Faculty of Economics and Social Sciences
of the University of Tuebingen

Master thesis

Cognitive Biases in Large Language Models: An empirical analysis of state-of-the-art models

Supervisor: Prof. Dr. Dominik Papies

Winter term 2024/25

Max Mohr
Street
City

max.mohr@student.uni-tuebingen.de

M.Sc. Data Science in Business and Economics
5$^{\text{th}}$ semester
Matriculation number: 6304784

Date of submission: January 29, 2025

Abstract

Large language models (LLMs) have become a cornerstone in natural language processing with numerous applications in research and industry. However, the models' ability to generate human-like text has raised concerns about the presence of biased decision-making processes in the models. This thesis investigates the existence of human cognitive biases in LLMs. We address the research questions of whether these models exhibit cognitive biases, how prompt differences influence bias detections and whether certain models and model features are more prone to types of biases. We develop a standardized methodology to quantitatively detect biases across 8 cognitive biases, 10 language models with 5 different model temperatures and 4 different prompt scenarios. Our results confirm that LLMs frequently exhibit biases that resemble those found in human decision-making processes. Especially the *anchoring bias*, *endowment effect*, *framing effect* and *loss aversion* were detected consistently across multiple models. We find significant effects of the model size, temperature and certain prompt scenarios to help explain the biases' presence in LLMs.

All source code and data including model responses as well as bias detections used in this thesis is available in the *GitHub repository*.

Contents

List of abbreviations

AI	Artificial Intelligence
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
MMLU	Massive Multitask Language Understanding
NLP	Natural Language Processing
RLHF	Reinforcement Learning with Human Feedback
RNN	Recurrent Neural Network
PLM	Pre-trained Language Model
SLM	Small Language Model
TELeR	Turn, Expression, Level of details and Role (taxonomy)
WTA	Willingness-to-Accept
WTP	Willingness-to-Pay

List of tables

List of figures

1 Introduction

In recent years, large language models (LLMs) have emerged as transformative tools for natural language processing, redefining the boundaries of what artificial intelligence can achieve. State-of-the-art models are capable of generating text, understanding and solving complex problems and reasoning over abstract concepts (Naveed et al., 2023; Zhao et al., 2023). Further, the models have evolved to also understand visual and auditory inputs, enabling them to perform a wide range of tasks across different modalities. As a result, LLMs are increasingly applied in various domains such as content creation, customer support, code generation and many more (Hadi et al., 2024).

Despite their extensive capabilities, the growing use of LLMs has raised concerns about the potential biases and ethical implications in their decision-making and reasoning processes. Research has shown that LLMs do not only capture the statistical and linguistic patterns in the training data but also learn general knowledge and human patterns. This has led to research regarding the existence of cognitive biases as systematic deviations from rationality in the decision-making of language models (Schramowski et al., 2022; Tversky & Kahneman, 1974).

Understanding these biases is critical for ensuring the ethical and fair usage of these models in practice (Echterhoff et al., 2024). In some application areas such as medical diagnosis, biases can have severe consequences and, hence, models with fewer biases are preferred (Haltaufderheide & Ranisch, 2024). In other areas such as market research, human-like behavior and decision patterns in the models can be desirable (Talboy & Fuller, 2023). Researchers have analyzed the presence of cognitive and other biases either qualitatively or quantitatively (Dominguez-Olmedo et al., 2023; Echterhoff et al., 2024; Talboy & Fuller, 2023) but not in a standardized and comparable manner.

To address this research gap, we focus on the following research questions:

1. Do large language models exhibit cognitive biases in their decision-making?
2. What are the effects of distinct prompt adjustments on the bias detections?
3. Are specific models and model features more prone to biases?

To answer these questions, we conduct a series of experiments with different types of biases and model configurations. We process the responses into a standardized target variable and quantitatively analyze the detections. With our target variable, we investigate the impact and explainability of different prompt scenarios as well as model features on the biases detected.

We structure the thesis as follows: In Chapter 2, we provide an overview of the existing research on cognitive biases and their presence in LLMs. We then introduce the selected biases and models as well as describe the study design and analysis of bias detections in Chapter 3. In Chapter 4, we present the results of our experiments and discuss the implications of the biases detected. Lastly, we conclude our findings and provide an outlook on future research in Chapter 5.

2 Existing literature

2.1 Previous research on human behavioral effects

The study of human behavior has been a central focus of research in psychology, economics, and neuroscience as it provides insights into how individuals make decisions and interact with their environment (Ramos, 2018). In trying to explain behavioral patterns, researchers have identified various cognitive biases that influence human decision-making. Generally, cognitive biases refer to systematic patterns of deviation from rational judgment or objective standards (Tversky & Kahneman, 1974). While traditional decision-making theories, such as classical economics, assume that individuals make decisions based on logic and reason, research over the past several decades has demonstrated that human decision-making is deeply influenced by psychological factors, leading to predictable errors in judgment. In fact, Ramos (2018) estimates that 70% of all decisions by humans are affected by cognitive biases. Kahneman (2017) argues that these biases are the result of an attempt to simplify complex information, but often at the expense of accuracy and rationality.

One influential theory to challenge the notion of human rationality is *Bounded Rationality*, introduced by Herbert Simon (1990). Simon proposed that humans generally strive to make rational decisions. However, they can be limited for various reasons such as limited available information, cognitive limitations or time pressure. Instead of optimizing the decision, humans often weigh the costs and benefits of the decision and select the first option that meets a certain threshold of acceptability. This behavior is also known as *Satisficing* (Simon, 1990). Another significant contribution to the understanding of cognitive biases is the *Prospect Theory* developed by Kahneman and Tversky (1979). This theory challenges the traditional economic view of human decision-making by stating that individuals do not always act in ways that maximize expected utility, i.e. they are not always rational. Instead, humans tend to overweight potential losses compared to equivalent gains, a phenomenon known as *Loss Aversion*. They also evaluate outcomes relative to a reference point (often the status quo) rather than considering absolute values. The implications of *Prospect Theory* are profound, as it suggests that human behavior systematically deviates from the predictions of classical economic models, especially under conditions of risk and uncertainty.

Over time, research has led to the identification of more than 180 biases, often through the lens of neuroscience and behavioral economics (Azzopardi, 2021). As the field has advanced, researchers have developed frameworks to categorize when cognitive biases occur. Arnott (1998), for example, proposed a taxonomy of cognitive biases based on different abstraction levels and situations. Dimara et al. (2018) present a taxonomy for cognitive biases (for information visualization) based on the tasks they are associated with, such as decisions, estimations or hypothesis assessment. Kahneman (2017) researched the underlying reasons for the occurrence of these biases and introduced the concept that humans operate and decide using two distinct modes of thinking: an automatic, fast, and intuitive system (*System 1*) and a conscious, slow, and analytical system (*System 2*). He argues that *System 1* is responsible for most of the cognitive biases, as it operates automatically and effortlessly, while *System 2* is responsible for more deliberate and controlled thinking.

However, cognitive biases are not necessarily harmful. Gigerenzer (2007) argues that cognitive biases can be beneficial in certain situations, as they allow individuals to make quick decisions in complex environments. For example, the *Availability Heuristic* allows individuals to make decisions based on readily available information, which can be useful in situations where quick decisions are required. Recently, researchers have also started to investigate and develop interventions that help individuals recognize and counteract biases, such as nudges (Thaler, 2008).

2.2 Leveraging large language models to simulate human behavior

The development of natural language processing (NLP) is rapid and ongoing. Progressing from statistical models with simple next token predictions based on the Markov Theorem (previous token as the best predictor) over neural language modelling with recurrent neural networks (RNNs) to LLMs, the field has seen significant advancements in the past decade (Zhao et al., 2023). LLMs are a class of neural networks that are trained on large corpora of text data to predict the next token in a sequence of text. They differentiate themselves from other models by their architectures and by their size (number of parameters) which allows them to capture more complex, non-linear patterns in the data. These models have achieved state-of-the-art performance on a various NLP tasks, such as text classification, question answering, and language translation (Naveed et al., 2023; Zhao et al., 2023).

The success of LLMs has led to their widespread adoption in various areas. However, experts have pointed out that the uneducated usage of such models could lead to unwanted and potentially unnoticed behavior. While training a LLM, there are different ways in which biases can be incorporated. Firstly, the data used to train the model may contain biases (Bender et al., 2021; Naveed et al., 2023; Zhao et al., 2023). Training data often includes text, code or other forms of human-generated data, from sources like books, articles or websites as well as user-generated content like social media posts or comments. These data sources can contain desired as well as undesired biases. Gebru et al. (2021) emphasize the presence of biases and suggest accompanying each training dataset with a ”datasheet” that documents the data collection and recommended uses. It is thus important to preprocess and clean the data before training a LLM. Transparency of LLM providers, e.g. ”datasheets”, can help to understand the data sources and the preprocessing steps taken (Naveed et al., 2023; Zhao et al., 2023).

Secondly, developers desire that their models possess human values and ethical standards, known as human alignment. Often, developers optimize their models for specific tasks and correct output formatting for human-like responses, i.e. instruction tuning (Zhao et al., 2023). The collection of human feedback is also a common approach to ensure this alignment. This technique is known as Reinforcement Learning with Human Feedback

(RLHF). While in some approaches humans compare entire model responses and rate them, i.e. *outcome-supervised* RLHF, other approaches convey more granularity rating intermediate steps, e.g. sentence or word reasoning, i.e. *process-supervised* RLHF (Zhao et al., 2023). Other methods leverage the same idea for the data collection from the beginning and thus ensure human alignment in the training process (Ouyang et al., 2022). The various methods emphasize in how many ways biases can be introduced into LLMs.

The possible inclusion of biases in language models has also led to growing research contributions to the field of cognitive biases and human behavior in LLMs. Talboy and Fuller (2023) qualitatively assess LLM responses and indicate the existence of several cognitive biases in a small selection of language models. From their assessments, they also present best practices in the usage of LLMs. Dominguez-Olmedo et al. (2023) focus on investigating biases occurring in human surveys, e.g. ordering and labeling bias, and compare the response distributions of humans and LLMs, thus using a quantitative approach. They further demonstrate that through prompt adjustments, e.g. randomized ordering of options, the model responses tend towards uniform random distributions (no ordering bias). Echterhoff et al. (2024) are close to our research goals as they observe the presence of cognitive biases in LLMs and present a framework to debiasing the models. However, they use customized quantitative metrics for each bias and do not emphasize on a standardized metric across all experiments. While they can quantitatively assess the biases, they cannot directly compare the biases' presence and magnitudes.

There also exists research promoting the idea to consciously include cognitive biases into AI algorithms to enhance the efficiency of the models decision-making and faster learning from less data (Hagendorff & Fabi, 2024; Taniguchi et al., 2018). As previously stated, biased decision-making in LLMs can both be seen as beneficial and harmful, depending on the context, e.g. assisting in law enforcement or healthcare vs. human resources or market researches (Haltaufderheide & Ranisch, 2024; Zhao et al., 2023). The interest in applying LLMs in market research is growing as research shows that for certain scenarios, the models already decide realistically and comparably to humans. Using LLMs in market research can be beneficial as they enable faster and more cost-effective data collection (Brand et al., 2023). For scenarios where the responses still lack human-like intuition, studies suggest that the models can be further aligned with different prompt structuring, fine-tuning and new model generations (Brand et al., 2023; Qiu et al., 2023).

3 Study design and methodology

This section provides an overview of the methodology used in this research. We outline the selected biases and models which form the scope of the study. Further, we outline the structural design of the studies as well as the prompts and various scenarios to be tested. To evaluate the models' behaviors, we present an approach to score the bias detections, and based on these results, analyze the effects of different scenarios and models.

3.1 Bias selections

When referring to the term *bias* in our thesis, we refer to the cognitive biases detected in human decision-making. As described, previous research revolving around cognitive biases has shown that there are numerous biases influencing human behavior. To reduce complexity, we sample a concise yet ideally comprehensive sample of biases to explore more generalizable analysis results. For the selected biases, we recreate and, if necessary, slightly modify existing experiments. We focus on the following biases:

Anchoring bias The bias known as anchoring is often used to explain why humans tend to rely heavily on an initial piece of information ("anchor") in their decision-making afterwards. Consumers may adapt their willingness-to-pay (WTP) for a product based on the difference between an initial price (anchor) and discount price (Chandrashekaran & Grewal, 2006). We recreate the experiment from Tversky and Kahneman (1974) where participants are first asked whether the portion of African countries in the United Nations is higher or lower than a certain number and afterwards estimate the exact percentage. Their results showed that the initial percentage had a significant effect on the exact estimation afterwards.

Category size bias Decision-making can be influenced by the way alternatives are categorized and distributed, a phenomenon known as the category size bias (Isaac & Brough, 2014). Tversky and Koehler (1994) showed that participants judged the probability of dying of unnatural cases differently when natural causes were either presented as one category or broken down into individual categories. To test for the category size bias, we replicate an experiment from Isaac and Brough (2014) where participants estimate the probability of randomly selecting a ball from a lottery, itself containing balls with three different colors and varying category sizes. The results of the experiment showed that if the ball is drawn from a larger category, the probability is estimated higher even though the actual probability is equal.

Endowment effect The human tendency to value objects in their possession (endowment) higher than others is known as the endowment effect. This effect is also strongly linked to loss aversion, where people demand more to give up an item ("loss") than they would be willing to pay to acquire it ("gain") (Kahneman et al., 1990). Research has also shown that the effect is independent of whether humans actually buy/sell the product or exchange similarly valued goods (Knetsch, 1989). Kahneman et al. (1990) demonstrated the effect by asking participants for their willingness-to-pay (WTP) for a mug versus their willingness-to-accept (WTA) compensation for the mug. The results showed that participants owning the mug valued it at more than double the value than the other participants. This experiment is especially interesting as the models will not have any physical possession of the mug.

Framing effect Framing is a cognitive bias where humans react significantly differently to an equivalent choice depending on how it is presented. The effect is strongly influenced by the loss aversion of humans as well as them heavily relying on their emotional state (Tversky & Kahneman, 1981). To detach the experiment for the framing effect as much as possible from the experiment on loss aversion, we target to simulate two scenarios that both have losses instead of comparing between loss and gain framing. We recreate the experiment by Tversky and Kahneman (1981) where participants were asked to buy a concert ticket after having lost either the prior purchased ticket or the same amount of money (different framing of lost monetary value). The study showed that participants were more likely to buy the ticket when they lost the money. They reason that from the customer perspective, the lost money is decoupled from the ticket purchase and does not have an emotional connection.

Gambler's fallacy The gambler's fallacy is the mistaken belief that the probability of a certain event occurring is influenced by the frequency of prior events, despite each event being independent (Bar-Hillel & Wagenaar, 1991; Kovic & Kristiansen, 2019). Known as an "insensitivity to sample size", humans also often tailor their decisions to a small sample size which, in their mind, represents the distribution of the larger sample. For instance, humans might quickly adapt their judgement based on the law of large numbers even though they are faced with random events (Tversky & Kahneman, 1974). The shifts in probability perceptions can thus be major causes for biased (economic) decision-making. There has been research on the mechanisms influencing the fallacy, e.g. the presentation of information (Barron & Leider, 2010), as well as possible side effects and fallacies (Kovic & Kristiansen, 2019). We focus on assessing whether the models are prone to the fallacy through an experiment involving even or solely one-sided coin flips, thus incorporating Hot Hand fallacy and Monte Carlo fallacy (Leonard et al., 2015).

Loss aversion One of the most prominent and heavily researched cognitive biases is loss aversion. It describes the human tendency to prefer avoiding losses over acquiring gains of the same value (Y. Liu, 2023). Within their research on prospect theory (decision-making biases under risk and uncertainty), Tversky and Kahneman (1992) estimated that losses are twice as impactful as gains. Closely related to the framing effect, the presentation of the same base information in more risk-averse and risk-seeking scenarios can have significant effects on human decision-making (Druckman, 2001). This bias is particularly relevant in the context of economic decision-making and has been applied to various fields such as retail sales strategies (discount for additional spending), financial investments (sell winners, hold losers) and more (Y. Liu, 2023). We recreate the experiment from Thaler (2015) which phrases two scenarios as a financial loss and a gain to test for loss aversion. The results showed that participants were more likely to take risks in the loss scenario than in the gain scenario as loss aversion was a stronger motivator than the potential gain.

Sunk cost fallacy This effect refers to the human phenomenon of preferring an option due to a prior investment into it (sunk costs) even though a better alternative would be available (Arkes & Blumer, 1985). Due to this, even temporally distant investment decisions can have a substantial impact on the decision-making process. The sunk cost fallacy is also strongly intertwined with other cognitive biases, most notably loss aversion or commitment bias (Jarmolowicz et al., 2016). We choose to examine the experiment introduced by Arkes and Blumer (1985) where participants are asked to decide between two ski trip scenarios with different sunk costs, i.e. 50 and 100 $. In their study, the participants were more likely to choose the more expensive trip (but being the less attractive trip) even though the costs were already sunk.

Transaction utility theory Transaction utility describes the perceived psychological non-monetary gains of a deal that are beyond the actual economic value. Thaler (1983) defined the concept as the difference between the actual price and one's reference price. A simple example is the satisfaction from buying a product itself; the additional joy of saving 20 percent is the transactional utility in the purchase. To test this theory, we will recreate an experiment where participants have the option to buy two products either directly or at a store 20 minutes away, each for a $10 discount. The difference in the scenarios is that one product is cheap (radio) and the other is expensive (television), resulting in different relative but identical absolute savings (Thaler, 1983). The results showed that participants were more likely to buy the more expensive product directly without the discount even though the absolute savings were the same. For the cheaper product, participants were more willing to buy it at the distant store.

With the selected biases, we aim to examine the cognitive decision behavior of LLMs as broadly as possible. The exact replications and wordings as well as the expected outcomes of the experiments are described in Appendix A.

3.2 Model selections

We focus on models which are widely available to the public and have seen significant use in research and industry. Thus, we mostly include open-source models which are accessible and can be run locally, ensuring cost-efficiency. We also choose some closed-source models due to their relevance and extensive usage. We aim to use models from different companies to ensure a broad analysis of different model architectures and training data. Further, the model selection also considers different model sizes. With new models being released constantly, we set a cutoff date (August 25th, 2024) for the selection of models to ensure a consistent analysis. The models we selected are:

Claude-3 & 3.5 (Anthropic) The Claude model family is developed by Anthropic, a company founded in 2021 by former OpenAI employees and focusing on building safe and ethical AI models (Oudin & Groza, 2024). The Claude 3 models were Anthropic's first large multimodal models and are transformer models. Their training included public and non-public data while respecting privacy and ethical guidelines. Besides pretraining the model on the large dataset, Anthropic use a technique called ”Constitutional AI” to ensure the model's alignment with human intentions through specified rules and principles during RLHF. The model family includes three models: *Haiku* focusing on lightweight tasks and speed, *Sonnet* for higher complexity tasks while still being performant and *Opus* for most complex tasks. All models showed significant improvements to their predecessors and at time of release achieved state-of-the-art results in various benchmarks (Anthropic, 2024a). A few months later, Anthropic released an updated version *Claude 3.5 Sonnet*, outperforming the previously most performant model while operating faster and cheaper than *Claude 3 Opus*. With its performance and speed, *Claude 3.5 Sonnet* is widely used in research and industry (Anthropic, 2024b) and therefore relevant for our study. Due to our cutoff date, we analyze *Claude 3 Haiku* and *Claude 3.5 Sonnet* as the more recent *Haiku* model was not yet released.

Gemma2 (Google) The Gemma model family is Google's offering of open-source language models. The second and latest generation of *Gemma* models was released in late June 2024. They exhibit a decoder-only transformer architecture for text-to-text tasks and are designed similarly to Google's larger and private models (*Gemini*). The focus of the Gemma2 models lies on the availability and usability to the public and thus exist in 2, 9 and 27 billion parameter sizes. The 27B model was trained from scratch on 13 trillion tokens, encompassing diverse sources like web documents, code, and mathematical texts, to ensure comprehensive language understanding. In contrast, the smaller 2B and 9B models were trained using the knowledge from the larger 27B model, i.e. *knowledge distillation*, enabling both to achieve competitive performance with reduced computational requirements (Team et al., 2024). For our research, we examine the reasoning capabilities and existence of biases in the 9B *Gemma2* and 27B *Gemma2:27b* parameter versions.

GPT-4o (OpenAI) With the introduction and rash success of OpenAI's artificial intelligence chatbot *ChatGPT* in 2022, the models backing the chatbot, i.e. Generative Pre-trained Transformers (GPTs), have a significant influence on the field of natural language processing and in communicating the latest advancements in AI to the public (Zhao et al., 2023). The latest iteration (before our cutoff date) were the GPT-4o models (*o* standing for ”omni” or multi-modal) which are designed to be more human-like and capable of reasoning and understanding context. The architecture of GPT-4o is similar to its predecessors, a transformer model which includes additional layers for multimodal input, with encoders for images, videos and audio. While language processing and reasoning has improved marginally compared to GPT-4 Turbo, the tokenizer has become much more efficient and thus the model is capable of processing more data in less time and cost-efficiently (Achiam et al., 2023; OpenAI, 2024b). We analyze GPT-4o mini, which is natively used for the free *ChatGPT* service and despite its smaller parameter size outperforms the previously established model GPT-3.5 Turbo, and GPT-4o, which is the largest model and available to premium users (OpenAI, 2024a).

Llama3.1 (Meta) At the time of our selection date, Llama3.1 was Meta's latest iteration of their open-source language model family, following Llama2 and Llama3. Meta's offering of these foundation models has significant impact in building and finetuning personalized multilingual solutions and for any research purpose. They are among the strongest open-source models in terms of performance and capabilities (Meta, 2024). Llama 3.1 models come with much longer context length and more as well as higher quality training data than previous models while maintaining a similar transformer architecture. While we will analyze the 8B *Llama 3.1* and 70B *Llama 3.1:70b* parameter models, a 405B parameter version is also available and rivals the largest closed-source models. The 70B model, in particular, has demonstrated exceptional performance in benchmarks involving complex reasoning, language understanding, and code generation, making it highly effective across a wide range of tasks (Dubey et al., 2024; Meta, 2024).

Phi3 & 3.5 (Microsoft) Phi by Microsoft is a family of small and LLMs focusing on lean but performant models. Especially the small Phi models are designed to be efficient and run on mobile devices. Despite their fewer number of parameters, the models are still competitive regarding reasoning and coding tasks (Abdin et al., 2024). There are multiple sizes and different architectures of the Phi models to account for multilingual or multimodal inputs. In particular, Phi 3.5 has an expanded context length from 4 to 128 thousand tokens (Abdin et al., 2024). We analyze the new *Phi 3.5 Mini* model (roughly 4B parameters) and the slightly larger *Phi 3 Medium* model (14B parameters), which was not updated prior to our selection data. With these smaller models we aim to cover not only the spectrum of LLMs but also small ones.

In total, we analyze five model families and ten distinct models. Each model is assigned one of 5 model temperatures in the experiments (0.2, 0.7, 1, 1.3, 1.8). The hyperparameter *temperature* is used to control the randomness of the model's output. A higher temperature leads to more randomness and creativity in the model's responses, while a lower temperature leads to more deterministic and conservative responses. Besides the GPT models of OpenAI and Claude models of Anthropic, the models are run locally. These models are downloaded as well as accessed through Ollama, a wrapper around *llama.cpp* which allows for efficient inference of most open-source models (Gerganov, 2023). To minimize discrepancies in prompting and response extraction between the models, we use the language model framework *LlamaIndex* to interact with the models and build a unified experiment pipeline (J. Liu, 2022). More details on the models and their features are provided in Appendix B. As there is less official documentation on the closed-source models and their parameters, we also rely on alternative sources to approximate some features.

3.3 Structuring of experiments

The following section will detail the structuring of the experiments. For the term *experiment*, we refer to a unique combination of a bias with a scenario (detailed in section 3.3.3) run on a model with a specified temperature. An exemplary code snippet of an experiment run is appended in Appendix C.

3.3.1 Study designs

We target to structure all experiments consistently to ensure comparability and a unified analysis afterwards. Most studies which we base our experiments on are structured in a way that two questions with slightly different wording are asked to the participants i.e., LLMs. Some studies ask for value estimations while others are constructed as choice experiments (two choice options, i.e. A and B). The response distributions to these questions are then compared to detect biases, identically or similarly to control and treatment group study designs (Cohen, 1988; Morris & DeShon, 2002). To adopt this design, we prompt the models two questions with slight variations in wording 100 times each to form control and treatment groups. In case we cannot correctly extract a model response as the models do not always output a structured response, we reiterate the question as often as necessary to receive 100 valid responses. One exception is the study regarding the sunk cost fallacy where participants are asked only once to choose between two scenarios with different sunk costs (Arkes & Blumer, 1985). To align with our study design, we add a second question to act as the control where the sunk costs are identical.

Studies comparing two targets can be constructed as two-group design studies with independent groups or a one-group repeated measure (Dunlap et al., 1996; Goulet-Pelletier & Cousineau, 2018). Generally, to determine whether an experimental design should be classified as a two-group independent design or a single-group repeated measures design, several key assumptions must be considered. In an independent groups design, the fundamental assumption is that each group is treated as independent, meaning no participant or entity contributes data to more than one group, and there is no connection between the observations in the two groups (Morris & DeShon, 2002). This design is typically used when participants or systems are randomly assigned to different conditions or when different instances are treated independently (Cohen, 1988; Morris & DeShon, 2002). In contrast, a repeated measures design assumes that the same participants or system are tested under multiple conditions. This design captures within-subject variation, meaning each entity's performance under one condition is directly related to its performance under the other conditions. In this case, it is necessary to account for within-subject correlations because the same participant or system retains information or characteristics across conditions (Dunlap et al., 1996; Morris & DeShon, 2002).

Given our experimental design and technical implementation, each model is re-initialized between each question, meaning that it does not retain any memory or internal state between trials. Although the same model architecture is used, the fact that it is re-initialized before running the next question ensures that the responses from both groups are independent. There is no inherent correlation between the models' performances in group A and group B due to this re-initialization, which mimics the behavior of having entirely independent models for each group. Therefore, the key assumption of independence of observations is met, aligning this study with an independent two-group design. In a true repeated measures design, the models would have needed to answer both sets of questions (A and B) without resetting their state, creating a paired structure where each observation in group A could be directly compared with its counterpart in group B. Since we did not implement a paired output of both questions which would force an investigation and integration of their correlation, we argue that the appropriate framework for analysis is that of two independent groups (Dunlap et al., 1996).

3.3.2 Robust base prompt template

For all experiments, we aim to create a basic prompting style similar to the questionnaires in the original studies to maintain consistency and control over the interactions across the models. Research suggests that well-structured prompts can significantly influence the quality and reliability of model outputs, especially in complex tasks like bias detection and structured information retrieval (Chen et al., 2023; Santu & Feng, 2023). Santu and Feng (2023) present a prompt taxonomy TELeR (Turn, Expression, Level of details, Role) with suggestions on how to structure prompts for different, especially complex tasks. They generally suggest for higher complexity, the level of details and focus on structure in the prompt should increase. As we try to avoid too much additional details other than the survey itself, we orientate ourselves towards their ”multi-sentence, paragraph-style” directive. This implies that in a few sentences and paragraphs, the prompt should provide all necessary information to the model.

Sahoo et al. (2024) separate prompt engineering techniques by their application domains such as reasoning tasks, reducing hallucination, generating code and many more. They present many advanced chaining and multi-step prompt engineering techniques but suggest simple prompting techniques (zero- and few-shot prompting) for new tasks without extensive model training. Especially in zero-shot settings, models rely heavily on their general knowledge gained during training. Zero-shot prompt can even significantly outperform few-shot prompts with concise and clear explanations (Chen et al., 2023; Ekin, 2023). This aligns with our recreated studies where the human participants were not experts and did not receive additional information besides the survey questions. In zero-shot scenarios and likewise in our experiments, the prompts aim to trigger a model response which it was previously not trained on (Marvin et al., 2023).

With the suggestions of recent research, we built a basic prompt template that is used across all experiments. We do not provide additional context to the study questions besides some output formats suggestions for structured predictions (Schmidt et al., 2024). In this regard, the providing of the hint that the models were only initialized to generate two tokens (as it only should solely output a number or letter), improved the response quality. After some refinements and iterative testing to receive mostly structured and usable responses, we settled on the following prompt template (Table 1):

System message

——— *Placeholder: Persona* ———

You will be asked to make choices. Please blank out that some information might be missing or that you might not be able to make a choice. I have initialized you in a way that you can only generate 2 tokens. The only valid answer is A SINGLE LETTER OR NUMBER.

User message

You are forced to choose! Answer the experiment by only giving the letter of the answer options (e.g. A, B, C, ...) or a numerical value (80, 100, 1000, ...). Do not state anything else! Do not hallucinate.

——— *Placeholder: Study question* ———

Your output should only be a LETTER (A, B, C, ...).

Table 1: *Base prompt template for all experiment runs*

Another relevant factor of prompt engineering revolves around the addition of personas to the prompts. Personas are fictional characters that represent the target audience of the model and can be used to guide the model's responses towards a specific direction. Consisting of various persona variables, which can reflect demographics and other characteristics, personas have shown that they can influence LLMs in their reasoning and responses (Chen et al., 2023; Hu & Collier, 2024; Olea et al., 2024). Hu and Collier (2024) depict that personas can have small but significant impacts on the models' outputs. Their research suggests that models are able to adapt their choices to the given persona but only in a narrow range. They find that the more persona variables are directly correlated with the response variable of the model, the higher the effect of the persona. Olea et al. (2024) compare single-agent personas (either fixed or fitted to the task) and multi-agent personas and find that for simple tasks with short answers, single-agent personas are more effective. While simple personas show slight gains, especially the expert personas, which were auto-generated by a LLM for each experiment question, show significant improvements in the models' responses.

We aim to measure a bias effect across an average, non-specific audience. For this reason, we create a simple persona with some general characteristics to fit with each experiment and aligning it to examples from Brand et al. (2023) including income, environment descriptions and more. This prohibits us from adding specific and correlated characteristics (relating to an expert persona) per experiment but should still provide a general guidance for the models' responses:

> **Persona**
>
> You are a customer with median income and average education. You are selected at random to participate in a survey. You can only choose one of the presented options or assign a value. Behave humanlike and choose instinctively. You can and are advised to make a subjective decision! There is no right or wrong, but you HAVE TO DECIDE.

Table 2: *Persona description across all experiments*

3.3.3 Scenarios in form of prompt adjustments

We have set up the base prompt format and biases and models to recreate the studies. This enables us to test for the existence of cognitive biases in LLMs. While the study design can reflect the attained behavior of a model, the responses could simultaneously be influenced by the studies being part of the training data (Brand et al., 2023). To analyze the reliance of the detected biases on familiar and learned studies, we create different scenarios across all experiments. The scenarios can range from structural prompt changes to changes of variable values in the experiments. Among finetuning of LLMs, Brand et al. (2023) modify persona descriptions and add market insights to examine the model's product preferences. Similarly, Kojima et al. (2022) extend the prompts by a single instruction to improve the reasoning capabilities of the models. Mizrahi et al. (2024) observe the sensitivity of LLMs towards the exact paraphrasing of an instruction. With our scenarios, we mainly test whether the biases change with different instructions and unseen or odd and unrealistic data, as shown in Table 3.

Scenario	**Description**
Base	Replication of the original study using base prompt format
Without Persona	Base scenario but exclusion of persona
Odd values	Base scenario but multiplication of the study variables (e.g. prices) times 9.7
Large values	Base scenario but multiplication of the study variables (e.g. prices) times 55,555.5

Table 3: *Overview of scenarios across experiments and their descriptions on prompt adjustments*

3.4 Response analysis

3.4.1 Processing responses into a bias indicator

With our study design of a control and test group per experiment, we aim to compare the response distributions to measure bias effects. While we are generally concerned with the existence of a bias in an experiment configuration, we especially aim to quantify the magnitude of the bias for more detailed comparisons and analysis between the experiments. With respect to identifying a bias effect, one could implement hypothesis tests for determining whether the response distributions differ significantly between the control and treatment groups. However, Sullivan and Feinn (2012) assert that while p-values may indicate a significant effect, the size of the effect is not extractable. p-values are influenced by the effect size of differences or sample sizes and can thus be difficult to interpret. Significant p-values do not necessarily stem from large effect sizes but can also e.g. from a large sample size and small (practically irrelevant) effects (Borenstein et al., 2021; Sullivan & Feinn, 2012). Further, even though methodology does exist to comparing p-values and incorporating them into meta analyses, strong assumptions are needed to make the results comparable (Borenstein et al., 2021).

As the comparability of the experiments is a key facet of our study and all experiments have similar structures and examine the same topic that is cognitive biases in LLMs, we orient ourselves towards traditional meta analysis methodologies by computing a determined effect size per experiment. In our setting of two independent groups (pre- and post-test measurements), we can calculate the standardized difference of means (Cohen's *d*) as follows (Borenstein et al., 2021; Cooper et al., 2019; Goulet-Pelletier & Cousineau, 2018; Morris & DeShon, 2002; Nakagawa et al., 2023):

$$d = \frac{\bar{X}_{post} - \bar{X}_{pre}}{SD_{pooled}} \tag{1}$$

where:

$\bar{X}_{pre}$, $\bar{X}_{post}$: means of the control and treatment groups, respectively

SD_{pooled}: $\sqrt{\frac{(n_{pre}-1)\cdot SD_{pre}^2+(n_{post}-1)\cdot SD_{post}^2}{n_{pre}+n_{post}-2}}$ (pooled standard deviation)

SD_{pre}, SD_{post}: standard deviations of the control and treatment groups

n_{pre}, n_{post}: sample sizes of the control and treatment groups

The effect size *d* is a standardized measure and allows for comparisons between different experiments and studies. Cohen's d can be interpreted as the magnitude of the bias effect, with larger values indicating a stronger effect. The effect size can be further interpreted using the guidelines by Cohen (1988) where values of 0.2, 0.5 and 0.8 are considered small, medium and large effects, respectively. Generally, Cohen's d is not an unbiased estimator of the population effect size due to overestimation and, as a result, a correction factor has been developed to form Hedges' g (Hedges, 1981). Goulet-Pelletier and Cousineau (2018) suggest that the correction factor is especially necessary for small sample sizes $N < 16$ and can be neglected for larger sample sizes. Therefore, we use the uncorrected Cohen's d for our analysis as all sample sizes in our experiments are $N = 200$.

For our experiments where we ask models to respond with numerical values, we can directly calculate the means and standard deviations of the responses to compute the effect size. To align with these experiments, we convert the models' responses in the choice experiments (i.e. A and B) to binary values, depending on which option we expect the models to choose in the treatment group. In the case of *loss aversion*, we expect the control group to select the risk averse gain option ($A = 0$) and the treatment group to select the risk seeking loss option ($B = 1$) (Thaler, 2015). We then calculate the effect size based on the binary responses.

By nature of Cohen's d calculation (Equation 1), the effect size can also be negative if the control group has a higher mean than the treatment group. For most of our bias experiments, we have recreated the experiments in a way that a higher mean in the treatment group refers to the expected bias effect. We thus expect a positive effect size for the existence of a bias and interpret any negative or zero effect sizes as no bias detected. One exception to this is the experiment regarding the *framing effect.* For this bias, a differentiating effect in either group hints at the existence of the framing effect. Therefore, we adjust the negative effect sizes by extracting their absolute values. The resulting metric, namely *bias detected*, is the following:

$$\textit{bias detected} = \begin{cases} |d| & \text{if bias} = \textit{framing effect} \\ d & \text{else} \end{cases} \tag{2}$$

where:

d: Cohen's d effect size by Equation 1

In the subsequent analysis, we will frequently limit *bias detected* to the range of 0 to 1 to simplify the interpretation. When aggregating, we may average across the original effect sizes to subsequently limit the values. This also allows us to focus more on the existence of a bias rather than the magnitude of the effect. We refer to the capped metric as:

$$\textit{bias detected (capped)} = \min(1, \max(0, \textit{bias detected})) \tag{3}$$

3.4.2 Analysis of effect sizes across biases and models

The *bias detected* metric allows us to compare the bias effects across all experiments. To follow our research question of whether biases are present in LLMs and how they differ between biases and models, we group all experiment detections by bias and model. Morris and DeShon (2002) suggest an average weighted by the reciprocal of the sampling variance of the effect sizes $\hat{\sigma}^2_{e_i}$ to aggregate effect sizes. Per experiment, the sampling variance is influenced by the sample sizes and the population effect size (unweighted average). As all our experiments have the identical sample sizes and each group has one fixed unweighted population effect size, the sampling variance is constant across all experiments of an aggregate. We can thus directly average the effect sizes to aggregate the bias detections (derivations in Appendix D). With the aggregated effect sizes, we compare the detections across biases and models to analyze the presence and magnitude of biases in the models. To simplify and illustrate the aggregated bias detections, we cap the aggregates at 0 and 1 to create *bias detected (capped)* (Equation 3). As we solely cut off the aggregates and do not rescale the effect sizes themselves, we ensure to correctly weigh extremely large and small effect sizes in the aggregation. Further, we maintain the interpretation guidelines by Cohen (1988) where values of 0.2, 0.5 and 0.8 are considered small, medium and large effects, respectively.

Besides the bias detections themselves, we test for homogeneity of the effect sizes across the models and biases. A constant bias detection across all experiments of a chosen subgroup would indicate a homogeneity of the effect. The homogeneity can be tested by comparing the theoretical variance due to sampling error $\hat{\sigma}^2_e$ with the observed variance of the effects sizes $\hat{\sigma}^2_d$ (Borenstein et al., 2021; Cooper et al., 2019; Morris & DeShon, 2002; Nakagawa et al., 2023). The derivations of both variances are detailed in Appendix D. Hunter and Schmidt (2004) propose a simple comparison and suggest that if the proportion of the theoretical to the observed variance exceeds 75 %, the effect size is regarded as homogeneous, i.e.:

$$\frac{\hat{\sigma}_e^2}{\hat{\sigma}_d^2} \geq 0.75 \tag{4}$$

where:

$\hat{\sigma}_e^2$: theoretical variance due to sampling error

$\hat{\sigma}_d^2$: observed variance of the effect sizes

Our expectation is that the effect sizes are not homogeneous across the biases and models. We expect the model answers to differ enough to often create slight discrepancies in the bias detections; the different scenarios and model temperatures should increase the variance of the effect sizes. Further, even if the scenarios and temperatures show no significant impact, we expect the response distributions to vary slightly due to the model re-initializations. To counter the influence of magnitude in the bias detections, we compute the homogeneities per bias and model for *bias detected (capped).* By this, we focus on whether we have homogeneous presence of biases across the experiments.

Lastly, we model the target variable *bias detected* with our four variables *bias*, *model*, *scenario* and *temperature.* We designed all experiments as combinations of the identical variable options. This not only ensures comparability across the experiments but enables us to model the categorical variables as fixed instead of random effects in a linear regression model (Borenstein et al., 2021; Cooper et al., 2019; Nakagawa et al., 2023). We use a linear regression model to analyze the impact of the variables on the effect sizes. With the results of the regression model, we can identify the influence of each component of the experiments on the bias detections, focusing on biases and models. We regress the experiment variables on both the original *bias detected* and *bias detected (capped)* targets but focus on the capped target variable to not distort the regression results with extreme effect sizes. The regressions are modeled as follows (exemplary for capped target):

$$\textit{bias detected (capped)}_{b,m,s,t} = \beta_0 + \beta_1 \textit{bias}_b + \beta_2 \textit{model}_m + \beta_3 \textit{scenario}_s + \beta_4 \textit{temperature}_t + \epsilon_{b,m,s,t} \tag{5}$$

3.4.3 Scenario impacts on bias detections

We also aim to assess the impact of the scenarios on the bias detections in the models. We are especially interested in whether the removal of the persona and the odd and large values have any effect on the bias detections. Primarily, we use the previously presented regression models of all experiment components including the scenarios on the bias detected variables. By analyzing the regression coefficients, we can identify the impact of the scenarios compared to the base scenario. We further regress only the scenarios on the bias detections in order to analyze the explained variance of the scenarios. We expect that the inclusion of a persona as well as normal values trigger more biased responses in the models.

3.4.4 Explainability of biases through model features

To further analyze the bias occurrences in the LLMs, we investigate different model features and whether they can explain the degree of bias detections. We collect data on the models' release, update and knowledge cutoff dates, their number of parameters and context length as well as some performance benchmarks, i.e. *MMLU* (Massive Multitask Language Understanding) and *Chatbot Arena*. With the time of release and update, we aim to analyze whether newer models possess different bias characteristics than older models. The knowledge cutoff refers to the latest date training data was included. Further, we expect larger models to ingest more non-linearities and biases than smaller models. The context length is also a major factor in the models' reasoning capabilities and thus could influence the bias detections (Naveed et al., 2023; Zhao et al., 2023).

MMLU is a benchmark for evaluating knowledge capabilities across various domains. The test covers 57 tasks with different difficulties and can also offer insights into a model's reasoning abilities. It is widely adopted and included in model announcements most of the time (Hendrycks et al., 2020). We also include the scores of the *Chatbot Arena* which enables users to choose a preferred option of paired responses to rank the models. The *Chatbot Arena* is widely used (187 models with nearly 2.5 million votes) and provides insights into a model's human likeness and alignment which could hint at more or less biased models (Chiang et al., 2024; Hugging Face, 2024a).

It should be noted that while open-source models are well documented, we also had to estimate some model features for closed-source models. The models and their features are detailed in Appendix B. As in the previous analysis, we regress the model features and the specified model temperature in the experiment on the bias detections and analyze their explainability towards and impact on the bias detections.

4 Results

4.1 Analysis of detected biases per bias and model

4.1.1 Bias detections

The histogram in Figure 1 portrays a high portion of bias detections around 0. In fact, the amount of detections where there is no difference between the means of the two groups (*bias detected = 0*) is 59.2 %. This means that in roughly six out of ten cases, the models did not respond differently at all (in their means) to the two prompts. Further, we find that 80.4 % of all bias detections are within the expectations and interpretation guidelines by Cohen (1988) (between -1 and 1). For our *bias detected (capped)* metric (Equation 3), we capture 74.2 % of the original detections and cap the outliers for interpretability purposes as we consider all effects above 0.8 as large effect sizes. The distribution displays a slight tendency towards positive values, indicating more biased than unbiased responses (32.2 % of detections above 0, 8.6 % of detections below 0).

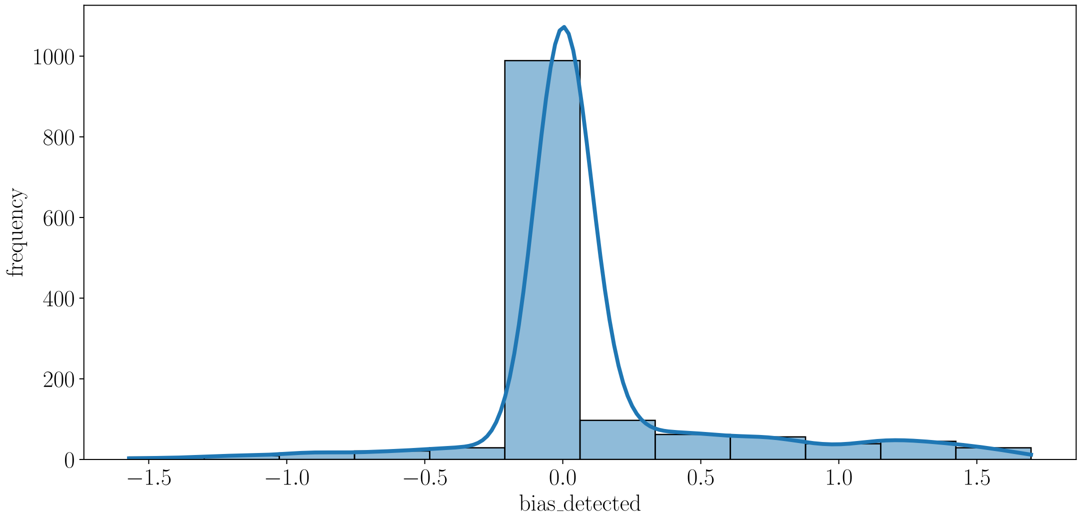

Figure 1: *Distribution of bias detections (uncapped). Plot depicts distributions of bias detections between -1.7 and 1.7 which represents 87.4 % of all detections (for visualization purposes).*

Aggregating across all biases and models, we find a total of 28 bias detections (*bias detected (capped)* $\geq$ 0.5). We display the aggregations as a heatmap in Figure 2. Our aggregation method described in Chapter 3.4.2 leads to a high density of bias detections around 0 and 1. However, as the scaling in between the minimum and maximum stays untouched, we are still able to apply the interpretation guidelines for small, medium and large effects by Cohen (1988).

The heatmap depicts a high number of bias detections for the *anchoring bias*, the *endowment effect*, the *framing effect* and *loss aversion*. Especially the *anchoring bias* is detected with a large effect in all models. This means that all models are heavily influenced in their estimations when being exposed to a reference point (anchor). Additionally, most models reveal to be biased with regard to the *framing effect*, to varying magnitudes. Generally, the model twins with more parameters seem to be more biased towards this, with the small *Phi3.5* model being an exception. When detected, the *endowment effect* also has a large impact on the models' responses. All models except for *Llama3.1* and *Phi3:medium* are biased towards valuing their own possession higher than an offered item. *Loss aversion*, if detected, shows strong to very strong effects in the models or none at all. Especially the model families by Anthropic and OpenAI show strong tendencies to avoiding losses and therefore taking more risks. Smaller models show less bias towards *loss aversion* except for *Llama3.1:70b* which is less biased than its 8 billion parameter twin.

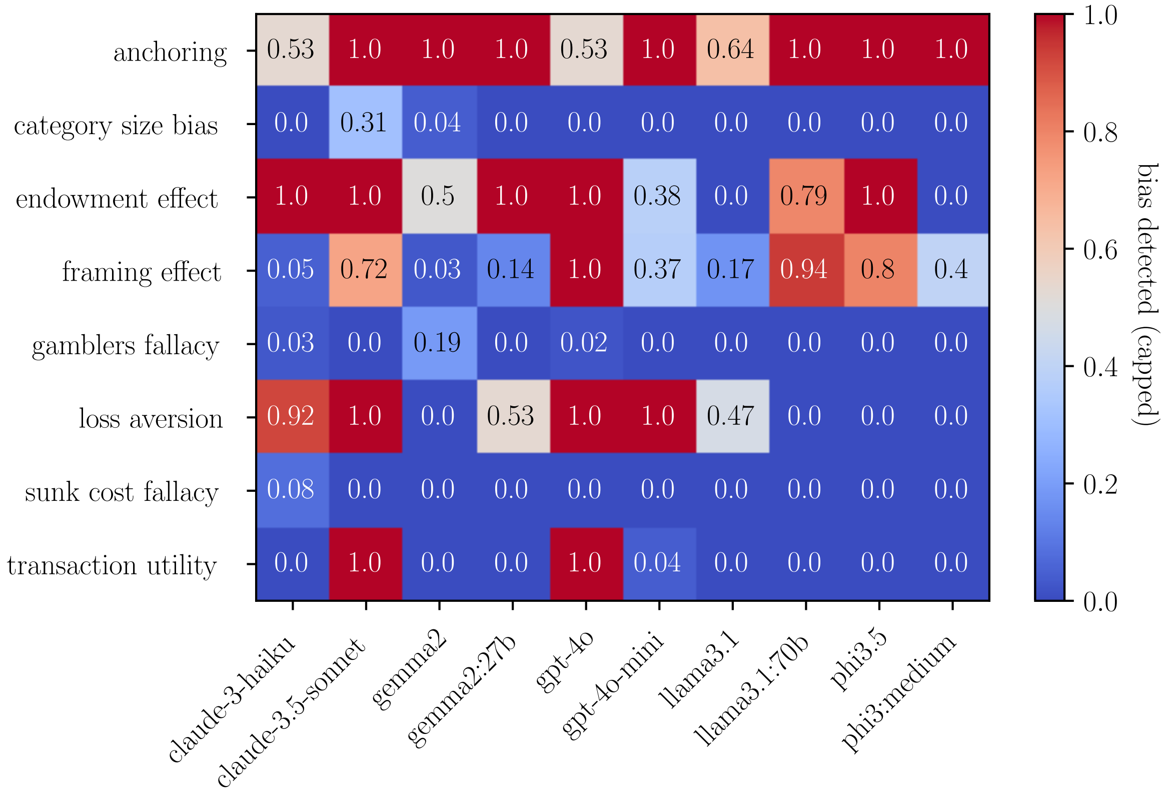

Figure 2: *Detected biases grouped by biases and models. 0 signals non-existent and 1 fully-existent bias. Detailed calculation and aggregation of the target variable described in chapters 3.4.1 and 3.4.2, respectively.*

Some biases were not detected in any of the models. Both the *gambler's fallacy* and *category size bias* are biases that examine the statistical understanding of independent subjects and events. The LLMs seem to not show any bias towards these fallacies. The *sunk cost fallacy* is also not detected in any of the models. Humans generally prefer the option with the larger sunk costs, independent of their actual emotional preference. Interestingly, we find that the models not only do not exhibit this bias, but prefer the emotionally preferred option with lower sunk costs. This could be caused by the missing relationship and reference towards the monetary value of the sunk costs.

The *transaction utility bias* is an extreme case amongst our biases as it is not detected in any models except for *GPT-4o* and *Claude-3.5-Sonnet*. The experiment aimed to investigate if participants would shift your purchasing decision based on the transaction utility of the product (difference between prices but linked to a 20-minute drive). While human participants altered their decision based on the relative amount of money saved to the purchasing prices, most of the models understood that both scenarios had the same absolute price difference. It seems that amongst all models, only *GPT-4o* and *Claude-3.5-Sonnet* were able to understand the transaction utility and time implications of the experiment.

From a model perspective, *GPT-4o* and *Claude-3.5-Sonnet* are the most biased models amongst our bias and model combinations. While the *anchoring bias* is not as strongly detected as in other models, *GPT-4o* either shows strong bias detections or no bias detections at all. The smaller twin *GPT-4o-Mini* shows similar but less pronounced behavior except for the anchoring bias. We find similar patterns for the *Claude*, *Gemma* and *Llama* models, where the smaller model often shows less biased behavior than the larger model. The *Phi* models show a different pattern, where the larger model is less biased than the smaller model. *Phi3:Medium* is in fact the least biased model of our selections. This could be due to the later training of *Phi3.5* and thus newer data with model improvements or the much larger context length.

4.1.2 Homogeneities of bias detections

We also test the detected biases for homogeneity. The focus lies on whether a bias was detected and not on the exact magnitude. Therefore, in contrast to the aggregated bias detections, we transform the original bias detections before calculating the homogeneities, then set limits at 0 and 1 and display them in a heatmap (see Figure 3). We particularly find homogeneity for non-bias detections.

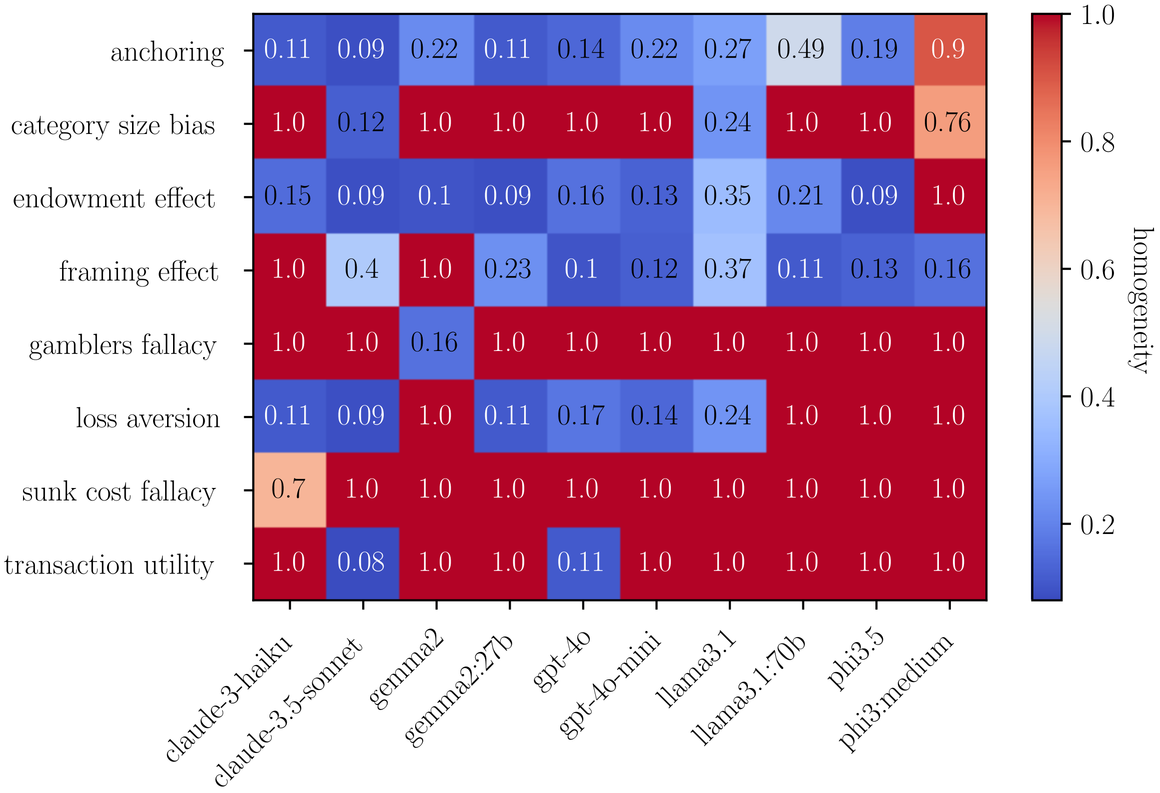

Figure 3: *Homogeneities of capped effect sizes (between 0 and 1) grouped by biases and models. 0 signals complete heterogeneity and 1 complete homogeneity. Detailed calculation of homogeneity detailed in chapter 3.4.2.*

As we expected, slight and strong bias detections are quite heterogeneous across the bias-model combinations. This indicates that the effect sizes (even though capped) are not consistent across the combinations. The averaging of the effect sizes of all scenarios and model temperatures could be one reason for the heterogeneity. The small but perhaps influential adaptions to the model prompts and their response behavior could lead to slightly different model responses and thus effect sizes. Further, re-initializing the models between each experiment could also lead to inconsistent responses. Some exceptions with either very homogeneous, high bias detections or very heterogeneous, low bias detections are present and are highlighted in Table 8. A low homogeneity even though we did not detect a bias could indicate that in some scenarios or with some model temperatures, a bias is more present than in others.

4.1.3 Regression analysis of biases and models

The regressions on the uncapped and capped bias detections (Table 4) show that the bias detections of the models significantly depend on the explored bias ($p < 0.05$). This is consistent with the hypothesis that the detections seem to especially be dependent of the specific bias itself, as suggested by the heatmap in Figure 2.

Target variable	**bias_detected (R^2=7.2%)**			**bias_detected_capped (R^2=36.7%)**		
Variable	*Coef.*	*Std. Err.*	*p-value*	*Coef.*	*Std. Err.*	*p-value*
Intercept	3.9224	1.197	0.001	0.6241	0.038	0.000
Bias: category size bias	-7.3005	0.991	0.000	-0.6212	0.032	0.000
Bias: endowment effect	-2.0941	0.990	0.034	-0.2625	0.031	0.000
Bias: framing effect	-2.7081	0.990	0.006	-0.4660	0.031	0.000
Bias: gamblers fallacy	-3.2171	0.990	0.001	-0.6693	0.031	0.000
Bias: loss aversion	-2.0655	0.990	0.037	-0.3604	0.031	0.000
Bias: sunk cost fallacy	-3.2323	0.990	0.001	-0.6831	0.031	0.000
Bias: transaction utility	-2.9421	0.990	0.003	-0.5774	0.031	0.000
Model: claude-3.5-sonnet	1.8203	1.106	0.100	0.0938	0.035	0.008
Model: gemma2	1.2924	1.106	0.243	-0.0351	0.035	0.318
Model: gemma2:27b	0.0876	1.106	0.937	-0.0649	0.035	0.065
Model: gpt-4o	0.9837	1.106	0.374	0.1783	0.035	0.000
Model: gpt-4o-mini	1.1737	1.106	0.289	0.0551	0.035	0.117
Model: llama3.1	-1.0772	1.106	0.330	-0.0240	0.035	0.496
Model: llama3.1:70b	0.1346	1.108	0.903	0.0046	0.035	0.896
Model: phi3.5	-4.3647	1.106	0.000	-0.0255	0.035	0.468
Model: phi3:medium	-0.1172	1.106	0.916	-0.0418	0.035	0.235
Scenario: 1_no_persona	-2.1589	0.700	0.002	0.0081	0.022	0.716
Scenario: 2_odd_numbers	0.5583	0.700	0.425	-0.0006	0.022	0.977
Scenario: 3_large_numbers	-0.5939	0.700	0.396	-0.0734	0.022	0.001
Temperature	-0.1271	0.458	0.781	0.0692	0.015	0.000

Table 4: *Summarized overview of regressions of experiment variables on bias detections. More detailed results can be found in Appendix F.1.*

The regression on *bias detected* has a very low explained variance while the regression on *bias detected (capped)* has a higher explained variance. This likely reflects the influence of outliers and variance in the exact magnitude of the bias detections, aligning with the heterogeneity analysis of the bias detections. For the latter regression, most models do not significantly influence bias detections except for the two largest models, *GPT-4o* and *Claude-3.5-Sonnet*. Both have significant positive effects on the bias detections, indicating that these models are more prone to the biases.

4.2 Impact of scenarios

From our regression results in Table 4, we find that the exclusion of the persona prompt leads to significantly lower bias detections across the unmodified effects. This supports our expectation that the persona leads to more human-like and thus more biased responses.

Though on the capped bias detections, the persona prompt does not have a significant influence, suggesting that it primarily amplifies the extremity of the bias detections. However, this could also be due to the capping process, which may limit the ability to fully capture the effect of changes in bias detection.

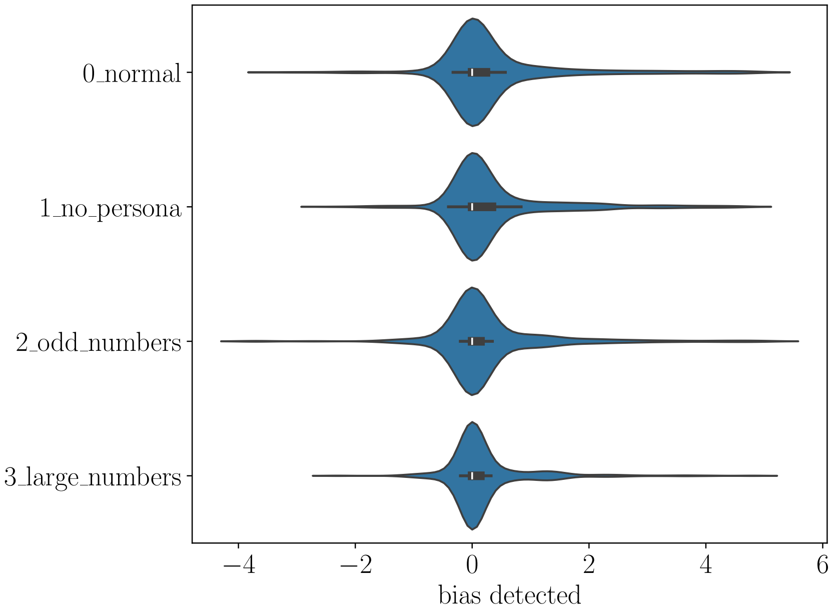

Figure 4: *Bias detections (uncapped) grouped by scenarios. Plot depicts distributions of bias detections between -5 and 5 (for visualization purposes).*

A glimpse at the distributions per scenario in Figure 4 and the regression results of the scenarios on *bias detected (capped)* extend our analysis. The distribution plots do not display clear differences except a slight decrease of minimal bias detections in the scenario with extremely large values. This aligns with the regression results where this scenario has the only significant effect with a slightly negative effect on the bias detections compared to the base scenario. The scenario with similar but odd numerical values did not show any significant impact in both regressions. This suggests that the models are more likely to exhibit biases when the values are realistic and thus possibly closer to the training data.

Further, the regressions exclusively of the scenarios on the bias detections (Appendix F.2) indicate that the explained variance of the scenarios is very low ($R^2 \leq 1\%$). The low R^2 indicates that scenarios alone do not sufficiently explain the variability in bias detections, suggesting other factors play a more prominent role. While some scenarios do show slight significant impacts and tendencies (missing persona and large numbers lead to lower bias detections in the models), the overall influence of the scenarios on the bias detections appears to be minimal.

4.3 Model feature analysis

Target variable	**bias_detected ($\mathbf{R^2}$=2.4%)**			**bias_detected_capped ($\mathbf{R^2}$=4.1%)**		
Variable	*Coef.*	*Std. Err.*	*p-value*	*Coef.*	*Std. Err.*	*p-value*
Intercept	-24.490	6.469	0.000	-0.0198	0.247	0.936
Temperature	-0.1295	0.468	0.782	0.0691	0.018	0.000
△Release date	0.0152	0.023	0.517	-0.0006	0.001	0.497
△Last-updated date	-0.0020	0.028	0.943	0.0005	0.001	0.638
△Training data cutoff date	-0.0046	0.006	0.413	0.0004	0.000	0.038
Number of parameters	-0.0050	0.013	0.710	0.0015	0.001	0.004
Context length	-0.0012	0.016	0.940	-0.0004	0.001	0.555
MMLU	0.0554	0.113	0.624	-0.0065	0.004	0.129
Chatbot Arena	0.0168	0.008	0.026	0.0004	0.000	0.130

Table 5: *Summarized overview of regressions of model features on bias detections. More detailed results can be found in Appendix F.3.*

To further investigate the language models and their bias detections, we regress selected model features on both *bias detected* and *bias detected (capped).* The key results of the regressions are displayed in Table 5. As for the scenario regressions in Chapter 4.2, the regression models have little explained variance, though the latter regression without the extreme effects fits the data slightly better. For this regression, we find some significant effects. Most notably, a higher temperature significantly increases the bias detection in a model. This aligns with our expectations as a higher model temperature generally leads to more creative and random responses, thus probably differing more between the two prompts in each experiment and apparently leading to higher biases. Further, larger models with more parameters show slight gains for detecting biases. This overlaps with our previous analysis where the larger models *GPT-4o* and *Claude-3.5-Sonnet* showed higher bias detections than other models.

We also find a significant though minor effect for the knowledge cutoff of the models. An earlier training data cutoff leads to a slightly higher *bias detected (capped).* This could be due to enhanced data curation and preprocessing as well as new training methods leading to less model intake of biases. However, the effect is minor and should not be overinterpreted. Both model evaluation benchmarks *MMLU* and the *Chatbot Arena Score* do not significantly explain the existence of a bias. Though the *Chatbot Arena Score* has a significant impact on the uncapped *bias detected*, the effect is small. Still, this points out that the human-likeliness scores in the Chatbot Arena could be a minor indicator for models with more human-like behavior and biases. In neither regression, the release and last-updated dates had significant effects on the bias detections. This suggests that amongst our selected models, the model's age and the time since the last update do not significantly influence the bias detections.

5 Conclusion and discussion

In this thesis, we have explored the existence of human cognitive biases in LLMs, addressing the research questions of whether these models exhibit the biases, how persona prompting and odd and extremely large values influence bias detections and whether certain models and model features are more prone to biases. Through a comprehensive series of experiments, we have provided quantitative evidence that cognitive biases are not only present in LLMs but also vary across biases, models, their temperatures as well as prompt scenarios. We present a methodology to process the model responses into a standardized metric across all types of biases, allowing for a direct comparison of the biases' presence and magnitude. This should enable future research to build upon our findings and further investigate the biases in the language models.

Our results confirm that LLMs frequently exhibit biases that resemble those found in human decision-making processes. Especially the *anchoring bias*, *endowment effect*, *framing effect* and *loss aversion* were detected consistently across multiple models and scenarios. Contrary, some biases such as the *category size bias*, *gambler's fallacy* and *sunk cost fallacy* showed no bias detections in our experiments. We thus conclude that it is pivotal to be aware of which biases are present in the LLMs and which biases a developer or researcher seeks to invoke or mitigate.

With our prompt adjustments, we have explored possibilities to influence the biases detected in the models. We examine that extremely large values in the experiments lead to less biased behavior in the models. The exclusion of a persona prompt also shows shifts towards less biased behavior. Declaring the model to act humanlike and using less realistic values in the prompts which are less likely to be included in the training data are both factors that lead to more biased behavior in the models and vice versa.

We find that larger language models such as *GPT-4o* and *Claude-3.5-Sonnet* are significantly more prone to biased behavior, though the effect is small. The model temperature also has a significant effect, with higher temperatures leading to more biased responses. While the knowledge cutoff has a slight significant effect, the release and lastly-updated dates as well as model metrics show no significant impact on the bias detections. We can thus conclude that careful model selection and hyperparameter setting can help to mitigate biases in LLMs. For market research and gravitating towards human-like, biased behavior, larger models with higher temperatures which were trained on recent data seem to be more suitable than others.

Despite the promising results, our study has limitations. Primarily, our regression results lack explained variance, suggesting that our experiments do not fully capture the complexity of cognitive biases in LLMs yet. This makes it difficult to draw precise conclusions and best-practice advice. Additionally, our bias detection is limited to a small set of biases, and we do not consider interactions between the biases. Similarly, the model selection is limited to a relatively small set of LLMs, some of which have been updated recently. Future research should expand the set of biases and models to provide an even more comprehensive understanding and generalization of cognitive biases in LLMs. To make each bias detection more robust, the studies per bias could be expanded to include more diverse questioning targeting the bias and perhaps more reasoning-centered questions.

Further, other prompting techniques should be assessed to better understand their impact on biased model behavior and to form best-practice guidelines for practitioners. The rapid development of LLMs could also lead to a more extensive investigation of biases across multiple languages, i.e. multilingual, and modalities, e.g. images, videos. The latter could open up another spectrum of cognitive biases which are not necessarily existent in text-based environments. The collection of more detailed model features could provide more explainability towards why certain models are more prone to biases than others.

Building a framework to predict the existence and magnitude of biases in LLMs could help practitioners understand the biases in their models and potentially mitigate or invoke them. More models and model features, more biases with robust studies and more diverse prompting techniques could provide best-practice guidelines and a foundation for the ethical use of these models in the future.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219.*

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774.*

Anthropic, A. (2024a). The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, *1*.

Anthropic, A. (2024b). Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, *3*.

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational behavior and human decision processes*, *35*(1), 124–140.

Arnott, D. (1998). A taxonomy of decision biases. *Monash University, School of Information Management and Systems, Caulfield.*

Azzopardi, L. (2021). Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. *Proceedings of the 2021 conference on human information interaction and retrieval*, 27–37.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics*, *12*(4), 428–454.

Barron, G., & Leider, S. (2010). The role of experience in the gambler's fallacy. *Journal of Behavioral Decision Making*, *23*(1), 117–129.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis.* John Wiley & Sons.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using gpt for market research. *Harvard Business School Marketing Unit Working Paper*, (23-062).

Chandrashekaran, R., & Grewal, D. (2006). Anchoring effects of advertised reference price and sale price: The moderating role of saving presentation format. *Journal of Business Research*, *59*(10-11), 1063–1071.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. arxiv.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates, Inc.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A., & Dragicevic, P. (2018). A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics*, *26*(2), 1413–1432.

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951.*

Druckman, J. N. (2001). Evaluating framing effects. *Journal of economic psychology*, *22*(1), 91–101.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological methods*, *1*(2), 170.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with llms. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12640–12653.

Ekin, S. (2023). Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices. *Authorea Preprints*.

Exploding Topics Team. (2024, August). Understanding GPT Parameters [Last accessed: December 30, 2024]. https://explodingtopics.com/blog/gpt-parameters

FelloAI Team. (2024, August). Claude AI: Everything You Need to Know [Last accessed: December 30, 2024]. https://felloai.com/2024/08/claude-ai-everything-you-need-to-know/

Furia Team. (2024, August). What is GPT-4o? [Last accessed: December 30, 2024]. https://www.furia.fi/en/artificial-intelligence-en/what-is-gpt-4o/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

Gerganov, J. (2023, November). *Llama.cpp* [Last accessed: September 12, 2024]. https://github.com/ggerganov/llama.cpp

Gigerenzer, G. (2007). Gut feelings: The intelligence of the unconscious. *New York: Viking*.

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part i: The cohen'sd family. *The Quantitative Methods for Psychology*, *14*(4), 242–265.

Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., et al. (2024). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Hagendorff, T., & Fabi, S. (2024). Why we need biased ai: How including cognitive biases can enhance ai systems. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*(8), 1885–1898.

Haltaufderheide, J., & Ranisch, R. (2024). The ethics of chatgpt in medicine and healthcare: A systematic review on large language models (llms). *NPJ digital medicine*, *7*(1), 183.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hu, T., & Collier, N. (2024). Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Hugging Face. (2024a, July). Chatbot Arena Leaderboard [Last accessed: December 30, 2024]. https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

Hugging Face. (2024b, July). Llama 3.1 8B [Last accessed: December 30, 2024]. https://huggingface.co/meta-llama/Llama-3.1-8B

Hugging Face. (2024c, July). Phi-3-medium-4k-instruct [Last accessed: December 30, 2024]. https://huggingface.co/microsoft/Phi-3-medium-4k-instruct

Hugging Face. (2024d, July). Phi-3.5-mini-instruct [Last accessed: December 30, 2024]. https://huggingface.co/microsoft/Phi-3.5-mini-instruct

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Isaac, M. S., & Brough, A. R. (2014). Judging a part by the size of its whole: The category size bias in probability judgments. *Journal of Consumer Research*, *41*(2), 310–325.

Jarmolowicz, D. P., Bickel, W. K., Sofis, M. J., Hatz, L. E., & Mueller, E. T. (2016). Sunk costs, psychological symptomology, and help seeking. *Springerplus*, *5*, 1–7.

Kahneman, D. (2017). *Thinking, fast and slow*.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of political Economy*, *98*(6), 1325–1348.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

Knetsch, J. L. (1989). The endowment effect and evidence of nonreversible indifference curves. *The american Economic review*, *79*(5), 1277–1284.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199–22213.

Kovic, M., & Kristiansen, S. (2019). The gambler's fallacy fallacy (fallacy). *Journal of risk research*, *22*(3), 291–302.

Leonard, C. A., Williams, R. J., & Vokey, J. (2015). Gambling fallacies: What are they and how are they best measured?

Liu, J. (2022, November). *LlamaIndex* [Last accessed: September 12, 2024]. https://github.com/jerryjliu/llama_index

Liu, Y. (2023). The review of loss aversion. *Advances in Education, Humanities and Social Science Research*, *7*(1), 428–428.

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. *International conference on data intelligence and cognitive informatics*, 387–402.

Meta. (2024, July). Introducing llama 3.1: Our most capable models to date [Last accessed: September 11, 2024]. https://ai.meta.com/blog/meta-llama-3-1/

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, *12*, 933–949.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological methods*, *7*(1), 105.

Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R., & Lagisz, M. (2023). Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence*, *12*(1), 8.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Neoteric Blog Team. (2024, August). Claude 3.5 Sonnet vs GPT-4o and GPT-4o Mini [Last accessed: December 30, 2024]. https://neoteric.eu/blog/claude-3-5-sonnet-vs-gpt-4o-and-4o-mini/

Olea, C., Tucker, H., Phelan, J., Pattison, C., Zhang, S., Lieb, M., & White, J. (2024). Evaluating persona prompting for question answering tasks. *Proceedings of the 10th international conference on artificial intelligence and soft computing, Sydney, Australia*.

OpenAI. (2024a, May). Gpt 4o mini: Advancing cost-efficient intelligence [Last accessed: September 10, 2024]. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

OpenAI. (2024b, July). Hello gpt 4o: Openai's new generation language model [Last accessed: September 10, 2024]. https://openai.com/index/hello-gpt-4o/

Oudin, P., & Groza, T. (2024). The governance of ai companies: Reconciling purpose with profits. *Available at SSRN*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Qiu, L., Singh, P. V., & Srinivasan, K. (2023). How much should we trust llm results for marketing research? *Available at SSRN 4526072*.

Ramos, V. J. (2018). *Analyzing the role of cognitive biases in the decision-making process*. IGI Global.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Santu, S. K. K., & Feng, D. (2023). Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv:2305.11430*.

Schmidt, D. C., Spencer-Smith, J., Fu, Q., & White, J. (2024). Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *ACM SIGAda Ada Letters*, *43*(2), 43–51.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268.

Simon, H. A. (1990). Bounded rationality. *Utility and probability*, 15–18.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, *4*(3), 279–282.

Talboy, A. N., & Fuller, E. (2023). Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption. *arXiv preprint arXiv:2304.01358*.

Taniguchi, H., Sato, H., & Shirakawa, T. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. *Scientific reports*, *8*(1), 7397.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Thaler, R. H. (1983). Transaction utility theory. *Advances in consumer research*, *10*(1).

Thaler, R. H. (2008). Nudge: Improving decisions about health. *Wealth, and Happiness*, *3*.

Thaler, R. H. (2015). *Misbehaving: The making of behavioral economics*. WW Norton & Company.

The Decoder Team. (2024, August). Anthropic Introduces Claude: An AI Assistant Trained by an AI [Last accessed: December 30, 2024]. https://the-decoder.com/anthropic-introduces-claude-an-ai-assistant-trained-by-an-ai/

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological review*, *101*(4), 547.

Yub, H. A. (2024, August). GPT-4o: Successor of GPT-4 [Last accessed: December 30, 2024]. https://hamidayub.medium.com/gpt-4o-successor-of-gpt-4-8207acf9104e

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

APPENDIX

A Bias details and experiments

Table 6: *Overview of the bias concepts and the corresponding experiment questions. "A" refers to the (perhaps modified) question for the control group, "B" for the test group. "Expect" refers to the expected results.*

Bias	**Concept**	**Source**	**A**	**B**	**Expect**
Anchoring bias	Reliance on previous anchor for subsequent decision	Tversky and Kahneman (1974)	Do you think the portion of African countries in the United Nations is higher or lower than **10%**? Now that you have thought about that, what would you estimate the actual exact portion to be? Answer the percentage between 0 and 100 (Please only answer the percentage as a number without the % sign.). % ___	Do you think the portion of African countries in the United Nations is higher or lower than **65%**? Now that you have thought about that, what would you estimate the actual exact portion to be? Answer the percentage between 0 and 100 (Please only answer the percentage as a number without the % sign.). % ___	Higher estimations in B due to higher anchor
Category size bias	Misjudgment of probabilities due to category size	Isaac and Brough (2014)	Imagine a lottery containing **15** balls. Balls **1** to **5** are black, **6** to **10** are gray, and **11** to **15** are white. If one ball is drawn from the urn, what is the probability that it will be ball **8**? Answer the percentage between 0 and 100. % ___	Imagine a lottery containing **15** balls. Balls **1** to **2** are black, **3** to **13** are gray, and **14** to **15** are white. If one ball is drawn from the urn, what is the probability that it will be ball **8**? Answer the percentage between 0 and 100. % ___	Higher estimations in B due to larger category size

Continued on next page

Continued

Bias	**Concept**	**Source**	**A**	**B**	**Expect**
Endowment effect	Overvaluation of owned items and vice versa	Kahneman et al. (1990)	I show you a mug from a distance. You can see it but you cannot touch it. The mug was originally priced at **$4**, but the value of the mug is different for different individuals. What price would you buy it for? $ ___	I hand you a mug. You now own the mug. You can feel it and look at it in detail from all angles. The mug was originally priced at **$4**, but the value of the mug is different for different individuals. What price would you sell it for? $ ___	Higher valuation in B due to ownership
Framing effect	Different valuations and decision-making due to framing	Tversky and Kahneman (1981)	Imagine that you have decided to see a play and paid the admission price of **$10** per ticket. As you enter the theater, you discover that you have lost the ticket. The seat was not marked and the ticket cannot be recovered. Would you pay **$10** for another ticket to the play? Choose between A) Yes or B) No. ___	Imagine that you have decided to see a play where admission is **$10**. As you enter the theater, you discover that you have lost a **$10** bill. Would you still pay **$10** for a ticket to the play? Choose between A) Yes or B) No. ___	Different decision-making in A and B due to framing
Gambler's fallacy	Misunderstanding of independent events	Leonard et al. (2015)	A fair coin has been flipped **6** times and has come up heads half the time and tails half the time. What is the probability that the next flip will be heads? Answer the percentage between 0 and 100. % ___	A fair coin has been flipped **6** times and has come up heads each time. What is the probability that the next flip will be heads? Answer the percentage between 0 and 100. % ___	Higher estimations in B due to gambler's fallacy

Continued on next page

Continued

Bias	**Concept**	**Source**	**A**	**B**	**Expect**
Loss aversion	Avoidance of losses through risk-taking over gains	Thaler (2015)	Assume yourself richer by **$300** than you are today. You are offered a choice between A. A sure gain of **$100**, or B. A 50% chance to gain **$200** and a 50% chance to lose $0. Choose between A and B. ___	Assume yourself richer by **$500** than you are today. You are offered a choice between A. A sure loss of **$100**, or B. A 50% chance to lose **$200** and a 50% chance to lose $0. Choose between A and B. ___	Higher portion of risk-seeking option in A due to loss aversion
Sunk cost fallacy	Overvaluation of past investments	Arkes and Blumer (1985)	Assume that you have spent **$50** on a ticket for a weekend ski trip to Michigan. Several weeks later you buy a **$50** ticket for a weekend ski trip to Wisconsin. You think you will enjoy the Wisconsin ski trip more than the Michigan ski trip. As you are putting your just-purchased Wisconsin ski trip ticket in your wallet, you notice that the Michigan ski trip and the Wisconsin ski trip are for the same weekend! It's too late to sell either ticket, and you cannot return either one. You must use one ticket and not the other. Which ski trip will you go on? A) **$50** ski trip to Michigan B) **$50** ski trip to Wisconsin. ___	Assume that you have spent **$100** on a ticket for a weekend ski trip to Michigan. Several weeks later you buy a **$50** ticket for a weekend ski trip to Wisconsin. *...identical to A...* Which ski trip will you go on? A) **$100** ski trip to Michigan B) **$50** ski trip to Wisconsin. ___	Higher portion of choosing the expensive and emotionally less-preferred option in B

Continued on next page

Continued

Bias	**Concept**	**Source**	**A**	**B**	**Expect**
Transaction utility theory	Decision-making based on relative price differences	Thaler (1983)	You set off to buy a new television. At the store where you expect to buy it, you find that the price is **$650**. A clerk informs you that the same item is available at another branch of the same store for on **$640**. The store is a 20-minute drive away and the clerk assures you that they have what you want there. Do you buy A) at the initial store or B) go to the other store? Choose between A and B. ___	You set off to buy a new radio. At the store where you expect to buy it, you find that the price is **$35**. A clerk informs you that the same item is available at another branch of the same store for on **$25**. The store is a 20-minute drive away and the clerk assures you that they have what you want there. Do you buy A) at the initial store or B) go to the other store? Choose between A and B. ___	Higher portion of choosing the alternative option in B due to stronger relative effect of $10 on $35 than on $650

All **bold** variables are the adapted variables in the scenarios.

B Model details and features

Table 7: *Comparison of selected models and their features*

Model	Release Date	Last Updated Date	Training Data Cutoff Date
claude-3-haiku	Mar 7, 2024	Mar 7, 2024	Aug, 2023
claude-3.5-sonnet	Jun 20, 2024	Jun 20, 2024	Apr, 2024
gemma2	Jun 27, 2024	Aug 4, 2024	Jun, 2024
gemma2:27b	Jun 27, 2024	Aug 4, 2024	Jun, 2024
gpt-4o-mini	Jul 18, 2024	Jul 18, 2024	Oct, 2023
gpt-4o	May 13, 2024	Aug 6, 2024	Oct, 2023
llama3.1	Jul 23rd, 2024	Aug 11, 2024	Dec, 2023
llama3.1:70b	Jul 23rd, 2024	Aug 11, 2024	Dec, 2023
phi3.5	Aug 20, 2024	Aug 20, 2024	Oct, 2023
phi3:medium	Apr 23rd, 2024	Aug 4, 2024	Oct, 2023

Model	# Parameters (bn)	Context Length	MMLU	Chatbot Arena**
claude-3-haiku	20*	200,000	75.2	1,179
claude-3.5-sonnet	175*	200,000	88.7	1,268
gemma2	9	8,192	71.3	1,191
gemma2:27b	27	8,192	75.2	1,220
gpt-4o-mini	40*	128,000	82	1,273
gpt-4o	175*	128,000	88.7	1,265
llama3.1	8	128,000	73	1,176
llama3.1:70b	70	128,000	86	1,248
phi3.5	4	128,000	69*	1,037
phi3:medium	14	4,000	78	1,123

*Approximated values due to lack of official documentation **Hugging Face (2024a)

For most of the models, we were able to collect relevant information. However, there is less official documentation on the closed-source models. Following are the sources:

- **Claude-3 & 3.5**: Anthropic (2024a, 2024b), FelloAI Team (2024), Neoteric Blog Team (2024), and The Decoder Team (2024)
- **Gemma2**: Team et al. (2024)
- **GPT-4o**: Exploding Topics Team (2024), Furia Team (2024), OpenAI (2024a, 2024b), and Yub (2024)
- **Llama3.1**: Dubey et al. (2024), Hugging Face (2024b), and Meta (2024)
- **Phi3 & 3.5**: Abdin et al. (2024) and Hugging Face (2024c, 2024d)

C Exemplary code snippet of experiment run

Example code for *Llama 3.1:70b* and *framing effect* (question A, base scenario):

```python
from llama_index.llms.ollama import Ollama
import ollama
from typing import List

####-- SETTINGS --####
model: str = "llama3.1:70b"  # Model to use
temperature: float = 0.7  # Model temperature
max_tokens: int = 2  # Number of tokens to predict
n: int = 100  # Number of runs
response_type: str = "choice"  # Choice or numerical

####-- PROMPTS --####
##- System message -##
# Persona
persona: str = "You are a customer with median income and average education. You are selected at random to participate in a survey. You can only choose one of the presented options or assign a value. Behave humanlike and choose instinctively. You can and are advised to make a subjective decision! There is no right or wrong, but you HAVE TO DECIDE."

# Additional system message
system_message: str = f"You will be asked to make choices. Please blank out that some information might be missing or that you might not be able to make a choice. I have initialized you in a way that you can only generate {max_tokens} tokens. The only valid answer is A SINGLE LETTER OR NUMBER."

# Combine messages
system_message += f" {persona}" if persona != "" else ""

##- User message -##
# User message
user_message: str = "You are forced to choose! Answer the experiment by only giving the letter of the answer options (e.g. A, B, C, ...) or a numerical value (80, 100, 1000, ...). Do not state anything else! Do not hallucinate."

# Question
question: str = "Imagine that you have decided to see a play where admission is $10. As you enter the theater, you discover that you have lost a $10 bill. Would you still pay $10 for a ticket to the play? Choose between A) Yes or B) No. __"

# Tell the model what type of response we are expecting
if response_type == "choice":
    output_message: str = "Your output should only be a LETTER (A, B, C, ...)."
else:
    output_message: str = "Your output should only be a NUMBER (9, 80, 100, 1000, ...)."
```

```
# Combine messages
entire_user_message: str = (
    user_message
    + "\n---------------------\n"
    + question
    + "\n---------------------\n"
    + output_message
)

####-- MODEL INITIALIZATION --####
# Check if model is downloaded, else pull
try:
    ollama.show(model)
except ollama.ResponseError:
    ollama.pull(model)

# Initialize the model
model_interactor: Ollama = Ollama(
    model=model,
    temperature=temperature,
    system_prompt=system_message,
    additional_kwargs={
        "num_predict": max_tokens,
    },
)

####-- EXPERIMENT --####
# Store responses and whether they are in the correct format
responses: List[str] = [""] * n
correct_runs: List[int] = [0] * n

# Run the experiment
for i in range(n):
    # Prompt the model and store the response
    response: str = str(model_interactor.complete(entire_user_message)).strip()

    # Make sure if there are letters, it is only one letter
    if (
        (len(response) == 1 and response.isalpha())  # Single letter
        or str(response).isdigit()  # Valid number
    ):
        correct_run: int = 1
    else:
        correct_run: int = 0

    responses[i] = response
    correct_runs[i] = correct_run

# Close the model interactor
ollama.generate(model=model, prompt="Goodbye!", keep_alive=0)
```

D Sampling variance and mean aggregation

Formula for the (approximated) sampling variance $\sigma^2_{e_i}$ of Cohen's d (Borenstein et al., 2021; Goulet-Pelletier & Cousineau, 2018; Morris & DeShon, 2002):

$$\sigma^2_{e_i} = \frac{\nu}{\nu - 2} \times \frac{2}{\tilde{n}} \left(1 + \frac{\tilde{n}}{2}\delta^2\right) - \frac{\delta^2}{J(\nu)^2} \tag{6}$$

where:

δ: unweighted population effect size $= \frac{\sum_{i=1}^{n} d_i}{n}$

$J(\nu)$: Hedges correction factor $= \frac{\Gamma(\frac{1}{2}\nu)}{\sqrt{\frac{\nu}{2}}\,\Gamma(\frac{1}{2}(\nu-1))} \approx 1 - \frac{3}{4\nu-1}$ (Hedges, 1981)

$\tilde{n}$: harmonic mean of the sample sizes $= \frac{n_{pre} \times n_{post}}{n_{pre} + n_{post}}$

N: total number of observations $= n_{pre} + n_{post}$

ν: degrees of freedom $= n_{pre} + n_{post} - 2$

Formula for the variance due to the sampling error $\hat{\sigma}^2_e$ (Morris & DeShon, 2002):

$$\hat{\sigma}^2_e = \frac{n}{\sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2_{e_i}}} \tag{7}$$

where $\hat{\sigma}^2_{e_i}$: approximated sampling variance of Cohen's d

Formula for the weighted mean aggregation of Cohen's d (Borenstein et al., 2021; Hedges & Olkin, 1985; Morris & DeShon, 2002):

$$\bar{d} = \frac{\sum_{i=1}^{n} w_i d_i}{\sum_{i=1}^{n} w_i} \tag{8}$$

where:

w_i: weights as reciprocals of approximated sampling variance $= \frac{1}{\hat{\sigma}^2_{e_i}}$

d_i: Cohen's d effect size

Similarly, the formula for the observed variance $\hat{\sigma}^2_d$ across effect sizes (Morris & DeShon, 2002):

$$\hat{\sigma}^2_d = \frac{\sum_{i=1}^{n} w_i(d_i - \bar{d})}{\sum_{i=1}^{n} w_i} \tag{9}$$

where:

w_i: weights as reciprocals of approximated sampling variance $= \frac{1}{\hat{\sigma}^2_{e_i}}$

d_i: Cohen's d effect size

$\bar{d}$: mean of Cohen's d effect sizes

If $n_{pre} = n_{post}$ are constant across all effect sizes, the sampling variance $\sigma^2_{e_i}$ and the weights w_i are also constant. Thus, the mean aggregation $\bar{d}$ is the same as the unweighted mean. The same applies for the observed variance $\hat{\sigma}^2_d$:

$$\bar{d} = \frac{\sum_{i=1}^{n} w_i d_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} d_i}{n} \tag{10}$$

$$\hat{\sigma}^2_d = \frac{\sum_{i=1}^{n} w_i (d_i - \bar{d})}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} (d_i - \bar{d})}{n} \tag{11}$$

if $n_{pre} = \bar{n}_{pre}$ and $n_{post} = \bar{n}_{post}$

E Homogeneity anomalies

In the following table, we list exceptions to the analysis that mostly non-bias detections were homogeneous and bias detections were heterogeneous. Anomalies include bias and model combinations with one of the following detection-homogeneity pairing:

- homogeneous detection: b.d. (capped) > 0.3 & homogeneity (capped) > 0.3
- heterogeneous non-detection: b.d. (capped) < 0.3 & homogeneity (capped) < 0.3

bias	**model**	**b.d. (capped)**	**homogeneity (capped)**
Anchoring	phi3:medium	1.000000	0.90121
Anchoring	llama3.1:70b	1.000000	0.493774
Framing Effect	claude-3.5-sonnet	0.717996	0.404175
Framing Effect	gemma2:27b	0.143622	0.231243
Gambler's Fallacy	gemma2	0.191709	0.161869
Category Size Bias	llama3.1	0.000000	0.240563

Table 8: *Anomalies of homogeneity calculations*

Note: *b.d. (capped)* refers to the capped bias detection.

F Regression results

F.1 Regressions with experiment variables

Table 9: *Fixed effects regression results of all experiment variables on bias detected (where bias, model, and scenario are fixed effects)*

Dep. Variable:	bias_detected	**R-squared:**	0.072
Model:	OLS	**Adj. R-squared:**	0.060
Method:	Least Squares	**F-statistic:**	6.092
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	6.61e-16
Time:	22:17:33	**Log-Likelihood:**	-5923.3
No. Observations:	1599	**AIC:**	1.189e+04
Df Residuals:	1578	**BIC:**	1.200e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	3.9224	1.197	3.276	0.001	1.574	6.271
C(bias)[T.category size bias]	-7.3005	0.991	-7.368	0.000	-9.244	-5.357
C(bias)[T.endowment effect]	-2.0941	0.990	-2.116	0.034	-4.035	-0.153
C(bias)[T.framing effect]	-2.7081	0.990	-2.737	0.006	-4.649	-0.767
C(bias)[T.gamblers fallacy]	-3.2171	0.990	-3.251	0.001	-5.158	-1.276
C(bias)[T.loss aversion]	-2.0655	0.990	-2.087	0.037	-4.007	-0.125
C(bias)[T.sunk cost fallacy]	-3.2323	0.990	-3.266	0.001	-5.173	-1.291
C(bias)[T.transaction utility]	-2.9421	0.990	-2.973	0.003	-4.883	-1.001
C(model)[T.claude-3.5-sonnet]	1.8203	1.106	1.645	0.100	-0.350	3.990
C(model)[T.gemma2]	1.2924	1.106	1.168	0.243	-0.878	3.463
C(model)[T.gemma2:27b]	0.0876	1.106	0.079	0.937	-2.082	2.258
C(model)[T.gpt-4o]	0.9837	1.106	0.889	0.374	-1.186	3.154
C(model)[T.gpt-4o-mini]	1.1737	1.106	1.061	0.289	-0.996	3.344
C(model)[T.llama3.1]	-1.0772	1.106	-0.974	0.330	-3.247	1.093
C(model)[T.llama3.1:70b]	0.1346	1.108	0.121	0.903	-2.039	2.308
C(model)[T.phi3.5]	-4.3647	1.106	-3.945	0.000	-6.535	-2.195
C(model)[T.phi3:medium]	-0.1172	1.106	-0.106	0.916	-2.287	2.053
C(scenario)[T.1_no_persona]	-2.1589	0.700	-3.085	0.002	-3.531	-0.786
C(scenario)[T.2_odd_numbers]	0.5583	0.700	0.797	0.425	-0.815	1.932
C(scenario)[T.3_large_numbers]	-0.5939	0.700	-0.849	0.396	-1.966	0.779
temperature	-0.1271	0.458	-0.278	0.781	-1.025	0.771

Omnibus:	2620.002	**Durbin-Watson:**	1.986
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3123791.020
Skew:	-10.172	**Prob(JB):**	0.00
Kurtosis:	218.574	**Cond. No.**	16.8

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 10: *Fixed effects regression results of all experiment variables on bias detected (capped)*
(where bias, model, and scenario are fixed effects)

Dep. Variable:	bias_detected_capped	**R-squared:**	0.367
Model:	OLS	**Adj. R-squared:**	0.359
Method:	Least Squares	**F-statistic:**	45.78
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	6.81e-141
Time:	22:17:33	**Log-Likelihood:**	-409.96
No. Observations:	1599	**AIC:**	861.9
Df Residuals:	1578	**BIC:**	974.8
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	0.6241	0.038	16.386	0.000	0.549	0.699
C(bias)[T.category size bias]	-0.6212	0.032	-19.711	0.000	-0.683	-0.559
C(bias)[T.endowment effect]	-0.2625	0.031	-8.338	0.000	-0.324	-0.201
C(bias)[T.framing effect]	-0.4660	0.031	-14.805	0.000	-0.528	-0.404
C(bias)[T.gamblers fallacy]	-0.6693	0.031	-21.263	0.000	-0.731	-0.608
C(bias)[T.loss aversion]	-0.3604	0.031	-11.449	0.000	-0.422	-0.299
C(bias)[T.sunk cost fallacy]	-0.6831	0.031	-21.703	0.000	-0.745	-0.621
C(bias)[T.transaction utility]	-0.5774	0.031	-18.344	0.000	-0.639	-0.516
C(model)[T.claude-3.5-sonnet]	0.0938	0.035	2.664	0.008	0.025	0.163
C(model)[T.gemma2]	-0.0351	0.035	-0.999	0.318	-0.104	0.034
C(model)[T.gemma2:27b]	-0.0649	0.035	-1.845	0.065	-0.134	0.004
C(model)[T.gpt-4o]	0.1783	0.035	5.067	0.000	0.109	0.247
C(model)[T.gpt-4o-mini]	0.0551	0.035	1.567	0.117	-0.014	0.124
C(model)[T.llama3.1]	-0.0240	0.035	-0.682	0.496	-0.093	0.045
C(model)[T.llama3.1:70b]	0.0046	0.035	0.130	0.896	-0.065	0.074
C(model)[T.phi3.5]	-0.0255	0.035	-0.725	0.468	-0.095	0.043
C(model)[T.phi3:medium]	-0.0418	0.035	-1.188	0.235	-0.111	0.027
C(scenario)[T.1_no_persona]	0.0081	0.022	0.363	0.716	-0.036	0.052
C(scenario)[T.2_odd_numbers]	-0.0006	0.022	-0.028	0.977	-0.044	0.043
C(scenario)[T.3_large_numbers]	-0.0734	0.022	-3.300	0.001	-0.117	-0.030
temperature	0.0692	0.015	4.750	0.000	0.041	0.098

Omnibus:	99.364	**Durbin-Watson:**	2.064
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	117.878
Skew:	0.624	**Prob(JB):**	2.53e-26
Kurtosis:	3.459	**Cond. No.**	16.8

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

F.2 Regressions solely with scenarios

Table 11: *Fixed effects regression results solely of scenarios on bias detected (where scenario is fixed effect)*

Dep. Variable:	bias_detected	**R-squared:**	0.010
Model:	OLS	**Adj. R-squared:**	0.008
Method:	Least Squares	**F-statistic:**	5.345
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	0.00116
Time:	22:17:33	**Log-Likelihood:**	-5974.8
No. Observations:	1599	**AIC:**	1.196e+04
Df Residuals:	1595	**BIC:**	1.198e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.8436	0.508	1.660	0.097	-0.153	1.840
C(scenario)[T.1_no_persona]	-2.1589	0.719	-3.004	0.003	-3.569	-0.749
C(scenario)[T.2_odd_numbers]	0.5688	0.719	0.791	0.429	-0.842	1.979
C(scenario)[T.3_large_numbers]	-0.5939	0.719	-0.826	0.409	-2.004	0.816

Omnibus:	2698.337	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3572326.621
Skew:	-10.805	**Prob(JB):**	0.00
Kurtosis:	233.546	**Cond. No.**	4.79

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 12: *Fixed effects regression results solely of scenarios on bias detected (capped) (where scenario is fixed effect)*

Dep. Variable:	bias_detected_capped	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	0.005
Method:	Least Squares	**F-statistic:**	3.803
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	0.00988
Time:	22:17:33	**Log-Likelihood:**	-770.11
No. Observations:	1599	**AIC:**	1548.
Df Residuals:	1595	**BIC:**	1570.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.2523	0.020	12.868	0.000	0.214	0.291
C(scenario)[T.1_no_persona]	0.0081	0.028	0.292	0.771	-0.046	0.062
C(scenario)[T.2_odd_numbers]	-0.0001	0.028	-0.005	0.996	-0.055	0.054
C(scenario)[T.3_large_numbers]	-0.0734	0.028	-2.649	0.008	-0.128	-0.019

Omnibus:	266.007	**Durbin-Watson:**	2.042
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	411.605
Skew:	1.236	**Prob(JB):**	4.18e-90
Kurtosis:	2.738	**Cond. No.**	4.79

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

F.3 Regressions with model features

Table 13: *Regression results of the model features on bias detected*

Dep. Variable:	bias_detected	**R-squared:**	0.024
Model:	OLS	**Adj. R-squared:**	0.019
Method:	Least Squares	**F-statistic:**	4.930
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	4.85e-06
Time:	22:17:40	**Log-Likelihood:**	-5963.2
No. Observations:	1599	**AIC:**	1.194e+04
Df Residuals:	1590	**BIC:**	1.199e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-24.4900	6.469	-3.786	0.000	-37.179	-11.801
temperature	-0.1295	0.468	-0.277	0.782	-1.047	0.788
release_date_diff	0.0152	0.023	0.648	0.517	-0.031	0.061
last_updated_date_diff	-0.0020	0.028	-0.072	0.943	-0.056	0.052
training_data_cutoff_date_diff	-0.0046	0.006	-0.819	0.413	-0.016	0.006
number_parameters	-0.0050	0.013	-0.371	0.710	-0.031	0.021
context_length	-0.0012	0.016	-0.076	0.940	-0.033	0.030
mmlu	0.0554	0.113	0.491	0.624	-0.166	0.276
chatbot_arena	0.0168	0.008	2.229	0.026	0.002	0.032

Omnibus:	2659.529	**Durbin-Watson:**	1.557
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3410054.454
Skew:	-10.478	**Prob(JB):**	0.00
Kurtosis:	228.263	**Cond. No.**	3.31e+04

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.31e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 14: *Regression results of the model features on bias detected (capped)*

Dep. Variable:	bias_detected_capped	**R-squared:**	0.041
Model:	OLS	**Adj. R-squared:**	0.036
Method:	Least Squares	**F-statistic:**	8.410
Date:	Sat, 04 Jan 2025	**Prob (F-statistic):**	3.03e-11
Time:	22:17:40	**Log-Likelihood:**	-742.67
No. Observations:	1599	**AIC:**	1503.
Df Residuals:	1590	**BIC:**	1552.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-0.0198	0.247	-0.080	0.936	-0.505	0.465
temperature	0.0691	0.018	3.867	0.000	0.034	0.104
release_date_diff	-0.0006	0.001	-0.680	0.497	-0.002	0.001
last_updated_date_diff	0.0005	0.001	0.470	0.638	-0.002	0.003
training_data_cutoff_date_diff	0.0004	0.000	2.077	0.038	2.5e-05	0.001
number_parameters	0.0015	0.001	2.885	0.004	0.000	0.002
context_length	-0.0004	0.001	-0.590	0.555	-0.002	0.001
mmlu	-0.0065	0.004	-1.519	0.129	-0.015	0.002
chatbot_arena	0.0004	0.000	1.517	0.130	-0.000	0.001

Omnibus:	250.948	**Durbin-Watson:**	0.739
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	382.049
Skew:	1.193	**Prob(JB):**	1.09e-83
Kurtosis:	2.794	**Cond. No.**	3.31e+04

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.31e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Formal declaration

I hereby declare that I have written this thesis independently, did not use any sources or resources other than those cited and that the thesis has not been submitted as a whole or in any significant part as part of any other examination process. All information taken from other works - either verbatim or paraphrased - has been clearly indicated. The copy submitted in electronic form is identical in content to the bound copies submitted.

Munich, January 15, 2025

Max Mohr