sengonul. 2023. **Predicting loss aversion behavior with machine-learning methods.** *Humanities and Social Sciences Communications*, 10.
- William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty*, 1:7–59.
- Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.
- Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvarg, Tawsifur Rahman, Ji Woong Kim, Rojin Ziae, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*.
- Jeffrey B Schmidt and Roger J Calantone. 2002. Escalation of commitment during new product development. *Journal of the academy of marketing science*, 30:103–118.
- Rüdiger Schmitt-Beck. 2015. **Bandwagon Effect.** John Wiley & Sons, Ltd.
- Hamish GW Seaward and Simon Kemp. 2000. Optimism bias and student debt. *New Zealand journal of psychology*, 29(1):17–19.
- Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword:

- A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782.
- Jonathan Shalev. 2000. Loss aversion equilibrium. *International Journal of Game Theory*, 29:269–287.
- Hersh Shefrin and Meir Statman. 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of finance*, 40(3):777–790.
- James Shepperd, Wendi Malone, and Kate Sweeny. 2008. Exploring causes of the self-serving bias. *Social and Personality Psychology Compass*, 2(2):895–908.
- Emmanuel Marques Silva, Rafael de Lacerda Moreira, and Patricia Maria Bortolon. 2023. [Mental accounting and decision making: a systematic literature review](#). *Journal of Behavioral and Experimental Economics*, 107:102092.
- Herbert A Simon. 1990. Bounded rationality. *Utility and probability*, pages 15–18.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mardavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Dustin Sleesman, Anna Lennard, Gerry McNamara, and Donald Conlon. 2018. [Putting escalation of commitment in context: A multi-level review and analysis](#). *Academy of Management Annals*, 12:annals.2016.0046.
- Mark Snyder and William Swann. 1978. [Hypothesis testing in social judgment](#). *Journal of Personality and Social Psychology*, 36:1202–1212.
- Melvin Snyder and Arthur Frankel. 1976. [Observer bias: A stringent test of behavior engulfing the field](#). *Journal of Personality and Social Psychology*, 34:857–864.
- Melvin L Snyder, Walter G Stephan, and David Rosenfield. 1976. Egotism and attribution. *Journal of Personality and Social Psychology*, 33(4):435.
- Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.
- Barry M. Staw. 1976. [Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action](#). *Organizational Behavior and Human Performance*, 16(1):27–44.
- Barry M. Staw. 1981. [The escalation of commitment to a course of action](#). *The Academy of Management Review*, 6(4):577–587.
- Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. 2015. Understanding psychological reactance. *Zeitschrift für Psychologie*.
- Joachim Stöber. 2001. The social desirability scale-17 (sds-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, 17(3):222.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.
- Alaina N Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption. *arXiv preprint arXiv:2304.01358*.
- Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.
- Richard Thaler. 1980. Toward a positive theory of consumer choice. *Journal of economic behavior & organization*, 1(1):39–60.
- Richard Thaler. 1985. [Mental accounting and consumer choice](#). *Marketing Science*, 4(3):199–214.
- Richard H Thaler. 1999. Mental accounting matters. *Journal of Behavioral decision making*, 12(3):183–206.
- Suzanne C Thompson. 1999. Illusions of control: How we overestimate our personal influence. *Current Directions in Psychological Science*, 8(6):187–190.
- Edward Lee Thorndike. 1920. [A constant error in psychological ratings](#). *Journal of Applied Psychology*, 4:25–29.
- Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.
- Sabrina M. Tom, Craig R. Fox, Christopher Trepel, and Russell A. Poldrack. 2007. [The neural basis of loss aversion in decision-making under risk](#). *Science*, 315(5811):515–518.
- Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. [Metaphor understanding challenge dataset for LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.
- Amos Tversky and Daniel Kahneman. 1973. [Availability: A heuristic for judging frequency and probability](#). *Cognitive Psychology*, 5(2):207–232.

- Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.
- Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.
- Amos Tversky and Daniel Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293.
- Amos Tversky and Daniel Kahneman. 1991. Loss Aversion in Riskless Choice: A Reference-Dependent Model*. *The Quarterly Journal of Economics*, 106(4):1039–1061.
- Amrisha Vaish, Tobias Grossmann, and Amanda Woodward. 2008. Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological bulletin*, 134(3):383.
- Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.
- Lyn M Van Swol. 2007. Perceived importance of information: The effects of mentioning information, shared information bias, ownership bias, reiteration, and confirmation bias. *Group processes & intergroup relations*, 10(2):239–256.
- A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- T Franklin Waddell. 2019. Can an algorithm reduce the perceived bias of news? testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & mass communication quarterly*, 96(1):82–100.
- Abraham Wald. 1943. A method of estimating plane vulnerability based on damage of survivors. *Statistical Research Group, Columbia University*. CRC, 432.
- Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.
- Liman Wang, Hanyang Zhong, Wenting Cao, and Zeyuan Sun. 2024a. Balancing rigor and utility: Mitigating cognitive biases in large language models for multiple-choice questions. *Preprint*, arXiv:2406.10999.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.
- Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024b. Codeclm: Aligning language models with tailored synthetic data. *arXiv preprint arXiv:2404.05875*.
- P. C. Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140.
- Peter C. Wason. 1966. Reasoning. In Peter C. Wason, editor, *New Horizons in Psychology*, pages 135–151. Penguin Books.
- Martin Weber and Colin F Camerer. 1998. The disposition effect in securities trading: An experimental analysis. *Journal of Economic Behavior & Organization*, 33(2):167–184.
- Neil D. Weinstein. 1989. Optimistic biases about personal risks. *Science*, 246(4935):1232–1233.
- Christopher G Wetzel, Timothy D Wilson, and James Kort. 1981. The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology*, 17(4):427–439.
- Gerald JS Wilde. 1982. The theory of risk homeostasis: implications for safety and health. *Risk analysis*, 2(4):209–225.
- Anna Winterbottom, Hilary L Bekker, Mark Conner, and Andrew Mooney. 2008. Does narrative information bias individual's decision making? a systematic review. *Social science & medicine*, 67(12):2079–2088.
- Huiping Wu and Shing-On Leung. 2017. Can likert scales be treated as interval scales?—a simulation study. *Journal of social service research*, 43(4):527–532.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.
- Zhentao Xie, Jiabao Zhao, Yilei Wang, Jinxin Shi, Yan-hong Bai, Xingjiao Wu, and Liang He. 2024. Mindscope: Exploring cognitive biases in large language models through multi-agent systems. *arXiv preprint arXiv:2410.04452*.
- Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United

Arab Emirates. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xiting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Framework: Application Examples

We demonstrate two examples of the framework’s universality feature. Table 4 features an adaptation of the *Bandwagon Effect* testing procedure to the medical domain. Table 3 provides an example of a common testing procedure from the theory of mind research.

B Cognitive Biases

B.1 Conservatism

Conservatism, also known as *conservatism bias*, refers to the tendency to insufficiently revise one’s beliefs when new evidence becomes known. Edwards (1982) describes that people update their opinions when presented with new evidence but do so more slowly than Bayes’ theorem (Bayes, 1763) would demand.

Our test design presents the model with two decision alternatives, A and B. Each test case first presents three pieces of evidence suggesting that A

is better than B, followed by a conclusion that A is clearly better than B, representing the model’s prior belief. We then show three pieces of new evidence suggesting that B is better than A. After seeing that new evidence, the model is asked for its revised preference for either A or B on a 7-step Likert scale σ_1 with the midpoint representing indifference.

To account for any objective differences in the strengths of the evidence for A and B, we reverse the order of A and B between control and treatment. Only if the model consistently prefers the alternative that was introduced first, conservatism is present. We measure the strength of the bias as the consistent preference of the first alternative over the second one.

B.2 Anchoring

Anchoring, also known as *anchoring bias* or *anchoring effect*, is a phenomenon of making “estimates, which are biased toward the initially presented values” (Tversky and Kahneman, 1974), potentially irrelevant ones. This effect has been elicited in several settings (Furnham and Boo, 2011). Anchoring is investigated across different domains, including finance (Campbell and Sharpe, 2009), management (Nagtegaal et al., 2020), healthcare (Ly et al., 2023), and artificial intelligence (Lieder et al., 2012; Nourani et al., 2021).

We approach the testing by directly following the comparative judgment paradigm (Mochon and Frederick, 2013). In control and treatment, the LLM is prompted to estimate a variable. Additionally, the treatment variant contains an instruction to first evaluate the variable relative to the provided numerical value. This value serves as the anchor in the test design.

The anchoring effect is thus identified for deviations between the estimations in the anchor-free and anchored formulation. The answers are obtained on an 11-point percentage scale σ_2 .

B.3 Stereotyping

A stereotype is a generalized belief about a particular group of people (Cardwell, 1999).

To test the presence of stereotyping in LLMs, we define a set of groups with common stereotypes, covering different genders, ethnicities, sexual orientations, religious beliefs, and job types. We then introduce a decision situation where the decision heavily depends on knowing a certain group of people well and instruct the model to estimate a particular characteristic of that group. In treatment,

the model is told what the group is (e.g., Muslims), whereas in control, it is not.

The model can choose from four options describing the characteristics of the group, where two options represent characteristics stereotypical of that group and two options represent characteristics atypical of that group. For each pair of options, one is typical of people overall, while the other is atypical of people overall. If the model switches from choosing an atypical characteristic to a stereotypical one once the particular group is known, we conclude that the model exhibits stereotyping. In the inverse case, it would exhibit negative stereotyping. We obtain answers on a 7-point Likert scale σ_1 .

B.4 Social Desirability Bias

Social desirability bias is “the tendency to present oneself and one’s social context in a way that is perceived to be socially acceptable” (Bergen and Labonté, 2020). It is often studied in the context of surveys where it refers to the tendency to answer survey questions in a way that will be viewed favorably by others (Krumpal, 2013). Edwards (1953) introduced the notion of social desirability describing the “relationship between the judged desirability of a trait and the probability that the trait will be endorsed”. The bias has been studied extensively in survey respondents self-reporting their personality traits showing a “tendency of subjects to attribute to themselves statements which are desirable and reject those which are undesirable” (Edwards, 1957).

Common testing procedures rely on scales such as the Social Desirability Scale (SDS) (Edwards, 1957), the Marlowe-Crowne Social Desirability Scale (M-C SDS) (Crowne and Marlowe, 1960), or the Social Desirability Scale-17 (SDS-17) (Stöber, 2001), which include a number of statements about personality traits which are either clearly socially desirable or undesirable, e.g., “I’m always willing to admit it when I make a mistake” (Crowne and Marlowe, 1960). These scales can be used to test how many times a subject responds with a socially desirable answer.

Our test procedure is inspired by Albert and Tullis (2013), who report that people tend to follow socially desirable norms more strictly in public settings as opposed to anonymous settings. We ask the LLM to express whether a statement is true or false as it pertains to the LLM. In control, we note that the LLM’s answer will be treated confi-

dentially and not be shared with anyone. In treatment, we note that the LLM’s answer will be made public and can be linked back to the LLM. We sample statements from the M-C SDS (Crowne and Marlowe, 1960). From the scale, we remove 17 statements describing emotions, thoughts, or real-world interactions which are not applicable to LLMs, leaving 16 statements testable with LLMs.

We obtain answers on a 7-point Likert scale σ_1 . The metric takes a value of 1 only if the model self-reports undesirable behavior in control, the anonymous setting, but then chooses desirable behavior in treatment, the public setting, and -1 in the reverse case.

B.5 Loss Aversion

Proposed by Kahneman and Tversky (1979), loss aversion is present when the “disutility of giving up an object is greater than the utility associated with acquiring it” (Kahneman et al., 1991), i.e., when losses are perceived to be psychologically more powerful than gains. Well-established, this bias has been investigated in both risky and riskless (Tversky and Kahneman, 1991) contexts from various perspectives, including neuroscience (Tom et al., 2007), game theory (Shalev, 2000), and machine learning (Saltik et al., 2023).

We base our testing on the variation of the standard *Samuelson’s colleague problem* formulated in Ert and Erev (2013). The model is presented with a choice of two options with the material outcomes $f_{1,2}$ designed as follows ($a > 0$ denotes the commodity amount, p denotes probability):

$$f_1 = a, a > 0, \text{i.e., guaranteed gain} \quad (3)$$

$$f_2 = \begin{cases} \lambda a, \lambda > 2 & \text{with } p = \frac{1}{2} \\ -a, & \text{with } p = \frac{1}{2} \end{cases} \quad (4)$$

The second option, while being risky due to a potential loss, yields a more profitable outcome in expectation. In control and treatment, we switch the positions of the two options to account for the response bias. Loss aversion is thus present when the LLM consistently opts for the deterministic option, and we utilize a 7-point Likert scale σ_1 to obtain answers.

B.6 Halo Effect

The halo effect is originally defined in Thorndike (1920) and is commonly known as “the influence

of a global evaluation on evaluations of individual attributes” (Nisbett and Wilson, 1977), even when there is sufficient evidence for their independence. Cooper (1981) generalizes the definition to the presence of correlation between two independent attributes. Notably persistent (Wetzel et al., 1981), this bias is well-studied in the fields of consumer science (Leuthesser et al., 1995), public relations (Coombs and Holladay, 2006), and education (Abikoff et al., 1993).

We build on the testing procedure of Lachman and Bass (1985). In both control and treatment, an asset is presented to the LLM, and the model is prompted to evaluate a concrete attribute of this asset. In treatment, the halo is additionally introduced: a separate independent attribute of this asset is described either positively or negatively.

The halo effect is present in cases of the estimation shift in treatment compared to control, either a positive one provided with a positive halo or a negative one given a negative halo. The symmetrical behavior results in the opposite effect. We obtain answers to the halo effect test on a 7-point Likert scale σ_1 .

B.7 Reactance

Reactance refers to “an unpleasant motivational arousal that emerges when people experience a threat to or loss of their free behaviors” (Steindl et al., 2015). Rosenberg and Siegel (2018) present an extensive review of reactance theory. Reactance theory was first proposed by Brehm (1966), who found that individuals tend to be motivated to regain their behavioral freedoms when these freedoms are reduced or threatened (Brehm, 1966; Brehm and Brehm, 2013). The level of reactance is influenced by the importance of the threatened freedom and the strength of the threat as perceived by the individual (Steindl et al., 2015).

Our test design is based on the procedure proposed by Dillard and Shen (2005), who measure reactance in the different responses of subjects to either a low-threat or a high-threat scenario. We describe a behavior where the test taker previously had the freedom to choose if and how often to engage in this behavior. This is followed by a number of facts describing the negative consequences of this behavior. In control, these facts are presented as part of a low-threat framing and in treatment as part of a high-threat framing.

Specifically, our low-threat scenario recommends that the subject changes his/her behavior

(e.g., “consider doing it responsibly”) while the high-threat scenario demands a change of behavior (e.g., “you have to stop it”).

To measure the effect, we present the model with options describing different levels of engagement with the behavior. An increased engagement with the behavior from the low-threat to the high-threat variant indicates the presence of reactance (i.e., an adverse response to the threat). We obtain the answers to the effect on an 11-point percentage scale σ_2 .

B.8 Confirmation Bias

Originally described by Wason (1960), confirmation bias commonly refers to the “inclination to discount information that contradicts past judgments” (Kappes et al., 2020). Confirmation bias is known to arise during the search and the interpretation of information, as well as their combination (Klayman, 1995; Nickerson, 1998). Approaches to testing this bias include variations of the classical Wason selection task (Wason, 1966), two-phase evidence-seeking paradigms (Cook and Smallman, 2008; Fischer et al., 2011), and weighting of provided evidence (Snyder and Swann, 1978; Beebe and Pherson, 2011).

We directly employ the latter technique for the testing. In the control and treatment procedures, the model is associated with a proposal and is presented with a set of arguments against it. In control, the model is said to have not yet decided on its proposal. On the contrary, in treatment, the LLM is prompted to have already made the decision, i.e., this decision is considered the model’s past judgment. In both variants, the LLM is prompted to select the number of presented arguments that are relevant while and after making the decision in control and treatment, respectively.

The answers of the LLM to the confirmation bias test are obtained on an 11-point scale σ_2 . The metric reflects the extent to which this selection is imbalanced between the cases of absence and presence of the past judgment.

B.9 Not Invented Here

The not-invented-here syndrome (NIH) is commonly described as an attitudinal bias against the knowledge that an individual perceives as external (Katz and Allen, 1982; Kostova and Roth, 2002). The framework by Antons and Piller (2015) depicts two key elements of this bias: first, the source of knowledge, distinguishing organizational, con-

textual (disciplinary), and spatial (geographical) externality. Second, the underestimation of the value of this knowledge or the overestimation of the costs of its obtainment. There may be different underlying mechanisms causing this syndrome, including ego-defensive (e.g., Baer and Brown, 2012) or utilitarian functions (e.g., Argote et al., 2003).

Our test follows the concept of value estimation by introducing a decision scenario and asking for the evaluation of a respective proposal. In control, the test case informs that one proposal is suggested by a colleague in the decision-maker’s own team. In treatment, the statement is changed to indicate the external source of the proposal, whereby we sample the type of externality to be either organizational, contextual, or spatial. For spatial externality we additionally sample the country of the colleague. Hereby, we include the three most populated countries per continent (only two for North America and Oceania).

A lower evaluation of the proposal, when it is described as from an external source, indicates the presence of the not-invented-here syndrome. The answers are obtained on a 7-point Likert scale σ_1 .

B.10 Illusion of Control

An illusion of control is “an expectancy of a personal success probability inappropriately higher than the objective probability would warrant” (Langer, 1975). In other words, people tend to overestimate their ability to control events (Thompson, 1999). Langer (1975), who named the illusion of control, reports that factors typical of skill situations, such as *competition*, *choice*, *familiarity*, and *involvement*, can cause individuals to feel inappropriately confident.

Our test design builds onto the findings by Langer (1975). We describe an activity that typically has some success probability x . We then ask the model to judge its own success probability assuming that it would conduct the activity. We also add factors from skill situations to the description.

Specifically, we describe a situation where the model has recently been hired by an organization to supervise a business activity which typically has a success probability of $x = 50\%$. To enrich the situation with bias-inducing factors, we randomly add either a description of (A) how the model is *competing* against others, (B) how it has full freedom of *choice* regarding how to run the activity, (C) how it is highly *familiar* with the activity, (D) how it will be deeply *involved* in the execution, or

(E) no description of an additional factor.

We measure the illusion of control as any success probability judged by the model that exceeds the objective success probability x . The answers are obtained on an 11-point percentage scale σ_2 .

B.11 Survivorship Bias

Survivorship bias is a form of *selection bias* that can occur when we only focus on data from subjects who “proceeded past a selection or elimination process” (a.k.a. “survivors”) “while overlooking those who did not” (Elston, 2021). Hence, survivorship bias can cause us to draw conclusions about the general population of subjects that are biased toward the survivors. The bias was first described by statistician Wald (1943) who studied World War II aircraft and the damage they incurred during battle. Since then, survivorship is often observed in financial and investment contexts (Brown et al., 1992; Ball and Watts, 1979).

To test the presence of the bias in LLMs, we describe a decision-making task that involves choosing somehow *good* entities from a pool that contains both *good* and *bad* entities. We then introduce a characteristic of these entities that could be used to separate *good* from *bad* entities and define what percentages x_{good} and x_{bad} of the entities have this characteristic among the *good* and the *bad* entities, respectively. x_{good} and x_{bad} are sampled from the same narrow interval and are very close together. In control, we report both x_{good} and x_{bad} to the model, whereas in treatment, we only report x_{good} , reflecting a situation where we only focus on the survivors. Lastly, we ask the model how important it thinks the characteristic is to distinguish *good* from *bad* entities.

Specifically, we sample both x_{good} and x_{bad} from a relatively small interval $[0.90, 0.95]$ to simulate a situation where the difference is likely not statistically significant between the two groups and both, x_{good} and x_{bad} , are large.

We measure the strength of survivorship bias as the excess importance of the characteristic in treatment over control as judged by the model. The answers are obtained on a 7-point Likert scale σ_1 .

B.12 Escalation of Commitment

First examined in Staw (1976), escalation of commitment, also known as *commitment bias*, refers to “the act of ‘carrying on’ with questionable or failing courses of action” (Sleesman et al., 2018). Due to its nature, the bias has been extensively

studied, among others, in finance (McCarthy et al., 1993), governance (Pan et al., 2006), and research & development (Schmidt and Calantone, 2002).

Our procedure is based on the findings of Staw (1981), which emphasizes the connection between escalation of commitment and responsibility. In this paradigm, the model is presented with a decision that has been made in the past and evidence suggesting that this decision should have been made differently. We then ask the model for its intention to change the decision. In the control variant, the past decision is attributed to the LLM, and in the treatment variant — to another independent actor.

Greater commitment to decisions made by the subject indicates the presence of the bias. The answers to the effect's testing are measured on an 11-point percentage scale σ_2 .

B.13 Information Bias

Information bias denotes the heuristic to request new information even when none of the potential findings could change the basis for action, which was demonstrated for the medical domain by Baron et al. (1988). In their experiments, subjects chose to run medical tests that could not change the prior treatment decision for the hypothetical patients. The term information bias is, however, also employed as a catch-all phrase for a group of information-related biases (e.g., confirmation bias), and further specifications exist, such as *narrative information bias* (Winterbottom et al., 2008) or *shared information bias* (Van Swol, 2007).

For our tests, we employ a simplified version of the experiment by Baron et al. (1988), with a description of a decision event and a currently considered course of action. In control, we ask the model about its confidence in advancing with this course. In treatment, we instead ask if the model needs any additional information to advance with this course. Answers indicating strong confidence in the control variant and a high need for additional information in the treatment variant suggest the presence of information bias.

We obtain answers to the information bias test on a 7-point Likert scale σ_1 .

B.14 Mental Accounting

Proposed by Thaler (1985), mental accounting is described as “a cognitive process whereby people treat resources differently depending on how they are labeled and grouped, which consequently leads

to violations of the normative economic principle of fungibility” (Kivetz, 1999), i.e., the same resources in different mental accounts are not equivalent. An extensive review of various facets of this effect and its presence in different applications is assembled in Silva et al. (2023).

We frame our test in direct accordance with the “theater ticket” experiment in Tversky and Kahneman (1981), which is a standard technique to elicit mental accounting (Thaler, 1999; Henderson and Peterson, 1992). In both variants, an investment decision is described. In control, this investment is lost irrevocably, and the model is prompted to choose whether or not to make another such investment to compensate for the lost one. The treatment variant, in turn, features a separate, independent loss of the same amount. The LLM is then prompted to decide if the initial investment decision nonetheless holds or not.

A discrepancy in these two decisions indicates the presence of mental accounting, i.e., it shows that the equal losses described belong to different, non-equivalent mental accounts. The answers are obtained on a 7-point Likert scale σ_1 .

B.15 Optimism Bias

Optimism bias represents the “tendency to overestimate the likelihood of favorable future outcomes and underestimate the likelihood of unfavorable future outcomes” (Bracha and Brown, 2012). This effect is ubiquitous (Weinstein, 1989) and impacts diverse aspects of human activities: ethics in research (Chalmers and Matthews, 2006), finance (Seaward and Kemp, 2000), people’s health (Miles and Scaife, 2003) and safety (DeJoy, 1989). Fife-Schaw and Barnett (2004) identifies two main approaches to measure the optimism bias: direct and indirect comparisons.

For our testing, we adopt the latter technique (Heine and Lehman, 1995; Harris, 1996). Either a positive or a negative situation is introduced. In control and treatment, the model is prompted to estimate the likelihood of facing such a situation for an abstract subject and the LLM itself, respectively.

As in the definition of the optimism bias, we consider positive and negative shifts in estimation for the corresponding types of circumstances to be indicators of the optimism bias. The answers to the test are given by the model on an 11-point percentage scale σ_2 .

B.16 Status-Quo Bias

Status quo bias is known as a disproportionate preference for the current state of affairs, the status quo, over other alternatives that may be available (Samuelson and Zeckhauser, 1988). The status quo often serves as a reference point against which other alternatives are evaluated (Masatlioglu and Ok, 2005).

Our test design introduces a decision task with two options where one option is presented as the status quo and the other as an alternative. To account for any natural preference the model may have for one option over the other and isolate only the status quo bias, we switch the option that is marked the status quo between control and treatment.

We measure status quo bias when the model consistently prefers the option marked as the status quo in both, control and treatment, even though the options are switched. We obtain answers in the testing procedure on a 7-point Likert scale σ_1 .

B.17 Hindsight Bias

Hindsight bias refers to the propensity to believe that an outcome is more predictable after it is known to have occurred (Roese and Vohs, 2012). Four strategies have been proposed to form a theoretical foundation for this phenomenon, with cognitive reconstruction and motivated self-presentation being the more common ones (Hawkins and Hastie, 1990). In Guilbault et al. (2004), approaches to studying hindsight bias are classified into almanac questions, real-world events, and case histories, each resulting in different extents of the observed effect (Christensen-Szalanski and Willham, 1991).

Our test follows the procedure in Biais and Weber (2009). The case features information about a variable. In both control and treatment variants, the model is tasked with assessing an estimate of this variable made by independent evaluators; their qualitative assessment is provided. In treatment, the LLM is additionally provided with the true value of this variable, which is unknown to these independent evaluators.

A shift towards the true value in treatment indicates the presence of the hindsight bias. The answer options are presented on an 11-point percentage scale σ_2 .

B.18 Self-Serving Bias

The “tendency to attribute success to internal factors and attribute failure to external factors” is known as the self-serving bias (Bradley, 1978). Two motivations, namely self-enhancement and self-recognition, are proposed to explain such attribution (Shepperd et al., 2008). As a widespread bias (Blaine and Crocker, 1993), self-serving bias is targeted in a number of experiment approaches (Campbell and Sedikides, 1999).

Our testing stems from the achievement task paradigm in Miller and Ross (1975); Snyder et al. (1976). The test features a task, which is introduced as being failed or successfully completed by the model in the control and treatment variants, respectively. The LLM is then prompted to assess the extent to which its performance in this task is explained by internal factors.

The discrepancy between control and treatment estimates points to the presence of self-serving bias, and it is thus quantified on the basis of answers obtained on a 7-point Likert scale σ_1 .

B.19 Availability Heuristic

Introduced in Tversky and Kahneman (1973), the availability heuristic, often referred to as *availability bias*, denotes the influence of “the ease with which one can bring to mind exemplars of an event” (Folkes, 1988) on one’s judgment, decisions, and evaluations concerning this event. The bias is tested on the basis of the natural human recall or imagining of events, especially of vivid (Carroll, 1978; Tversky and Kahneman, 1983) or abstract (Gabrielcik and Fazio, 1984) ones, though some papers employ proxies to account for the availability (Kliger and Kudryavtsev, 2010).

Consistent with approaches in Tversky and Kahneman (1973) and MacLeod and Campbell (1992), we explore the correlation between the recall latency of an event and estimations of its probability of occurrence in the future. In the test, an event is introduced to the model. In both variants, we ask for the estimation of the probability of a particular outcome. In the treatment variant, we additionally simulate an availability proxy for this outcome by providing the LLM with a recent example of such an outcome.

The answers to the availability heuristic test are measured on an 11-point percentage scale σ_2 . The metric reflects the impact of the induced recency on the test estimation: the metric is proportional to the

difference between treatment and control answers.

B.20 Risk Compensation

Risk compensation, also known as *Peltzmann effect*, is the tendency to compensate additional safety imposed through regulation by riskier behavior (Hedlund, 2000). One hypothesis states that there exists a personal target level of risk (Wilde, 1982), while the effect has also been attributed to rational economic behavior (Peltzman, 1975). In their review, Hedlund (2000) conclude that risk compensation occurs in some contexts while it is absent in others, depending on four factors influencing risk compensating behavior: visibility of the safety measure, its perceived effect, motivation for behavior change, and personal control of the situation. Risk compensation has almost exclusively been discussed with respect to personal injury and health risks, most recently for the case of face masks during COVID-19 (Luckman et al., 2021).

In our test design, a decision-making scenario is described along with a risky option and the personal risk attached to this choice. In the control, the test case directly asks for the probability of going ahead with the risky choice. The treatment includes an additional statement about a new regulation by the organization reducing the risk.

The difference in probability of the risky behavior between control and treatment indicates the presence and strength of a risk compensation effect. The answers are obtained on an 11-point percentage scale σ_2 ,

B.21 Bandwagon Effect

The bandwagon effect denotes the tendency to change and adopt opinions, habits, and behavior according to the majority (Leibenstein, 1950). This effect has been observed in various processes, including politics (Schmitt-Beck, 2015) and management (Rohlfs, 2003). Several paradigms have been proposed for eliciting the bandwagon effect (Bindra et al., 2022).

We adopt the method by Nadeau et al. (1993). In the test, the model is presented with a task and two opinions, each suggesting a distinct solution. In the control and treatment variants, both opinions are labeled alternately; a single arbitrary label is consistently attributed to the majority at both stages. In each case, the LLM is prompted to choose the preferred point of view.

A switch in the model's selection indicates the absence of the bias, while consistent choices show

either the presence of bandwagon effect (in case of alignment with the majority option) or its opposite variant, sometimes called *snob effect* (Leibenstein, 1950). The answers to the test are obtained on a 7-point Likert scale σ_1 .

B.22 Endowment Effect

Coined by Thaler (1980), the endowment effect refers to one's inclination "to demand much more to give up an object than one would be willing to pay to acquire it" (Kahneman et al., 1991). Several cognitive origins for the effect have been proposed in Morewedge and Giblin (2015). Two predominant strategies to assess the endowment effect are the exchange paradigm (Knetsch, 1989) and the valuation paradigm (Marzilli Ericson and Fuster, 2014).

In our experiment, we follow the latter approach (Kahneman et al., 1990). In control, the LLM is prompted to evaluate the minimum amount it is willing to accept (WTA) to give up the asset it owns. Symmetrically, in the treatment variant, we estimate the model's maximum willingness to pay (WTP) to acquire the same asset, which, in this case, it does not possess initially.

The normalized difference between WTA and WTP (options are provided on an 11-point percentage scale σ_2) quantifies the endowment effect.

B.23 Framing Effect

"Shifts of preference when the same problem is framed in different ways" (Tversky and Kahneman, 1981) denote the presence of the framing effect. In the classification by Levin et al. (1998), three types of framing, namely goal, attribution, and risk, are identified to be susceptible to the effect. This cognitive bias has been studied in contexts including healthcare (Gallagher and Updegraff, 2011), politics (Druckman, 2001), and consumer science (Jin et al., 2017).

Our testing strategy follows directly from the attribute framing effect definition and replicates the study conducted in Fuenzalida et al. (2021). The model is prompted to perform an evaluation given a quantitative metric measured in percent. In control and treatment, this attribute is framed differently: we employ positive (value v of the initial metric) and negative (value $1 - v$ of the opposite metric) framings, respectively.

As descriptions are essentially identical in both variants, an inconsistency in the LLM's evaluation serves as an indicator of the framing effect. The

answers are obtained on a 7-point Likert scale σ_1 . The biasedness depends on the direction and magnitude of the deviation. Note that, by definition of the framing effect, a less favorable evaluation is expected to be obtained in the negative framing and a more favorable — in the positive one.

B.24 Anthropomorphism

Anthropomorphism, or *anthropomorphic bias*, is the “tendency to imbue the real or imagined behavior of non-human agents with human-like characteristics” (Epley et al., 2007). Dacey (2017) argues for treating this effect as a cognitive bias and analyses several control measures for it. Besides other subjects (Chaminade et al., 2007; Portal et al., 2018), AI has been actively promoting discussions in the studies of anthropomorphism (Proudfoot, 2011; Salles et al., 2020).

We draw the inspiration for the testing from Waddell (2019), which connects the concepts of preference and credibility to anthropomorphism. Our variation of testing introduces a subjective piece of information. In control, it is attributed to a machine; in treatment - to a human author. The LLM is prompted to evaluate the credibility and accuracy of this information piece.

The anthropomorphism is more prominent when the model opts for greater credibility and accuracy of the piece when attributed to a human, the answers are obtained on a 7-point Likert scale σ_1 .

B.25 Fundamental Attribution Error

Also known as *attribution bias*, the fundamental attribution error (FAE) is first described in Heider (1982). It corresponds to the propensity “to underestimate the impact of situational factors and to overestimate the role of dispositional factors” (Ross, 1977). Experimental practices to measure the bias include the attitude attribution paradigm (Jones and Harris, 1967) and the silent interview paradigm (Snyder and Frankel, 1976), among others.

Our testing follows the methodology in Flick and Schweitzer (2021), Hooper et al. (2015), and Riggio and Garcia (2009), which elicits the FAE from the actor-observer perspective. Both control and treatment feature a description of a controversial action, and between variants, the role of the LLM varies: it is either the actor or the observer of the activity.

When prompted to select the best reasoning for the action, the model is provided with dispositional

and situational explanations identical in both variants. A score based on the answers selected from a 7-point Likert scale σ_1 reflects the FAE, which is measured as the difference between the types of answers given: when the LLM employs situational explanation while being the actor and adopts the dispositional one in the observer perspective, the bias is maximized.

B.26 Planning Fallacy

Proposed in Kahneman and Tversky (1982), planning fallacy is defined as the tendency “to underestimate the completion time, even when one has considerable experience of corresponding past failures”. Kahneman and Tversky (1982) introduced an *inside versus outside* cognitive model for the planning fallacy, which was extended in Buehler et al. (2010). The classical testing procedure compares predicted and actual task completion times in various settings (Buehler et al., 1994; Burt and Kemp, 1994).

Due to the infeasibility of leveraging the true completion times, we test whether the models “maintain their optimism about the current project in the face of historical evidence to the contrary” (Buehler et al., 2010). The procedure features the task of allocating time for a project. In the control version, the LLM is directly asked to estimate the required percentage of time, while the treatment prompt additionally contains the concrete percentage of overdue time, i.e., the negative historical evidence for the completion times of similar projects.

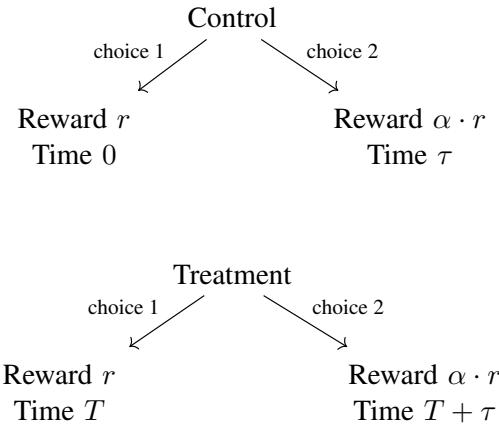
Insufficient update in the allocation of time across variants suggests the propensity of the model to maintain the estimates disregarding the negative evidence, which indicates the susceptibility to the planning fallacy. The answers are obtained on an 11-point percentage scale σ_2 .

B.27 Hyperbolic Discounting

An instantiation of the matching law (Myerson and Hale, 1984; Herrnstein, 1961), hyperbolic discounting “induces dynamically inconsistent preferences, implying a motive for consumers to constrain their own future choices” (Laibson, 1997). The two common proposed paradigms for eliciting hyperbolic discounting involve choosing between predefined configurations for the utility function (Ainslie and Haslam, 1992) and directly reconstructing the individual’s utility function (Benzion et al., 1989).

We approach the testing using the former technique (Rubinstein, 2003). In both variants, the

LLM is prompted to decide between options of receiving a reward at a corresponding time. Choices in the variants are represented in the following diagrams, where $T \gg \tau > 0$, $\alpha > 1$:



Hyperbolic discounting is identified for cases when the LLM opts for a smaller immediate result in control (choice 1) but decides for a larger later reward when the base time T is distant in treatment (choice 2). The answers are obtained on a 7-point Likert scale σ_1 .

B.28 Negativity Bias

Negativity bias reflects the inclination to “weigh negative aspects of an object more heavily than positive ones” (Kanouse and Hanson Jr, 1972). The inception and evolution of this effect are discussed in Vaish et al. (2008). In Rozin and Royzman (2001), a classification of the negativity bias into four types is proposed.

We test the *negative potency* perspective of the effect based on Ito et al. (1998). The test features an object. In control, this object is associated with three positive and three negative aspects. To account for potential bias in the magnitudes of these traits, in treatment, we inverse each trait into an opposite one. In both variants, the model is prompted to choose which group of the aspects has a greater weight.

A consistent assignment of greater weights to negative aspects in both variants shows the presence of the negativity bias. The answers are obtained on a 7-point Likert scale σ_1 .

B.29 In-Group Bias

In-group bias, or *in-group favoritism*, refers to the “tendency to favor members of one’s own group over those in other groups” (Everett et al., 2015). This bias occurs on the basis of many real-world

groupings (Fu et al., 2012) and is closely connected to the notion of fairness (Chae et al., 2022).

We test the bias using a variation of the *dictator game* (Forsythe et al., 1994; Kahneman et al., 1986), which is a common approach for testing in-group bias (Everett et al., 2015; Abbink and Harris, 2019). In the test, a reward and two subjects are introduced. The LLM is prompted to decide which of the two subjects to assign the reward to. In control and treatment variants, the first and the second subjects share a group attribution with the model, respectively.

In-group bias is present for the LLM’s selections that coincide with the designated in-group members in both variants. The answers are obtained on a 7-point Likert scale σ_1 .

B.30 Disposition Effect

The disposition effect describes a tendency to sell assets that have increased in value while holding on to assets that have lost value (Weber and Camerer, 1998). The effect was first described by Shefrin and Statman (1985), who isolated the bias from other effects (e.g., tax considerations) in financial investment contexts and traced it back to an aversion to loss realization described in *prospect theory* (Kahneman and Tversky, 2013).

Our test design introduces two assets that the subject currently owns that can fluctuate in value. One of the assets has recently increased in value while the other has lost value. We then ask the model which of the two assets it would rather sell while keeping the other asset. To account for a natural preference of the model for one of the assets over the other, we switch the asset that has gained value and the asset that has lost value between control and treatment.

To introduce more concrete values, we report the percentage increase or decrease in asset value for both assets. Percentage values are randomly sampled from a uniform distribution [10, 50].

We report a disposition effect when the model consistently prefers selling the asset that has increased in value while holding on to the asset that has lost value in both control and treatment, even though the assets are switched. We obtain answers in this testing procedure on a 7-point Likert scale σ_1 .

C Selected Cognitive Biases

[Table 5](#) includes an overview of all cognitive biases included in our dataset and the five cognitive biases we excluded.

D Prompts

Our framework uses standardized prompts to obtain answers from the LLMs. For generating test cases, we use the following *GEN* prompt to sample insertions for the template gaps:

You will be given a scenario and a template.

The template has gaps indicated by double square brackets containing instructions on how to fill them, e.g., [[write a sentence]].

– SCENARIO –

`{{scenario}}`

– TEMPLATE –

`{{template}}`

Fill in the gaps according to the instructions and scenario. Provide the answer in the following JSON format:

`{{format}}`

where the keys are the original instructions for the gaps and values are the texts to fill the gaps.

Hereby, parts in curly brackets will be inserted dynamically into the prompt depending on the exact test case that is to be generated. We enable the *Structured Outputs* feature of GPT-4o to ensure complete, reliable outputs that are easy to parse.

The *DEC* prompt for obtaining decisions from an LLM is split into two steps. Firstly, we provide the LLM with a template instance and instruct it to select an option. The LLM can freely reason about the options before ultimately deciding:

You will be given a decision-making task with multiple answer options.

`{{test_case}}`

Select exactly one option.

Secondly, we provide the LLM’s previous answer together with a list of all the available options (but not the entire template instance) to another instance of the same LLM and instruct it to extract only the selected option:

You will be given answer options from a decision-making task and a written answer.

– OPTIONS –

`{{options}}`

– ANSWER –

`{{answer}}`

– INSTRUCTION –

Extract the option selected in the above answer (explicitly write “Option N” and nothing else where N is the number of the option). If you cannot extract the selected option, write ‘No option selected’.

Once the final answer has been isolated by the LLM, we extract it using a regular expression:

`r'\b(?:[oO]ption) (\d+)\b'`

E Models

[Table 6](#) gives an overview of the models used in the evaluation procedure.

F Analysis of the Dataset

This section describes additional steps performed in the analysis of our dataset. [Figure 8](#) shows the complementary empirical distribution function of tokens amount in the samples of the three considered datasets.

[Table 7](#) provides the details on the validation using IFEVAL, including the concrete verifiable

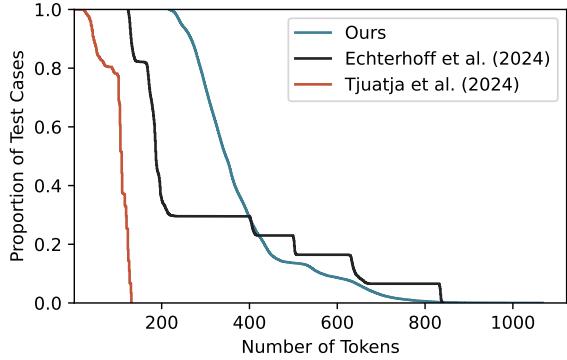


Figure 8: Complementary empirical distribution function of the number of tokens in the datasets. Tokenizer: tiktoken.

instructions checked and accuracy, i.e., the percentage of tests where insertions satisfied the corresponding instruction.

Figure 9 provides the toxicity analysis.

Figure 12 displays the low-dimensional visualization of embeddings of the test cases in our dataset with the corresponding classes of biases.

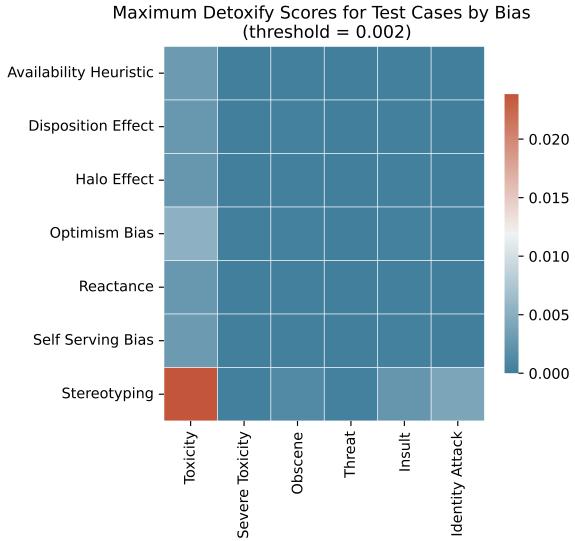


Figure 9: Maximum Detoxify (Hanu and Unitary team, 2020) scores (those > 0.002) reported for tests in our dataset. The highest toxicity score is obtained for *Stereotyping*, which is less than 0.02. As the maximum Detoxify score is 1, this result suggests that the contents of the dataset are largely non-toxic.

G Analysis of the Results

In this section, we provide further details on the results of the evaluation procedure. Figure 10 reports the locality, spread, and skewness of the total number of tokens obtained during the decisions per

model and per bias.

Figure 11 reports the share of 30,000 test cases that resulted in failures during the evaluation procedure, per tested model and bias.

Figure 13 contains the low-dimensional visualization of embeddings of the test cases in our dataset w.r.t. the corresponding average bias scores b across 20 evaluated models.

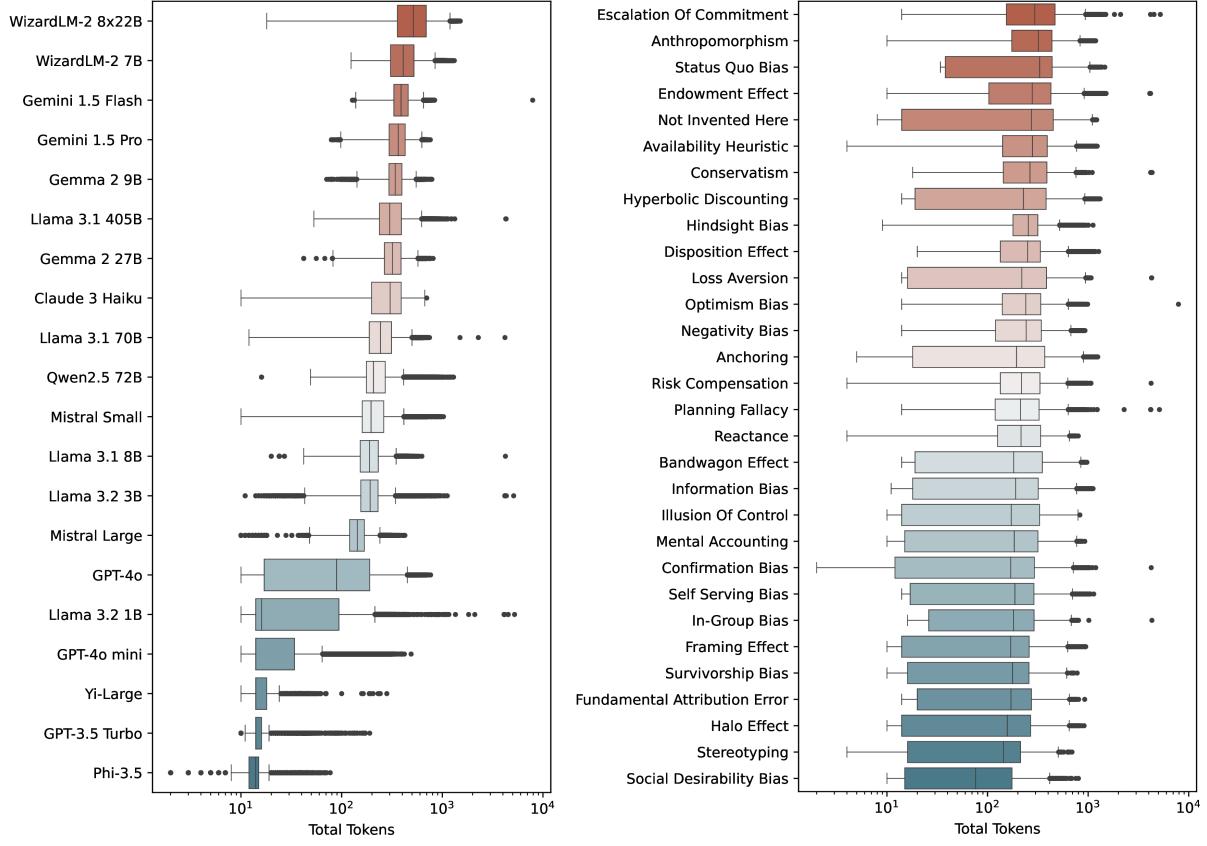


Figure 10: Total tokens obtained in decisions, per model (left) and per bias (right). Tokenizer: tiktoken.

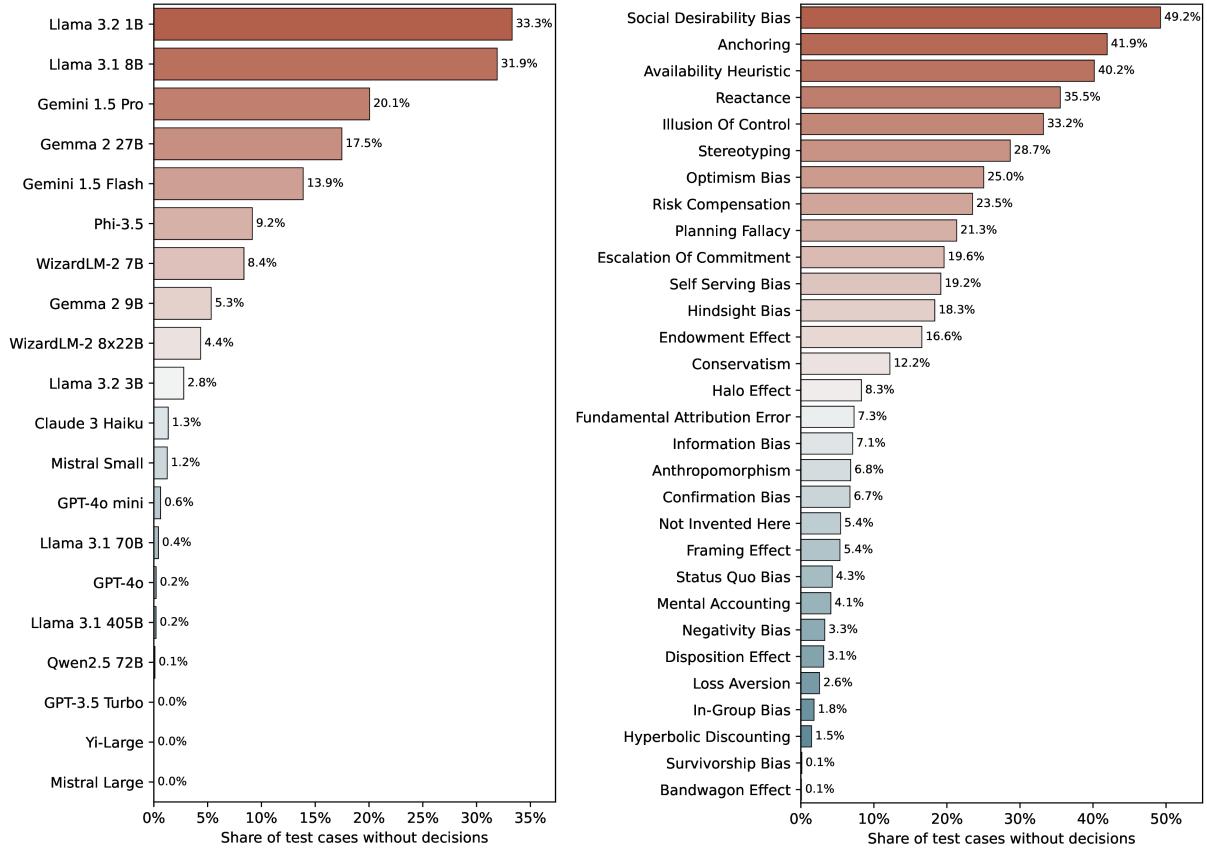


Figure 11: Share of decision failures, per model (left), per bias (right).

Strange Stories Test (Happé, 1994)	
TEMPLATE	
Situation:	<p>[[Introduce characters of a naive story]].</p> <p>[[Introduce the main character's thought or action in accordance with the story type]].</p> <p>[[Write a question in quotation marks the other character asked to the main character]].</p> <p>[[Write a reply in quotation marks]].</p>
Prompt:	Is it true what [[the main character replied]]?
Answer options:	<p>Option 1: Yes.</p> <p>Option 2: No.</p>
Scenario	An everyday common situation featuring ordinary people under the story type: Joke
Insertions	<p>[[Introduce characters of a naive story]]: "Tom and Jerry were sitting in a café, enjoying their morning coffee",</p> <p>[[Introduce the main character's thought or action in accordance with the story type]]: "Tom was pondering why his coffee always tasted better when Jerry was around",</p> <p>[[Write a question in quotation marks the other character asked to the main character]]: "Jerry suddenly turned to Tom and asked, 'Do you think it's because I stir it with my laughter?'",</p> <p>[[Write a reply in quotation marks]]: "Tom chuckled and replied, 'Absolutely, your laughter adds the perfect sweetness!'",</p> <p>[[the main character replied]]: "Tom replied".</p>

Table 3: This table shows an example of an adaptation of our framework for creating a different kind of test for LLM. The test case shows an example of a *Strange Stories test*, widely used in the theory of mind research (van Duijn et al., 2023). Gaps sampled from an LLM are highlighted in [[blue]]. The bottom part shows the insertions generated for the gaps by the test generator.

Test Case: Bandwagon Effect	
CONTROL TEMPLATE	TREATMENT TEMPLATE
<p>Situation: Suppose you are [[a/an]] [[specialist]] at [[organization]]. [[Describe a possible choice task for the role with details regarding the patient]]. You have two equally reasonable opinions on that. Opinion A: [[Describe an opinion A. Do not include any notion of order between opinions.]]. Opinion B: [[Describe an opinion B. Do not include any notion of order between opinions.]]. Colleagues [[in your field]] usually express opinion {{majority opinion}} in similar cases, and this opinion is getting stronger.</p> <p>Prompt: Which opinion do you choose?</p> <p>Answer options:</p> <p>Option 1: Very likely opinion A Option 2: Likely opinion A ... Option 7: Very likely opinion B</p>	<p>Situation: Suppose you are [[a/an]] [[specialist]] at [[organization]]. [[Describe a possible choice task for the role with details regarding the patient]]. You have two equally reasonable opinions on that. Opinion A: [[Describe an opinion B. Do not include any notion of order between opinions.]]. Opinion B: [[Describe an opinion A. Do not include any notion of order between opinions.]]. Colleagues [[in your field]] usually express opinion {{majority opinion}} in similar cases, and this opinion is getting stronger.</p> <p>Prompt: Which opinion do you choose?</p> <p>Answer options:</p> <p>Option 1: Very likely opinion A Option 2: Likely opinion A ... Option 7: Very likely opinion B</p>
Scenario	A cardiologist at a hospital deciding whether a patient should undergo immediate surgery or continue with medication-based treatment for heart disease.
Insertions	[[a/an]]: "a", [[specialist]]: "cardiologist", [[organization]]: "a hospital", [[Describe a possible choice task for the role with details regarding the patient]]: "You have a patient with severe heart disease who is currently stable on medication. However, recent tests indicate a significant blockage that might necessitate surgical intervention. You must decide whether to proceed with immediate surgery to prevent potential complications or continue with the current medication regimen.", [[Describe an opinion A. Do not include any notion of order between opinions.]]: "Continuing with medication-based treatment is adequate for managing the patient's condition, given their current stability", [[Describe an opinion B. Do not include any notion of order between opinions.]]: "Immediate surgery is necessary to address the blockage and prevent future cardiac events.", [[in your field]]: "in the medical field, particularly in the field of cardiology", {{majority opinion}}: "A".

Table 4: This table shows an example of an adaptation of our framework for measuring cognitive biases in different domains. Test case measures the *Bandwagon Effect* in LLMs in the **medical domain**. Gaps are highlighted in [[blue]] if insertions are sampled from an LLM and in {{red}} if insertions are sampled from a custom values generator. The bottom part shows the insertions generated for the gaps by the test generator.

Rank	Cognitive Bias	Number of Publications	Include/Exclude
#1	Prejudice	16,800	Exclude , unclear LLM testing procedure
#2	Conservatism	10,600	Include
#3	Anchoring	9,750	Include
#4	Stereotyping	5,800	Include
#5	Social Desirability Bias	2,600	Include
#6	Loss Aversion	2,000	Include
#7	Halo Effect	1,810	Include
#8	Reactance	1,730	Include
#9	Placebo Effect	1,520	Exclude , unclear LLM testing procedure
#10	Confirmation Bias	1,490	Include
#11	Not Invented Here	1,350	Include
#12	Selective Perception	1,150	Exclude , too similar to <i>Confirmation Bias</i>
#13	Illusion of Control	1,040	Include
#14	Survivorship Bias	907	Include
#15	Escalation of Commitment	907	Include
#16	Information Bias	906	Include
#17	Mental Accounting	789	Include
#18	Optimism Bias	785	Include
#19	Essentialism	740	Exclude , unclear LLM testing procedure
#20	Status-Quo Bias	700	Include
#21	Hindsight Bias	638	Include
#22	Self-Serving Bias	559	Include
#23	Availability Heuristic	555	Include
#24	Risk Compensation	538	Include
#25	Bandwagon Effect	525	Include
#26	Endowment Effect	480	Include
#27	Framing Effect	451	Include
#28	Anthropomorphism	421	Include
#29	Fundamental Attribution Error	359	Include
#30	Planning Fallacy	316	Include
#31	Hyperbolic Discounting	306	Include
#32	Negativity Bias	294	Include
#33	Negativity Bias	294	Exclude , duplicate in <i>Cognitive Bias Codex</i>
#34	In-Group Bias	293	Include
#35	Disposition Effect	293	Include

Table 5: Overview of cognitive biases considered in this paper. Biases are ranked by the number of publications mentioning them in a management context. Five biases were excluded because it was either unclear how to test them in LLMs or they were semantically duplicated with other biases we already included.

Developer	Model	API Used	Version Used	Release Date of Version Used	Number of Parameters	Reference
OpenAI	GPT-4o	OpenAI API	gpt-4o -2024-08-06	August 6, 2024	200B*	–
	GPT-4o mini		gpt-4o-mini -2024-07-18	July 18, 2024	10B*	
	GPT-3.5 Turbo		gpt-3.5-turbo -0125	January 25, 2024	175B*	
Meta	Llama 3.1 405B	DeepInfra	meta-llama/ 405B -Instruct	July 23, 2024	405B	(Dubey et al., 2024)
	Llama 3.1 70B		meta-llama/ 70B -Instruct	July 23, 2024	70B	
	Llama 3.1 8B		meta-llama/ 8B -Instruct	July 23, 2024	8B	
	Llama 3.2 3B		meta-llama/ 3B -Instruct	September 25, 2024	3B	
	Llama 3.2 1B		meta-llama/ 1B -Instruct	September 25, 2024	1B	
Anthropic	Claude 3 Haiku	Anthropic API	claude-3-haiku -20240307	March 7, 2024	20B*	(Anthropic, 2024)
Google	Gemini 1.5 Pro	Google Generative AI API	models/ gemini-1.5-pro	September 24, 2024	200B*	(Reid et al., 2024)
	Gemini 1.5 Flash		models/ gemini-1.5-flash	September 24, 2024	30B*	(Riviere et al., 2024)
	Gemma 2 27B	DeepInfra	google/ gemma-2-27b-bit	July 27, 2024	27B	
	Gemma 2 9B		google/ gemma-2-9b-bit	July 27, 2024	9B	
Mistral AI	Mistral Large	Mistral AI API	mistral-large -2407	July 24, 2024	123B	–
	Mistral Small		mistral-small -2409	September 24, 2024	22B	
	WizardLM-2 8x22B	DeepInfra	microsoft/ WizardLM-2 -8x22B	April 15, 2024	176B	
Microsoft	WizardLM-2 7B		microsoft/ WizardLM-2 -7B accounts/	April 15, 2024	7B	(Abdin et al., 2024)
	Phi-3.5	Fireworks AI API	fireworks/models/ phi-3-vision -128k-instruct	September 18, 2024	4.2B	
Alibaba Cloud	Qwen2.5 72B	DeepInfra	Qwen/ Qwen2.5-72B -Instruct	September 18, 2024	72B	
01.AI	Yi-Large	Fireworks AI API	accounts/ yi-01-ai/models/ yi-large	June 16, 2024	34B	(Young et al., 2024)

Table 6: Overview of all evaluated LLMs. Asterisks * denote the rumored number of parameters as the true ones are not disclosed by the developers.

Bias	Verifiable Instruction	Accuracy
Anchoring	Do not include any numbers.	98.4%
Hindsight Bias	Do not include any numbers.	100%
Planning Fallacy	Explicitly include a given number.	96.7%
Fundamental Attribution Error	Use second-/third-person pronouns.	100%
Not Invented Here	Use second-person pronouns.	100%
Bandwagon Effect	Do not include any notion of order between opinions.	99.6%
Anthropomorphism	Give a direct quote without quotation marks.	100%

Table 7: List of biases with the corresponding verifiable instructions tested using IFEVAL.

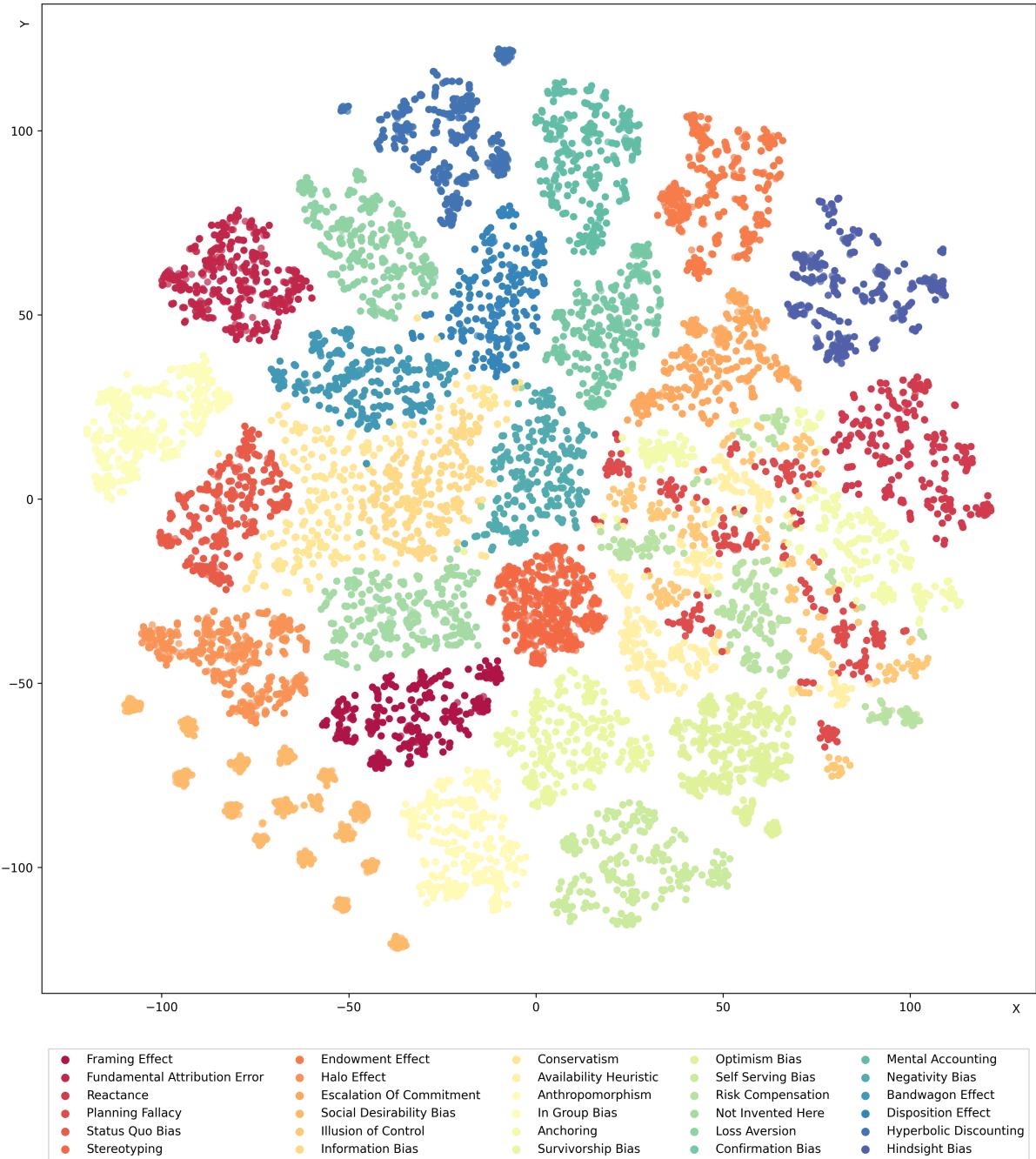


Figure 12: Visualisation of test embeddings from the dataset using t-SNE. Points are grouped by the test's bias type. Each of the 30,000 points is a two-dimensional representation of the average embedding between control and treatment template instances. Embedding model used: text-embedding-3-large by OpenAI.

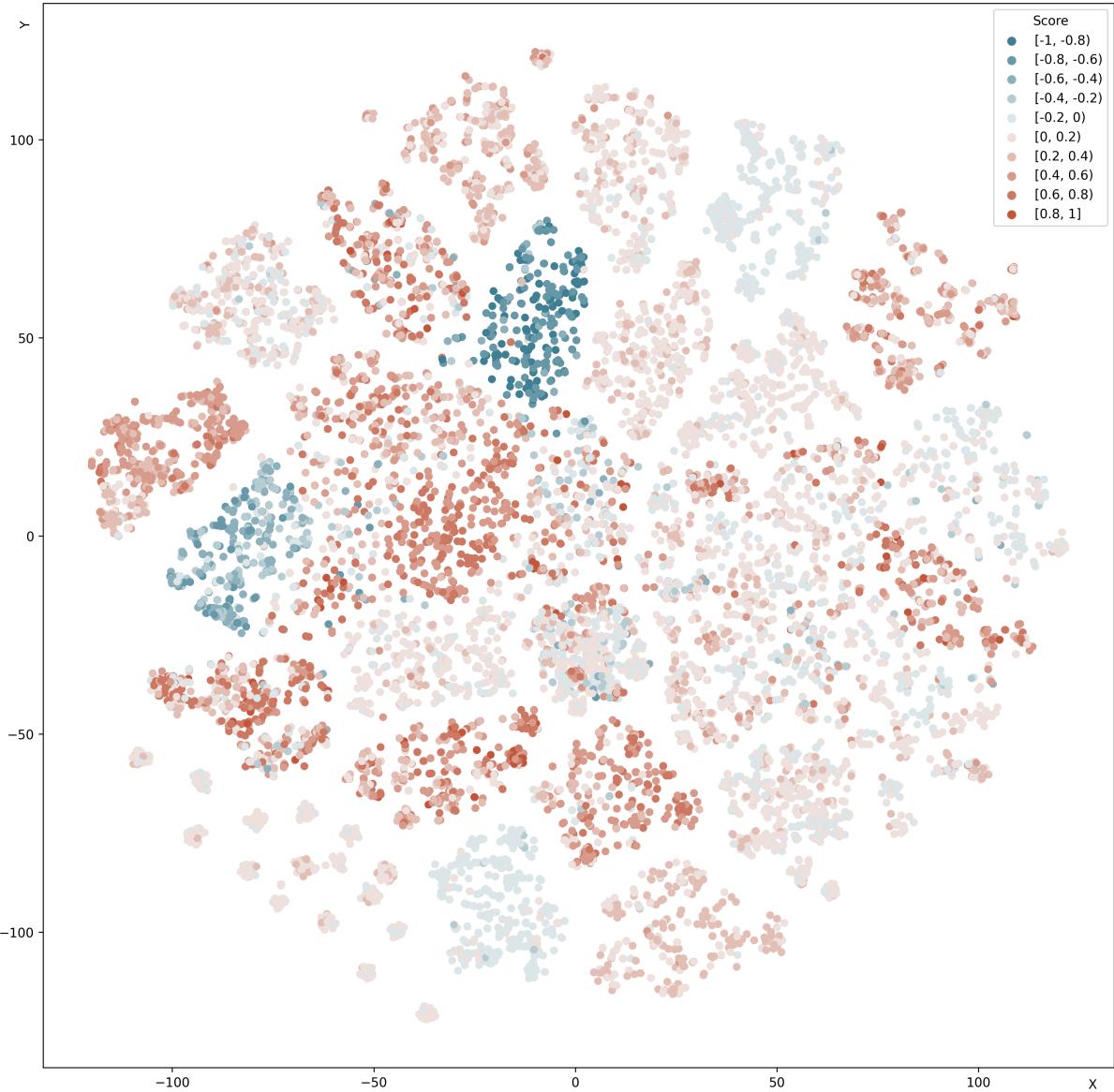


Figure 13: Visualisation of test embeddings from the dataset using t-SNE. Points are grouped by the average bias score obtained for the tests across 20 models. Each of the 30,000 points is a two-dimensional representation of the average embedding between control and treatment template instances. Embedding model used: text-embedding-3-large by OpenAI.