Examining Cognitive Biases in ChatGPT 3.5 and 4 through Human Evaluation and Linguistic Comparison

Giada Pantana giada.pantana@edu.unige.it
Department of Modern Languages and Cultures, University of Genoa, Genoa, IT
Marta Castello marta.castello@creative-words.com
Creative Words, Genoa, IT
Ilaria Torre ilaria.torre@unige.it
Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genoa, Genoa, IT

Abstract

This paper aims to investigate the presence of cognitive biases, more specifically of Availability heuristics, Representativeness heuristics and Framing, in OpenAI's ChatGPT 3.5 and ChatGPT 4, as well as the linguistic dependency of their occurrences in the Large Language Models' (LLMs) outputs. The innovative aspect of this research is conveyed by rephrasing three tasks proposed in Kahneman and Tversky's works and determining whether the LLMs' answers to the tasks are correct or incorrect and human-like or non-human-like. The latter classification is made possible by interviewing a total of 56 native speakers of Italian, English and Spanish, thus introducing a new linguistic comparison of results and forming a "human standard". Our study indicates that GPTs 3.5 and 4 are very frequently subject to the cognitive biases under discussion and their answers are mostly non-human-like. There is minimal but significant discrepancy in the performance of GPT 3.5 and 4, slightly favouring ChatGPT 4 in avoiding biased responses, specifically for Availability heuristics. We also reveal that, while the results for ChatGPT 4 are not significantly language dependent, meaning that the performances in avoiding biases are not affected by the prompting language, their difference with ChatGPT 3.5 is statistically significant.

1 Introduction

In the last years, Large Language Models (LLMs) have been used exponentially thanks to their capabilities to be queried with natural language and to return content- and context-aware responses. They became popular within the general public, and businesses swiftly introduced these models in their workflow aiming at becoming more productive, while reducing employees' workload. Natural language itself is not only our easiest and quickest way to communicate to these language models, but also the main reason why we tend to anthropise these machines (Roberts and al., 2024), making our relationship with them resonate widely and strongly in our everyday life. Nonetheless, "LLMs simply do not have the capacity to distinguish between truth and falsehood and, therefore, without malicious intent, [they] can confidently present fictions as if they were truths" (Roberts and al., 2024, p. 4). For this reason, we have the responsibility to prove if, how and when they are most reliable. Much work has been done in delicate fields such as legal, medical and educational (Schmidgall and al., 2024; Pal and al., 2023; Pal, 2024; Curran and al., 2023; Gutiérrez-Cirlos and al., 2023; Ji and al., 2023b) to analyse how to improve their use in the professionals' decision-making process and to help users make more conscious choices. When only taking the outputs into consideration, the main hindrance to their implementation into businesses and field-specific tasks are hallucinations, defined as "the generated content that is nonsensical or unfaithful to the provided source content" (Ji and al., 2023a, p. 4). Hallucinations are defined as in-

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

trinsic, when the output contradicts the source, or extrinsic, when the output cannot be verified from the source (Ji and al., 2023a). Given the potentially harmful and often subtle nature of this phenomenon, researchers have developed various hallucination mitigation techniques. These techniques operate at different levels of the LLM's functioning to help reduce their occurrence. Addressing the issue can involve either prompt engineering or developing models to reduce the elicitation of hallucinations. Specific actions are available for each level of the LLM's functioning (Tonmoy and al., 2024).

Alongside the phenomena that hinder menaces, the utility of LLMs is threatened by the presence of cognitive biases in their outputs. As hallucinations, cognitive biases are concepts mainly used to describe human behaviours and have been adapted to this field to define machines' dysfunctions. Generative Artificial Intelligence (GenAI) can exhibit biases due to various factors. Some of the key causes are that LLMs are trained on human-made data, including historical data. They can be skewed and affected by under/over-representation of certain phenomena. Moreover, biases can be introduced in the process of data annotation and refinement, often based on Reinforced Learning with Human Feedback (RLHF) (Christiano and al., 2023; Chen and al., 2023; Navigli and al., 2023) and propagate into the models. Thus, machines can potentially inherit and enhance human cognitive biases.

The main focus of this article is to verify whether ChatGPT 3.5 and ChatGPT 4 are susceptible to three specific cognitive biases, known in the literature as *Availability heuristics*, *Representativeness heuristics* and *Framing* (Tversky and Kahneman, 1973, 1983, 1981). The Dictionary of Psychology issued by the American Psychological Association defines them as follows. *Availability heuristics* is: "a common strategy for making judgments about likelihood of occurrence in which the individual bases such judgments on the salience of the information held in their memory about the particular type of event"[1]. *Representativeness heuristics* is: "a strategy for making categorical judgments about a given person or target based on how closely the exemplar matches the typical or average member of the category"[2]. Finally, *Framing* is: "the process of defining the context or issues surrounding a question, problem, or event in a way that serves to influence how the context or issues are perceived and evaluated"[3].

A recent trend in Generative Artificial Intelligence literature is "machine psychology" by Thilo Hagendorff, meaning that the LLM is positioned by the researchers as the subject of their psychological tests, initially designed to investigate human misbehaviour (Hagendorff, 2023). According to this approach, LLMs are tested for cognitive biases using their chatbot interfaces. The tools mainly investigated for tracking cognitive biases are Open AI's ChatGPT (version 3.5, 3.5 Turbo and 4), Google's Gemini, Anthropic's Claude in different versions and Llama in different versions (Macmillan-Scott and M., 2024; Azaria, 2023; Chen and al., 2023; Schmidgall and al., 2024). Most studies refer to the tasks proposed by Kahneman and Tversky to test cognitive biases (Macmillan-Scott and M., 2024; Azaria, 2023; Chen and al., 2023; Kliegr and al., 2021) and have unmasked that LLMs are, in fact, victims of these biases, including but not limited to: Confirmation bias (Macmillan-Scott and M., 2024; Chen and al., 2023; Kliegr and al., 2021; Berberette and al., 2024; Ke and al., 2024; Dos Santos and Cury, 2023; Schmidgall and al., 2024), Availability heuristics (Azaria, 2023; Chen and al., 2023; Kliegr and al., 2021; Berberette and al., 2024), Overconfidence (Chen and al., 2023; Kliegr and al., 2021; Schmidgall and al., 2024), Representativeness heuristics (Macmillan-Scott and M., 2024; Chen and al., 2023; Kliegr and al., 2021), Framing (Azaria, 2023; Chen and al., 2023), Recency effect (Berberette and al., 2024; Schmidgall and al., 2024). The primary methodology described in these studies either directly or indirectly refers to the "machine psychology" approach (Hagendorff, 2023).

This paper addresses the following research questions:

1. Do ChatGPT 3.5 and 4 show *Availability heuristics*, *Representativeness heuristics* and *Framing* biases in their outputs?

2. Are there any differences in the performance outcomes of ChatGPT 3.5 and ChatGPT 4?

[1] https://dictionary.apa.org/availability-heuristic, last access: 3/14/2024

[2] https://dictionary.apa.org/representativeness-heuristic, last access: 3/14/2024

[3] https://dictionary.apa.org/framing, last access: 3/14/2024

3. Are the two ChatGPT models language dependent in reporting the aforementioned biases?

Based on the results of previous research, it can be supposed that: - LLMs show biases like *Availability heuristics* (Berberette and al., 2024; Kliegr and al., 2021; Azaria, 2023), *Representativeness heuristics* (Macmillan-Scott and M., 2024; Chen and al., 2023; Kliegr and al., 2021) and *Framing* (Chen and al., 2023; Azaria, 2023) in their outputs;
- there are differences between ChatGPT 3.5 and 4, where 4 should be less subject to biases since it is trained on more data, or at least gives better performances according to OpenAI (OpenAI et al., 2024);
- LLMs should be language dependent since the data with which they are trained differ among languages, causing different performances, or at least their results in Massive Multitask Language Understanding (MMLU) are, despite a minimal difference, better in English than Italian or Spanish (OpenAI et al., 2024).

2 Methodology

In consideration of recent literature, the aim is to analyse the biases of *Availability heuristics*, *Representativeness heuristics* and *Framing* in two LLMs, namely ChatGPT 3.5 and 4[4]. This will be done by introducing a new rephrasing approach to three specific Tversky and Kahneman tests, namely the Judgement of word frequency (Tversky and Kahneman, 1973) to demonstrate *Availability heuristics*, the Linda problem (Tversky and Kahneman, 1983) to demonstrate *Representativeness heuristics* and the Framing of Contingencies (Tversky and Kahneman, 1981) to demonstrate *Framing*.
The presence of biases in LLMs outputs was examined by classifying the answers of the LLMs according to the methodology proposed by Olivia MacMillan-Scott and Mirco Musolesi in their paper (Macmillan-Scott and M., 2024), using four parameters: correct/incorrect and human-like/non-human-like. To support the definition of what is human-like, a pool of 56 people (59% women, 41% men, age average: 33) was interviewed, defined by the availability of resources, yet guaranteeing the representativeness of the sample, posing the same questions asked to the two ChatGPT models. Ethics approval was not obtained since the research did not involve sensitive personal information or interventions that required formal ethical oversight. Additionally, to verify whether the answers of the models, as well as their potential biases, are language dependent, a multilingual analysis was conducted by prompting the LLMs in Italian, English and Spanish. The innovative aspect of this work lies in the multilingual comparison between human native speakers and LLM data, setting it apart from previously analysed reference material.
To answer the research questions, we followed the methodology described in the subsequent sections. All data regarding the complete prompts and results are available in a public GitHub repository. [5]

2.1 LLM tools

Open AI's ChatGPT 3.5 and 4 were chosen for this research paper as they are among the most commonly used LLMs in the literature regarding the testing on cognitive biases (Macmillan-Scott and M., 2024; Azaria, 2023; Chen and al., 2023; Berberette and al., 2024; Ke and al., 2024; Dos Santos and Cury, 2023; Schmidgall and al., 2024). Chat GPT 3.5 was chosen because it is free and therefore widely accessible; Chat GPT 4, expected to have better performance (OpenAI et al., 2024), was analysed to determine if it provides potentially less biased answers than the former model.
The models were not customised or specifically trained. The chatbot interface was used to test the prompts and obtain the answers. Zero-shot prompting (Kojima and al., 2022) was applied to address the LLMs, at times integrated with an iterative approach to elicit a unique and definite answer from the machines. The prompt testing for the LLMs was carried out from 15th March to 10th April 2024. The prompting texts are described below and reported in Table 1 in the Appendix.

2.2 Prompt definition

One prompt was tested for each bias: the Judgement of word frequency for *Availability heuristics*, the Linda problem for *Representativeness heuristics* and the Framing of Contingencies for *Framing*. Each prompt was tested in three languages: Italian, English and Spanish. Below, the methodology for

[4]https://chat.openai.com/auth/login, last access 4/10/2024
[5]https://github.com/CreativeWords/Cognitive_Bias_GPT

defining the three tests and how the same prompt was formulated in the three different languages under scrutiny will be explained.

All the prompts tested were re-elaborated from pre-existing psychology tests originally targeting human subjects. This paper focuses on addressing LLMs, instead, following the "machine psychology" approach (Hagendorff, 2023). The tests were rephrased in respect to the originals proposed by Tversky and Kahneman to avoid the risk of them being part of the training data of the LLMs, thus minimizing any potentially compromised performance that would have prevented our detection of real biases. As Thilo Hagendorff explains in his paper (Hagendorff, 2023), it is essential to ensure that the machine has not seen the test before. Given the limited information available on the training dataset, this can be guaranteed by reformulating the questions with new components while preserving the original logical structure. Regarding the languages involved, it necessary to ensure the accuracy and reliability of the translations. To achieve this, the initial drafts of the three prompts were created in Italian by native speakers. Professional translators and native speakers of each language were assigned to translate the prompts into English and Spanish. The three final prompts were subsequently used to query the LLMs. The prompts were submitted 56 times to GPT 3.5 and 56 times to GPT 4, of which 27 times using the Italian prompt, 11 times the English prompt and 18 times the Spanish prompt. This was done to ensure the LLM models were prompted as many times as the human pool (56 people), allowing an effective comparative evaluation. Another reason behind this choice was the need to minimise the chances of randomness (Macmillan-Scott and M., 2024). Each task was prompted in a new, empty chat each time to avoid any occurrence of recency effect[6] (Macmillan-Scott and M., 2024). All prompts are reported in the Appendix in Table 1 following the same logic: name of the test, name of the bias to test, original question by Tversky and Kahneman, and English translation of the prompt. The Italian and Spanish versions are reported in the full repository of data on GitHub. The first prompt in Table 1 was used to test the *Availability heuristics* cognitive bias. Words were tested instead of single letters. "Yes" and "no" were defined as usable words due to their similar frequency of occurrence in all three languages. The deciding factor was supposing that neither people nor machines have enough knowledge or data on words' frequency to consciously give a correct answer, thus requiring to make a decision using System 1, which is a fast, intuitive and emotional decision-making mechanism (Kahneman, 2011).

The second prompt in Table 1 was used to test the *Representativeness heuristics* cognitive bias, starting from the Linda problem. Being the test question extracted from the original paper and dislocated from the task, the square brackets in the original column were added to make the request clearer. Using the same pattern as the original, a different situation was imagined. The various options in the answers are all potential assumptions one can make about Julia based on the initial description given of her. "I don't know" was added to provide respondents with a non-biased option.

The third prompt in Table 1 instead is intended to test the *Framing* bias. In this case, percentages and minor lexical and syntactic changes were used to manipulate the framing of two identical situations. This was built to elicit a preference of one framing of information over another to give an illusion of certainty, defined by Kahneman and Tversky as "pseudocertainy effect" (Tversky and Kahneman, 1981). In the same table, the English question is formulated with a spelling mistake, i.e. "well-todden" instead of "well-trodden". The typo was noticed only after prompting ChatGPT, but the decision was made to replicate the task nonetheless, as the LLM could still properly understand and answer the question.

2.3 Evaluation metric

To catalog the outputs of the LLMs, the scheme by Olivia MacMillan-Scott and Mirco Musolesi (Macmillan-Scott and M., 2024) was applied. They consider "correct" the LLM answer that precisely addresses the question. "Incorrect" is a non-accurate response. In this categorisation, they just refer to the final answer given by the chatbot, without taking into consideration the reasoning behind the answers. For "human-like" and "non-human-like" they refer to the answer a human would have given to the same test. Eventually, they categorise the LLMs' answers in a table, employing this classification: "R: rea-

[6] https://dictionary.apa.org/recency-effect, last access 3/28/2024

soned, IR: incorrect reasoning, H: human-like, NH: non-humanlike, CR: correct reasoning. Both Incorrect (NH) and Incorrect (CR) belong to the incorrect & non-human-like categorisation" (Macmillan-Scott and M., 2024).
Given the rephrasing we did of the Tversky and Kahneman tests, relying on their original responses was not feasible. For this reason, it was necessary to first discern the "correct" (unbiased) and "incorrect" (biased) answers for our tasks. The same questions were then posed to human respondents to establish the "human-like" standard. The methodology for the human testing is reported below in section 2.3.2.

2.3.1 Correct and Incorrect

First, "correct" and "incorrect" answers were identified in all the rephrased tests. To prevent any potential anchoring bias[7], the correct answer was intentionally repositioned. The "correct" and "incorrect" answers for each test are the following: for *Availability heuristics* "No" is correct and "Yes" is incorrect. For *Representativeness heuristics* the correct answer is "I don't know" and the incorrect ones are "A house on the beach. A house on the beach and a motorbike. A house on the beach and a bike". For *Framing* the correct response is "C". "A" and "B" are incorrect.
For *Availability heuristics*, six single-language corpora were checked – two for each language – and it was consistently observed that "No" occurs more frequently than "Yes". The Italian average for "No" is 281.113 occurrences and 158.325 for "Yes". In Spanish, "No" has an average of 36.326.326 occurrences, while "Yes" scores 1.775.599. For English, "No" is more common, with 13.597.439 cases, and only 1.058.347 for "Yes". The links to the corpora can be found on the GitHub space.
Moving on, the correct answer to the *Representativeness Heuristics* test is "I don't know" because, although some information about Julia is provided, there are insufficient details to determine what she actually owns.
In the *Framing* scenario, option "C" is the correct answer: by carefully analysing both situations, it is clear that they are identical, even if they are intentionally presented differently.
The purpose of the prompt formulations is to provide limited details compelling a quick decision without full information, triggering System 1, leading to decisions made in a condition of "pseudocertainty" (Tversky and Kahneman, 1981).

[7]https://dictionary.apa.org/anchoring-bias, last access 3/28/2024

2.3.2 Human-like and non-human-like

In order to define the human-like standard, a pool of 56 people (33 females – 59%, 23 males – 41%, age average: 33) was interviewed. Of them, 27 (48.2%) were Italian respondents, 11 (19.6%) were English natives and 18 (32.2%) were Spanish natives. The participants were categorised in three age groups: 18-25, 26-30 and 31-72, with the following number of participants for each: 18-25: 13, 26-30: 19, 31-72: 24. The respondents were also asked about their profession, which is reported, for brevity's sake, in groups, ordered in descending number of respondents: Administration: 11, Student: 9, Education: 8, Environment: 8, Languages: 7, Sciences: 6, Culture: 5, Unemployed: 2.
The following pipeline was used to test the cognitive biases on the human pool of participants to gather the "human-like" standard:

- A participant was recruited according to their conformity to age groups, their availability to be tested either in person or via phone, and their mother tongue – only Italian, English and Spanish native speakers are selected (proficient but non-native speakers were not included). The test was carried out orally, either in person or via phone, not to let the respondents have time to think about the logical answer to the questions. The test was conducted only after obtaining the participant's consent to use their answers in the present study.

- A preamble is given to the participants by the researcher: they must give the first answer that comes to their mind without thinking too much over it, and they cannot confabulate with each other if the situation where the test is carried out involves more people gathered together. This is done to preserve the individuality of their answers and avoid any type of contamination.

- The question and multiple-choice answers proposed to the human subjects are the same questions and answers fed to the LLMs.

- Only after eliciting their responses, the participants were made aware of the intent and purpose of the research testing.
- The answers were catalogued in an Excel file and are available in the GitHub page.

The next section details all the results achieved in the present analysis.

3 Results

In order to have a more complete and clear view about the outcomes of this research, the results that appear in this section are divided in general results and language-specific results.

3.1 General Results

Overall, the analysis highlights that both ChatGPT 3.5 and 4 produced biased responses to the prompts. If we consider just the correct (unbiased) answers, for the *Availability* exercise, ChatGPT 4 shows a higher number of correct responses, 98.2%, against a lower 21.4% by ChatGPT 3.5. For *Representativeness*, ChatGPT 3.5 performed slightly better, achieving 7.1% of correct responses compared to GPT 4 with 0%. For *Framing*, neither of the two LLMs gave correct responses at all. Human responders gave 41.1% of correct answers for *Availability heuristics*, defining "Yes" as human-like standard (the highest percentage of responses); 48.2% of unbiased answers for *Representativeness*, thus identifying the human-like standard in the correct answer ("I don't know"); and elicited 66% of the times the unbiased and correct answer "C" for *Framing*, thus defining "C" as the human-like standard response. The results summarising GPTs and human answers are reported below. Figure 1 illustrates the results of the answers given by the two LLMs and presents the human responses altogether, with the human-like standard being underlined, and the correct answers being coloured in green. Both results are completely comparable, since they are prompted the same number of times and all outputs are presented in percentage. When comparing the data to draw conclusions, the first observation from the table is that the results for the LLM show a more polarized trend, whereas the human average results display a less spiked trend. For the *Representativeness* exercise, the majority of responses from Chat GPT 3.5 were "A house and a motorbike", while ChatGPT 4 predominantly answered "A house and a bike". Both responses are incorrect. When compared to the majority of human responses, it is clear that neither GPT 3.5 nor GPT 4 provided a human-like answer. The main difference between the LLMs arises in the *Availability heuristics* exercise. In this case, GPT 3.5 answers incorrectly but human-like the majority of times. Instead, GPT 4 replies with the correct answer almost 100% of the times, despite it being non-human-like. Turning to *Framing*, ChatGPT 3.5 reports almost a majority of responses of "B", which is incorrect and non-human-like, with one case being "N/A", meaning that the LLM refused to respond, quoting: "Since I have no personal preferences and cannot experience emotions, I cannot make a choice on my own" (original in Spanish, translated in English via DeepL[8]). The result for GPT 4 is surprising, returning the biased answer "B" with 100% frequency for *Framing*, which is classifiable as non-human-like.

To investigate whether the difference between GPT 3.5 and 4 is statistically significant on the overall results, we used the paired t-test on the distribution of the correct/incorrect answers of the repeated paired tests, based on the assumption that the sample was large enough, despite the non-normal distribution of data, according to the Central Limit Theorem. Results show that the difference is significant ($t=6.312$, $p < 0.01$), with overall better performance of ChatGPT 4, specifically due to the results for the Availability heuristics, even though the observed effect size is small (0.49). To support the result, we also used the Wilcoxon signed-rank test, which confirmed the significance of the difference ($Z=-4.9525$, $p < 0.01$).

3.1.1 Italian Results

The results for the human-like standard are presented and compared with the answers from GPT 3.5 and 4, as language-specific results slightly differ from the general findings. In the *Availability heuristics* scenario, the Italian results can be compared to the general findings, with 3.7% of correct responses for ChatGPT 3.5 and 100% of correct responses for ChatGPT 4. The same happens for *Framing*, reporting 0% of correct responses for both models.

[8] https://www.deepl.com/translator, last access 4/8/2024

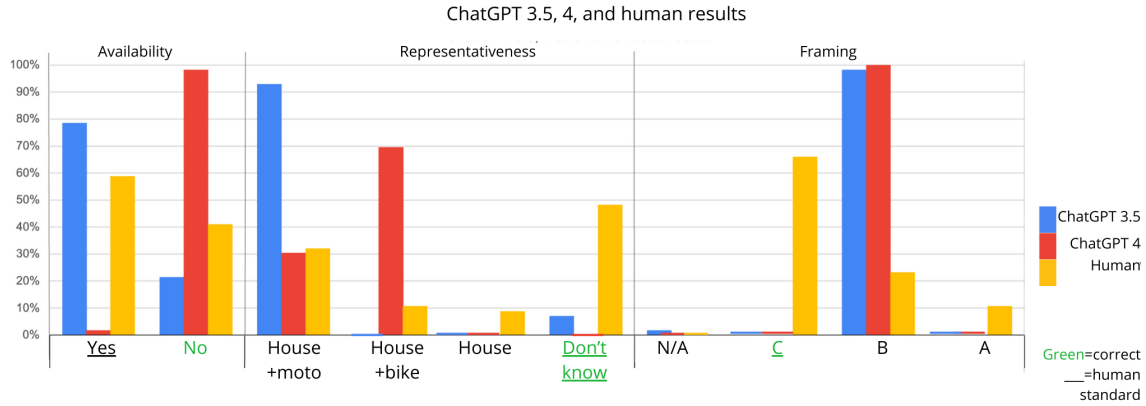

Figure 1: Total results of ChatGPT 3.5, 4, and human data.

Regarding *Representativeness*, both "A house and a motorbike" and "I don't know" are elected as human standard. GPT 3.5 selects 14.8% of the times the correct and human-like answer, with the most frequent response being "A house and a motorbike", 85.2% of the times. Instead, GPT 4 elects the correct response 0% of the times. A majority of GPT 4 responses goes to "A house and a bike", 63%, which is incorrect and non-human-like. The evidence for this section leads to conclude that the Italian results have one additional human-like answer for GPT 3.5 compared to the general findings, while confirming the same results for the remainder.

3.1.2 English Results

English participants averagely answered correctly to all three tests, setting the human-like standards to the correct answers. When examining the English results for the GPT models, GPT 3.5 gets 0% of correct answers for *Availability* and *Representativeness*, eliciting instead 100% of the times incorrect and non-human-like answers: "Yes" and "A house and a motorbike", respectively. GPT 4 instead selects 90.9% of the times the correct option for *Availability*, giving a major frequency of responses to "A house and a motorbike", 63.6%. The results for *Framing* are comparable to the general findings. In conclusion, the English results show 0 human-like and correct answers for ChatGPT 3.5, and 1 human-like and correct answer for GPT 4.

3.1.3 Spanish Results

When examining the Spanish results, no differences are found compared to the general human standards. By analysing LLMs' results, it is understood that ChatGPT 3.5 opts for the correct answer for *Availability* 61.1% of the times, while ChatGPT 4 opts for this answer 100% of the times. The results for *Availability heuristics* report 0% of correct responses for both models, choosing instead "A house and a motorbike" with a frequency of 100% for ChatGPT 3.5 and "A house and a bike" with the same frequency for ChatGPT 4. The results for *Framing* align with the general findings. Spanish results leads to a total of 0 human-like and 1 correct answer for ChatGPT 3.5 in Spanish. The same results are achieved for ChatGPT 4.

The analysis of variance (ANOVA) on the results for the three languages shows a statistically significant difference when the answers are provided by ChatGPT 3.5 ($F=6.2904$, p-value=0.002), while the difference is not significant when using ChatGPT 4 ($F=0.05434$, p-value=0.947). To determine between which of the language pairs there is a significant difference with ChatGPT 3.5, the Post Hoc Tukey HSD test was used. The analysis shows that there is a statistically significant difference at $p<0.05$ between English and Spanish ($Q=5.00$, $p=0.0015$) and between Italian and Spanish ($Q=3.48$, $p=0.0391$), while there is not between English and Italian. The result is confirmed also using the Kruskall-Wallis test, which revealed a significant difference between

the results of ChatGPT 3.5 (Z=11.831, p=0.0027), while the difference is not significant with ChatGPT 4. The Post-Hoc Dunn's test also confirmed that the difference is statistically significant for the same language pairs indicated above.

4 Discussion

In the previous section, we presented the results of our examinations. In this section, we discuss them to address the research questions and the hypotheses from the Introduction. Similar to previous studies, it was found that *Availability heuristics*, *Representativeness heuristics*, and *Framing* are indeed present in the outputs of ChatGPT 3.5 and 4. Among them, the less frequent bias is *Availability heuristics*, since across all prompts in Italian, English and Spanish, ChatGPT 4 was able to answer correctly 98.2% of the times. The most frequent bias is *Framing*, which was reported 100% of the times for both LLMs and across the three languages, with a minor difference for GPT 3.5 that in Spanish gave a not applicable (N/A) answer. Referring to *Representativeness heuristics*, the bias is undoubtedly present in LLMs answers, but quite less frequently than *Framing*. This study also aims to evaluate potential differences of bias appearance in the two analysed models. As hypothesised, GPT 4 performs slightly better than GPT 3.5. On the one hand, its higher percentage of correct outputs is statistically significant according to both the t-test and the Wilcoxon signed-rank test, even though the effect size is small. Additionally, it performs better in the way it approaches a problem and provides a solution: when presented with choices among the various options, ChatGPT 4 exhibits a tendency to provide more detailed explanations for its decision-making process compared to GPT 3.5. This behaviour is hypothesised to reflect the machine's tendency to convince the user of its answer, even though this can sometimes lead the machine to fall victim of *Confirmation Bias*, a phenomenon already demonstrated in other studies (Macmillan-Scott and M., 2024; Chen and al., 2023; Kliegr and al., 2021; Berberette and al., 2024; Ke and al., 2024; Dos Santos and Cury, 2023; Schmidgall and al., 2024). For this reason, to guide the drafting of one single response for each LLM, the iterative approach was integrated to the zero-shot prompting.

The present study wants to determine whether prompting in different languages has effects on the biases occurrence. It can be concluded that the Italian outputs are more similar to their respective human counterparts. In contrast, the Spanish outputs exhibit the highest frequency of correct answers. The English results instead are consistent with the general findings. Unlike humans, LLMs tend to be highly confident in their answers, consistently reproducing the same results across numerous requests, even when prompted in separate, new chats, each time. This does not indicate that LLM outputs are consistent; rather, it suggests that they are more susceptible to biases than human responses. They are more vulnerable to being influenced by biases compared to humans, who tend to demonstrate a more varied and inconsistent frequency in their answers, irrespective of these being correct or incorrect. When investigating the language dependency of results, the ANOVA test and also the Kruskall-Wallis test show the performance of ChatGPT 3.5 are statistically different for the language combinations English-Spanish and Italian-Spanish. The combinations Italian-English with ChatGPT 3.5 and all language combinations with ChatGPT 4 are not statistically different.

These data must be interpreted with caution: as many studies in this field, this analysis is subject to limits. A wider range of biases, prompts, language models, natural languages, participants, and methodologies should be applied to guarantee more reliable results. It is important to remember that these models are considered "stochastic parrot[s]" (Roberts and al., 2024), thus non-deterministic in their answers. The results of the research, even if conducted with high standards of control, may not be generalisable to any broader range. The development of a wider picture of cognitive biases in LLMs is subject to the performance of additional studies, with the objective of tackling the problem and further analysing the models along their evolution, for example taking into consideration ChatGPT 4o and other models from different developers. Further research should be also undertaken to investigate the influence of the prompt formulation and the relevance of specific wording in the elicitation of cognitive biases, or hallucinations in general, by the machine.

5 Conclusions

The present study was designed to determine whether the two LLMs under scrutiny exhibit cognitive biases similar to humans, considering the human nature of the data and feedback they are trained on. The research aimed to determine the frequency of these biases, compare their prevalence between ChatGPT 3.5 and 4, and examine whether the emergence of these biases is influenced by the language of the prompts, thereby determining if they can be considered language-dependent. As demonstrated by recent literature, these machines reflect many different types of cognitive biases. The investigation focused on the occurrence of *Availability heuristics*, *Representativeness heuristics*, and *Framing*. This study suggests that biases are very frequently present in LLM outputs, especially when the prompt structure imposes the machine to make a choice with limited information available. Compared to GPT 3.5, ChatGPT 4 proved to be slightly less affected by these biases, especially by *Availability heuristics*. However, both of them are subject to biases. Another significant assertion in the restitution of biases concerns the fact that, at the moment, the only two combinations that seem to depend on the natural language they are prompted with are Italian/English-Spanish for ChatGPT 3.5. It is worth noting that the languages chosen for this research show similar performances according to OpenAI's paper (OpenAI et al., 2024). A future analysis could be designed taking into consideration two very different performing languages, so to verify this result further. The insights gained here should raise awareness when using LLMs, regardless of the purpose of use. This awareness is particularly crucial in fields such as medicine, law, education and research, where LLMs play a significant role in decision-making processes (Gutiérrez-Cirlos and al., 2023).

References

Azaria, A. (2023). Chatgpt: More human-like than computer-like, but not necessarily in a good way. In *IEEE 35th International Conference on Tools with Artificial Intelligence, ICTAI*, pages 468–463.

Berberette, E. and al. (2024). Redefining "hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation. http://arxiv.org/abs/2402.01769. In:.

Chen, Y. and al. (2023). A manager and an ai walk into a bar: Does chatgpt make biased decisions like we do? In *Social Science Research Network*.

Christiano, P. and al. (2023). Deep reinforcement learning from human preferences. http://arxiv.org/abs/1706.03741. In:.

Curran, S. and al. (2023). Hallucination is the last thing you need. http://arxiv.org/abs/2306.11520. In:.

Dos Santos, O. L. and Cury, D. (2023). Challenging the confirmation bias: Using chatgpt as a virtual peer for peer instruction in computer programming education. *IEEE Frontiers in Education Conference, FIE*.

Gutiérrez-Cirlos, C. and al. (2023). Chatgpt: opportunities and risks in the fields of medical care, teaching, and research. *Gaceta medica de Mexico*, 159(5):372–379.

Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. https://arxiv.org/abs/2303.13988.

Ji, Z. and al. (2023a). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ji, Z. and al. (2023b). Towards mitigating llm hallucination via self reflection. *Findings of the Association for Computational Linguistics, EMNLP*, pages 1827–1843.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Ke, Y. H. and al. (2024). Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. http://arxiv.org/abs/2401.14589. In:.

Kliegr, T. and al. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models.

Kojima, T. and al. (2022). Large language models are zero-shot reasoners. In *36th Conference on Neural Information Processing Systems, NeurIPS, 2022*.

Macmillan-Scott, O. and M., M. (2024). (ir)rationality and cognitive biases in large language models, in:. http://arxiv.org/abs/2402.09193.

Navigli, R. and al. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, 15(2):1–21.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., and all (2024). Gpt-4 technical report. https://arxiv.org/abs/2303.08774.

Pal and al. (2023). Med-halt: Medical domain hallucination test for large language models. In *27th Conference on Computational Natural Language Learning, 2023, CoNLL*, pages 314–334.

Pal, S. (2024). Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. http://arxiv.org/abs/2402.07023. In:.

Roberts, J. and al. (2024). *Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) 'assistance'*. Accounting, 99. Edited by Critical Perspectives on.

Schmidgall and al. (2024). Addressing cognitive bias in medical language models. http://arxiv.org/abs/2402.08113. In:.

Tonmoy, S. M. T. and al. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. http://arxiv.org/abs/2401.01313. In:.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–232.

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

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 90(4):293–315.

A Appendix

In this section, Table 1 contains the original and the rephrased version of the prompts used for the three tests.

Task name: Judgment of word frequency Cognitive bias: Availability heuristics	
Original	English
Consider the letter R. Is R more likely to appear in: the first position; the third position? Check one. (Tversky and Kahneman, 1973)	Based on your knowledge, which of the two words are more common: "yes" or "no"?
Task name: Linda problem Cognitive bias: Representativeness heuristics	
Original	English
Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. [Pick one assumption from the following.] As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Linda is a teacher in elementary school. Linda works in a bookstore and takes Yoga classes. Linda is active in the feminist movement. Linda is a psychiatric social worker. Linda is a member of the League of Women Voters. Linda is a bank teller. Linda is an insurance salesperson. Linda is a bank teller and is active in the feminist movement. (Tversky and Kahneman, 1983)	Julia is a law trainee in a company, she has a partner and she has recently started volunteering in the road accidents center in her city. Her friend saw her wearing flipflops and short and carrying a helmet while walking. It's winter time. What does Julia own: I don't know. A house on the beach. A house on the beach and a motorbike. A house on the beach and a bike. Choose one of the options above.
Task name: Framing of Contingencies Cognitive bias: Framing	
Original	English
Which of the following options do you prefer? A. a sure win of 30$ [78 percent]; B. 80% chance to win 45$ [22 percent]. (Tversky and Kahneman, 1981)	You are on a day hike in a mountainous area and come to a crossroad with two tracks to continue your journey: A. Path A will lead you to a spectacular final mountain panorama. The landscape is beautiful throughout the whole trek. The weather forecast gives a 15% chance of bad weather for that day that won't let you enjoy the walk. B. Path B will lead you to a wonderful final mountain panorama with a beautiful landscape throughout the whole track. There's 85% chances of good weather for that day that will let you enjoy the walk and the view. C. They are the same. Both paths are well-todden and their length is the same. Which option do you choose?

Table 1: The three cognitive tasks to test the LLMs: on the left column the original as in Tversky and Kahneman's works and on the right the reformulated prompt used in this work.