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
Grünwalder Straße 14
81547 Munich

max.mohr@student.uni-tuebingen.de

M.Sc. Data Science in Business and Economics
5th semester
Matriculation number: 6304784

Date of submission: January 29, 2025

Abstract

Xxx

Contents

List of Abbreviations

XXX	Xxx
XXX	Xxx

List of Tables

List of Figures

1 Introduction

Xxx

2 Theoretical background

2.1 Past studies on human behavioral effects

Humans are constantly exposed to decision making. Decisions can vary between very simple and complex ones. In studying the decision processes of humans, researchers started seeing the human species as a rational species that makes decisions based on logic and reasoning (Juárez Ramos, 2018). However, gaps in these theories such as missing information access were identified quickly. This led to the development of the bounded rationality theory by Herbert Simon simon1955behavioral. The theory suggests that humans are not always rational and that they make decisions based on the information available to them. This theory was further developed by Daniel Kahneman and Amos Tversky, who introduced the concept of cognitive biases kahneman1974judgment. Cognitive biases are systematic errors in thinking that affect the decisions and judgments that people make. It has been estimated that 70% of all decisions by humans are affected by cognitive biases (Juárez Ramos, 2018).

”Relationship between Cognitive Biases in Decision-Making” (Yeung et al., 2023)
”Effect Of Select Cognitive Biases On Financial And General Decision Making” (Gupta, 2018)
”The Impact of Cognitive Biases on Professionals’ Decision-Making: A Review of Four Occupational Areas” (Berthet, 2022)

2.2 Leveraging large language models to simulate human behavior

A) Recent developments in large language model
B) The exposure of human behavioral patterns in the models
How could models pick up biases? A) Data (texts of humans e.g.), B) Training and Learning (RLHF)

”Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption” (Talboy and Fuller, 2023)

Does exactly what we want to do

”Questioning the Survey Responses of Large Language Models” (Dominguez-Olmedo et al., 2023)

”Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT” (Hagendorff et al., 2023)

”Cognitive bias in large language models: Cautious optimism meets anti-Panglossian meliorism” (Thorstad, 2023)

2.3 Meta analysis techniques

3 Methodology

3.1 Experiments

3.1.1 Studies

Some are choice experiments, some expect a number and then compare the answers between different questioning types.

3.1.2 Scenarios

- Normal (replication of original study)
- Random values
- Explicitly prompt to behave humanlike

Also describe how the normal prompt is structured.

3.2 Bias selections

As described, previous research revolving around cognitive biases has shown that there are numerous biases influencing human behavior. Thus, we have to sample a concise yet ideally comprehensive sample of biases to explore more generalizable analysis results. To narrow down the range of biases, we focus on biases affecting (economic) decision-making processes. In particular, we aim to recreate and perhaps slightly modify existing experiments.

Anchoring bias The bias known as anchoring is often used to explain why people tend to rely heavily on an initial piece of information (”anchor”) in their decision-making afterwords. 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 is also influenced by the way alternatives are categorized and distributed, a phenomenon known as the category size bias (Isaac & Brough, 2014). For example, an investor’s expectation about the performance of a particular stock could be influenced by the number of other stocks in the portfolio that belong to the same industry. Similarly, Tversky and Koehler (1994) showed that participants judged the probability of dying of unnatural cases differently when natural causes were 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 output 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 if they did not own them is known as the endowment effect. This effect is strongly linked to loss aversion, where people demand more to give up an item than they would be willing to pay to acquire it (Kahneman et al., 1990). Research has also shown that the effect is independent of whether sellers actually 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 for a mug versus their willingness to accept compensation for the mug. The results showed that participants owning the mug valued it at more than double the value than the other participants.

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 on framing effect as well as possible from the experiment on loss aversion, we target to simulate two scenarios that both have losses (not comparison between loss and gain framing). We recreate the experiment by Tversky and Kahneman (1981) where participants were asked to buy another concert ticket after having lost either the original 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 as from the human perspective, the lost money is decoupled from the ticket purchase.

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. While there has been research on the mechanisms influencing the fallacy (for example 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 coin flips (similar to the Monte Carlo fallacy).

Loss aversion Loss aversion 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 impact 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 sale strategies (discount for additional spending), financial investments (sell winners, hold losers) and more (Y. Liu, 2023). We recreate the experiment from R. H. 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 gain scenario than in the loss scenario.

Sunk cost fallacy This effect refers to the human phenomenon to 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. 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 perceived psychological non-monetary gains of a deal that are beyond the actual economic value. R. 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 20 percent discount. The difference in the scenarios is that one product is cheap (radio) and the other is expensive (television) (R. Thaler, 1983). The results showed that participants were more likely to buy the cheaper product directly without the discount even though the absolute savings were the same.

With the selected biases, we aim to examine the cognitive decision behavior of large language models as broadly as possible. The exact replications and wordings of the experiments are described in the appendix.

3.3 Model selections

The development and publication of new models and model architectures is ongoing and rapid. We focus on models which are widely available to the public and have seen significant use in research and industry. The models we selected are those of major technology companies and have some similarities but also differences in their architecture, size and training data. We aim to include models from different companies to ensure a broad analysis of the models' behavior. The models we selected are:

Gemma2 (Google[1]) The Gemma model family is Google's offering of open language models. The latest and second generation Gemma models were 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 two, seven and 25 billion parameter sizes (Team et al., 2024). For our research, we examine

[1] Google, founded in 1998, is a global technology company specializing in internet-related services and products, including search engines, online advertising, and AI; in 2023, its parent company Alphabet reported a revenue of over $305 billion (Alphabet, 2023).

the reasoning capabilities and existence of biases in the seven and 27 billion parameter versions.

GPT-4o (OpenAI[2]) With the introduction and rash success of OpenAI's artificial intelligence chatbot *ChatGPT* in 2022, the models backing the chatbot (GPTs) have a significant impact on the field of natural language processing and in communicating the latest advancements in AI to the public. The latest iteration are 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 however 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 (Achiam et al., 2023; OpenAI, 2024c). 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, 2024b).

Llama3.1 (Meta[3]) Llama3.1 is Meta's latest iteration of their open-source language model family, following the success of Llama2 and 3. Meta's offering of these foundational models has significant impact in building and finetuning personalized multilingual solutions and for any research purpose. Llama 3.1 models come with much longer context length and more as well as higher quality training data than previous models while maintaining the similar transformer architecture of previous iterations. While we will analyze the eight and 70 billion parameter models, a 405 billion parameter version is also available and rivals the largest closed-source models (Dubey et al., 2024; Meta, 2024b).

Phi3.5 (Microsoft[4]) The Phi models by Microsoft are considered SLMs (small language models) and focus on efficiency and speed. While this does lead to less factual

[2] OpenAI, founded in 2015 as a non-profit but now a capped for-profit, is a leading AI research and deployment company. OpenAI's primary products include the GPT series of language models as well as text-to-image and text-to-vision models (OpenAI, 2024a).

[3] Meta, formerly known as Facebook, is a multinational technology conglomerate founded in 2004. It rebranded as Meta in 2021 to reflect its expansion into the metaverse and artificial intelligence. In 2023, Meta reported a revenue of $134 billion (Meta, 2024a).

[4] Microsoft, founded in 1975, is a global technology company known for its software products (revenue of $211 billion). With recent large investments such as in OpenAI, Microsoft plays a significant role in the AI space (Microsoft, 2024).

knowledge in the models, they are still competitive regarding reasoning and coding tasks even when implemented and running on phones. There are multiple sizes and different architectures of the Phi models to account for multilingual or multimodal inputs. In particular, Phi 3.5 has a much expanded context length limit from 4 to 128 thousand tokens. We analyze the new Phi 3.5 Mini model (roughly 4 billion parameters) and the slightly larger Phi 3 Medium model (14 billion parameters). With these smaller models we hope to cover not only the spectrum of large language models but also small ones (Abdin et al., 2024).

In total, we analyze four model families and eight distinct models. Besides the GPT models of OpenAI, 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 local 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 configurations are provided in the appendix.

3.4 Response analysis

Somewhere describe what the expected output of the models should look like (structured prediction) and what I do if it is different.

3.4.1 Analysis of detected biases

Original studies and compare results. Perhaps a ”bias detected” number. If always 100 experiment runs, bias detected is the percentage of runs where the model acted biased. (Between 0 and 1, perhaps normalize) Perhaps Chi-Square Test of Independence to see if there are significant differences between the biases.

3.4.2 Scenario impact

Average treatment effects of treated on randomized values as well as explicitly prompting humanlike behavior. The control group is the normal prompt. Perhaps with Repeated Measures ANOVA (RM-ANOVA) to see if there are significant differences between the scenarios. Perhaps Mixed-Effects Models to see if there are significant differences between the scenarios. Friedman test? https://datatab.de/tutorial/friedman-test

3.4.3 Model analysis

Are there any trends between newer/larger models? Temperature? Ggf nur für das normale Szenario da es sonst die Runs sprengt

Train XGBoost model, get feature importances and see if there are important features that can explain whether a model is more biased or less.

4 Results

Xxx

4.1 Analysis of detected biases

Xxx

4.2 Scenario impact

Xxx

4.3 Model analysis

Xxx

4.4 Interactions

Xxx

5 Discussion and outlook

Xxx

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*.

Alphabet. (2023). Alphabet annual report 2023, 63. https://abc.xyz/assets/5a/ae/29f710e646b49ee3d6b63c4dc3a0/goog-10-k-2023.pdf

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

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.

Berthet, V. (2022). The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in psychology*, *12*, 802439.

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*.

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

Gupta, D. (2018). Effect of select cognitive biases on financial and general decision making. *Proceedings of International Academic Conferences*, (7010051).

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, *3*(10), 833–838.

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.

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

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.

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

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

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.

Meta. (2024a). About meta [Last accessed: September 11, 2024]. https://about.meta.com/company-info/

Meta. (2024b, 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/

Microsoft. (2024). Microsoft about [Last accessed: September 11, 2024]. https://www.microsoft.com/en-us/about

OpenAI. (2024a). About openai [Last accessed: September 10, 2024]. https://openai.com/about/

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

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*.

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

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

Thorstad, D. (2023). Cognitive bias in large language models: Cautious optimism meets anti-panglossian meliorism. *arXiv preprint arXiv:2311.10932*.

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.

Yeung, Y., et al. (2023). Relationship between cognitive biases in decision-making. *Journal of Sociology and Ethnology*, *5*(6), 28–32.

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 29, 2025

Max Mohr