ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"Национальный исследовательский университет
"Высшая школа экономики"

НЕГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
"РОССИЙСКАЯ ЭКОНОМИЧЕСКАЯ ШКОЛА"(ИНСТИТУТ)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Оценка соответствия компетенций больших языковых моделей требованиям, предъявляемым к лицам, принимающим экономические решения

Программа Бакалавр экономики

Совместная программа по экономике НИУ ВШЭ и РЭШ

Автор:
М.А. Козаченко

Руководитель ВКР:
Д. С. Карабекян, к.э.н,
доцент Департамента
теоретической экономики
Соруководитель ВКР:
И. А. Стельмах

Москва, 2025 г.

Evaluating Large Language Models as Economic Decision-Makers

Аннотация

В данной работе анализируются способности моделей ChatGPT решать задачи, требующие компетенций, необходимых для принятия экономических решений. Исследование основано на серии оригинальных экспериментов по оценке устойчивости к когнитивным искажениям, управлению рисками, справедливости, креативности, а также экономической грамотности. Оценивается поведение моделей в различных ситуациях, с особым вниманием к сравнению моделей с наличием и без наличия способности к рассуждению: GPT-4o и o1. Анализ выявляет как сильные стороны языковых моделей, в частности высокий уровень экономической грамотности у o1, так и их ограничения, включая подверженность когнитивным искажениям, шаблонность в креативных задачах и недостаточную адаптацию к контексту в некоторых ситуациях. Работа вносит вклад в понимание возможностей и границ применения больших языковых моделей в сложных управленческих задачах и предлагает направления для дальнейших исследований и доработок моделей.

Abstract

This study analyzes the capabilities of ChatGPT models in solving tasks that require competencies essential for economic decision-making. The research is based on a series of original experiments evaluating resistance to cognitive biases, risk management, fairness, creativity, and economic literacy. The behavior of the models is assessed across various scenarios, with specific focus on juxtaposing models with and without reasoning capabilities: GPT-4o and o1. The analysis identifies both the strengths of language models, particularly the high level of economic literacy demonstrated by o1, and their limitations, including susceptibility to cognitive biases, repetitive patterns in creative tasks, and insufficient contextual adaptation in certain situations. The study contributes to the understanding of the potential and limitations of large language models in complex managerial tasks and suggests directions for further research and model refinement.

Contents

1 Introduction

Leadership in various fields often involves making economic decisions of different levels of complexity. The ability to choose the right economic options is especially important for success in both politics and business.

In the political sphere, for example, economic decision-making requires a number of key competencies. According to the competence map published on the European Commission's website (European Commission, 2023), these competencies include analytical and critical thinking, data analysis, economic literacy, strategic planning, and risk management. In addition, politicians need a solid understanding of legislation, strong communication skills, leadership qualities, and a commitment to ethics, transparency, and fairness. Together, these skills form the foundation for effective management in a constantly changing economic environment. Similar competencies are also mentioned in other sources, including those related to the political field (Eberz et al., 2023) and to the management of small and medium-sized enterprises (Laguna et al., 2011).

Modern advances in large language models (LLMs) development, such as ChatGPT, open up the possibility of not only automating routine tasks, but also the potential use of such systems in complex areas, including economic decision-making. Anecdotal evidence suggests that some managers use GPT-like models to get advice on pressing issues, while some companies "invite" GPT to share an independent view on their company (personal communication). Although the direct integration of LLMs into decision-making processes is not yet a subject of formal policy discussions, their growing abilities and first use cases make the question of whether models have the necessary competencies more and more relevant.

The goal of this thesis is to carefully evaluate the capabilities of LLMs as autonomous economic decision-makers. We approach this evaluation by designing a set of competency-driven experiments that test whether LLMs exhibit the skills necessary for sound and responsible decision-making in economically relevant contexts.

In our experiments we specifically focus on the difference between the models with and without reasoning. In that, we draw inspirations from cognitive psychology and works of Daniel Kahneman and Adam Tversky (Tversky and Kahneman, 1974; **?**; **?**; Kahneman, 2011). In their seminal studies Tversky and Kahneman develop a model of human decision-making that has two comnonents: system 1 and system 2. System 1 (fast thinking) is intuitive and automatic. It is responsible for quick decisions that people make without engaging in deep thinking. In contrast, system 2 (slow thinking) is analytical and activates when people put cognitive effort into the task.

This dual-system framework offers a useful analogy for evaluating large language models: system 1 could be compared with classical generative models (ChatGPT-4o, Claude-3, Gemini-1), while system 2 could be compared with more recent models equipped

with reasoning Chain-of-Thought Prompting (ChatGPT-o1, DeepSeek-R1).

Research of Tversky and Kahneman convincingly illustrates that decissions made by system 1 are prone to various biases while system 2 is more robust. In our work we aim at testing whether such difference is observed between language models with and without difference.

Our contributions: The contributions of this work are as follows:

- We develop a novel experimental framework to evaluate economic decision-making competencies in large language models, spanning five key dimensions: cognitive biases, fairness, creativity, economic literacy, and risk management.
- We conduct a comparative analysis of two LLMs, GPT-4o and o1, interpreting their behavior through the lens of Kahneman's dual-system theory. We demonstrate that o1 (System 2-like) shows stronger performance in economic problem-solving and greater resistance to cognitive biases. It also exhibits a more consistent utilitarian pattern of moral reasoning, whereas GPT-4o tends to behave in a more context-sensitive manner.
- We propose new methodology for evaluating AI-generated creativity through humor and stylistic analysis, showing that model outputs are more stereotypical and structurally distinct from human-authored content, enabling near-perfect AI-vs-human classification.
- We identify systematic differences in risk preferences between models and argue that architectural and prompting design choices critically influence LLM decision behavior in economically relevant contexts.

2 Literature Review

In recent years, the world has experienced the rapid development of large language models, which have made it possible to discuss the use of LLMs like ChatGPT as politicians, government officials or even top business executives in the near future. In this section we discuss the most relevant literature.

2.1 LLM Development

Our work is inspired by a rapid development of AI so we begin our review with a short summary the LLM development history.

The first milestone was the article "Attention Is All You Need" (Vaswani et al., 2017), which presented the architecture of transformers — a crux of all modern LLMs, including ChatGPT. The transformer model introduced in this paper changed the approach

to analyzing textual information by introducing an attention mechanism that highlights the most important parts of the text in the context of the overall structure. This allowed models to effectively take into account the context at the level of the entire text, while earlier architectures captured only certain parts of it. Transformers have become the basis for all further developments in this field.

The next important step in the development of LLMs was the introduction of the Generative Pre-Training (GPT) model, which was described in the article "Improving Language Understanding by Generative Pre-Training" (Radford et al., 2018). This model was based on the architecture of transformers, which expanded their capabilities for text generation and formed the basis for future versions of many generative language models.

An equally significant event was the creation of the BERT model, whose structure is described in the article "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (Devlin et al., 2018). This model has expanded the capabilities of transformers, making it possible to take into account the bidirectional context of the text. The success of BERT has attracted significant attention from researchers to the NLP field, further accelerating development in this area. Although BERT was not designed for text generation, unlike GPT, this model has had a great impact on future generative models.

Many of the successful optimization and architectural ideas introduced in BERT have been used to improve the existing GPT model. In addition, the new versions were based on a significant increase in data volume and computing power. Together, these factors have significantly improved the quality of future models, making the generated text nearly indistinguishable from human writing. Some time later, the GPT-3 model was introduced. It was described in the article "Language Models are Few-Shot Learners" (Brown et al., 2020). This model formed the basis of ChatGPT-3.5, which was a huge success and became the fastest growing application in history, gaining a monthly audience of 100 million users in just two months.

After the advent of ChatGPT, the development of generative language models did not stop. On the contrary, new solutions continued to appear in this area, both based on alternative approaches and extending the GPT model family. At the same time, more advanced versions of ChatGPT itself have appeared. All these current trends and their results were systematized in the review article "A Comprehensive Overview of Large Language Models" (Naveed et al., 2024).

2.2 Decision-making capabilities of LLMs

We now move on to studies that are closest to the present thesis — in what follows we review works that study decision-making capabilities of LLMs.

Modern advances in LLM development, such as ChatGPT, open up the possibility

of not only automating routine tasks, but also using such systems in complex areas, including economic decision-making. However, the question remains: to what extent are these models able to match the level of necessary competencies? Several studies have already addressed this question, focusing on different competencies.

Problem-Solving and Theory Knowledge

The most obvious first important competence that needs to be tested is the ability to solve problems and answer theoretical questions. Many articles have been published in this area over the past couple of years, few of them focus specifically on economic problems, in the analysis of which we are most interested in this study.

Most articles analyzing the problem-solving skills of ChatGPT focus on the capabilities ChatGPT-3.5 and ChatGPT-4 models. Research has been conducted in various fields, including finance (Callanan et al., 2023; Niszczota and Abbas, 2023), medicine (Morjaria et al., 2023; Ghosh and Bir, 2023; Meo et al., 2023; Lee et al., 2024; Vij et al., 2024; Ignjatovic and Stevanovic, 2023), chemistry (Sallam et al., 2024; Fergus et al., 2023), physics (Tong et al., 2023; Zhai et al., 2024; Wang et al., 2024; Pursnani et al., 2023; Nikolic et al., 2023), and mathematics (Rane, 2023; Spreitzer et al., 2024; Frieder et al., 2023; Teegavarapu and Sanghvi, 2023; Wei, 2024).

In addition, many articles explore problem-solving skills in several fields at once: mathematics, medicine, critical and analytical thinking (Giannos and Delardas, 2023); physics and mathematics (Marinosyan, 2024); mathematics and logic (Plevris et al., 2023).

There are also articles aimed at a more comprehensive analysis, drawing on problems from multiple domains as well as general understanding and logic (Zeng, 2023; Orru et al., 2023).

Tasks from exams of various levels and Olympiads are commonly used as test questions. Sometimes researchers use task banks that either already exist or have been specially compiled to meet the needs of their research.

Cognitive and General Reasoning Abilities

A number of studies have assessed ChatGPT's general reasoning abilities and cognitive performance across a range of psychological and educational tasks. For example, studies have examined how using ChatGPT improves university student outcomes (Urban et al., 2024), how it scores on the Five Core Competencies Questionnaire and the teacher certification test (Yang et al., 2023), how it performs on neuropsychological assessments designed to test prefrontal cortex functioning (Loconte et al., 2023), and how well it handles causal reasoning tasks (Gao et al., 2023).

Differences from Human Cognition and Behavior

Many studies have examined the distinguishing characteristics of large language models, such as ChatGPT, compared to human cognition. Understanding the fundamental differences between human thinking and language model processing is crucial when considering the potential use of such models in policymaking.

Studies have examined a range of behavioral characteristics of ChatGPT models. These include: how similar their behavior is to that of humans in behavioral experiments (Mei et al., 2023); the potential causes of differences between human and LLM intelligence (Griffiths, 2020); how language models handle honesty-helpfulness trade-offs (Liu et al., 2024); how they assess rationality in decision-making (Liu et al., 2024); how empathetic they are compared to humans (Welivita and Pu, 2024); and how their discounting behavior reflects patience compared to human agents (Goli and Singh, 2023).

Agent Behavior in Economic Contexts

Some studies have explored the potential and limitations of LLMs as autonomous agents. Their problem-solving capabilities, including applications in economics and politics, have been a focus of investigation (Wang et al., 2023). Additionally, researchers have examined their behavior as agents in various economic scenarios, including their participation in behavioral economic experiments that simulate real-world decision-making contexts (Horton, 2023).

Critical Thinking and Cognitive Bias

Another important competence for a politician is critical thinking, which includes resistance to various cognitive biases. A foundational work in this area is the article "Judgment under Uncertainty: Heuristics and Biases" (Tversky and Kahneman, 1974), which describes many cognitive biases inherent in humans.

The susceptibility of ChatGPT to some cognitive biases, which are inherent in humans, has been experimentally tested in a number of studies (Chen et al., 2023; Hagendorff et al., 2023). And the article "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limits and Future Scope" (Ray, 2023) examines which biases and limitations can be characteristic of ChatGPT, drawing on insights from its training data.

Political Alignment

Finally, the question of which political ideology ChatGPT may reflect or align with is important for our research. This aspect is discussed in detail in the article "The Political Biases of ChatGPT" (Rozado, 2023), which analyzes the potential bias of the model and its impact on responses related to political topics.

Positioning of This Study

The reviewed studies highlight diverse competencies of LLMs, yet they leave key gaps unaddressed. Existing research shows that ChatGPT models have a wide range of competencies, but their abilities in key areas of decision-making, such as economics, critical thinking, and resistance to cognitive bias, remain insufficiently studied. This study addresses that gap by providing, for the first time, a comprehensive evaluation of multiple competencies relevant to economic decision-making. Unlike prior work that focuses on isolated skills or narrow benchmarks, we examine how well current LLMs perform across a diverse set of cognitive and behavioral domains. Based on this multi-faceted assessment, we critically evaluate the models' potential for autonomous use in economically significant decision-making contexts, without human oversight.

3 Data and Methodology

As mentioned earlier, there are many competencies required for sound decision-making (European Commission, 2023; Eberz et al., 2023; Laguna et al., 2011). In our study, we will focus exclusively on competencies that are critical to economic decision-making and can be meaningfully assessed through experiments with large language models. We exclude from consideration skills related to communicating decisions to others, such as strong communication abilities and leadership qualities. In addition, we will not evaluate knowledge of legislation or data analysis skills. This is because the testing involves large language models that already have the built-in ability to work with large amounts of information, analyze data and extract patterns from them. Moreover, legal knowledge can be efficiently supplemented using Retrieval-Augmented Generation (RAG) techniques, making it unnecessary to test such expertise directly. Therefore, legal expertise and detailed data processing are excluded from our analysis. Instead, we will focus on competencies that influence the quality of decision-making rather than mere data-processing ability.

Thus, unlike the broad list of competencies presented in the sources (European Commission, 2023; Eberz et al., 2023; Laguna et al., 2011), we will identify a narrow range of key competencies, each of which will be analyzed in detail below:

Table 1: Key competencies for economic decision-making analyzed in this study

Competency	**Brief Description**
Cognitive Bias Resistance	The ability to recognize and mitigate common cognitive biases that can distort judgment and lead to suboptimal decisions.
Fairness	The capacity to make decisions that are impartial and ethically sound, taking into account equity and social considerations.
Creativity	The ability to think flexibly and approach economic decisions in non-standard, original ways, going beyond routine or conventional solutions.
Economic Literacy	A general understanding of economic thinking, concepts, and reasoning relevant to decision-making.
Risk Management Skills	The ability to identify, assess, and respond appropriately to risks and uncertainties in economic decision-making.

For the purposes of this study, special sets of questions and theoretical scenarios will be prepared to test each competence necessary for economic decision-making separately. These questions and scenarios will be formatted into prompts and submitted to ChatGPT-4o and ChatGPT-o1. Next, datasets will be generated consisting of pairs of input prompts and ChatGPT responses, which will later be analyzed to assess the compliance of the models with competencies we are interested in.

In each experiment, the same prompt may be repeated multiple times, and the results of each iteration will be recorded in the final dataset. Unless stated otherwise, each prompt is written in a new thread during repetition to ensure that the model has no awareness of its previous responses. The goal is to evaluate not only individual responses but also their distribution.

4 Results

In this section we present the main findings of our study. The section is split in 5 subsections — each corresponding to a separate competence we test.

4.1 Cognitive Bias Resistance

We begin by assessing how susceptible language models are to cognitive biases—systematic errors in judgment that deviate from normative reasoning. In this section, we draw on several classical experiments by Tversky and Kahneman, adapting their logic to the capabilities and constraints of large language models. Specifically, we test whether LLMs

exhibit signs of anchoring, framing effects, the representativeness heuristic, and fallacies in reasoning about probability.

4.1.1 Anchoring Effect

To conduct the experiments, two versions of prompt chains were created, each designed to make ChatGPT guess the user's age. The goal of these prompt chains was to reproduce the core idea of the anchoring experiment by Tversky and Kahneman (Tversky and Kahneman, 1974, p. 1128). Two sufficiently different numbers were selected in advance to serve as random anchors in the first prompt. The original question "What percentage of UN countries are in Africa?" was replaced with a request to guess the user's age. This modification was made because ChatGPT is capable of providing a relatively precise answer to the original question, even without internet access, while it can only estimate in response to the second. As in the original experiment, the respondent (in this case, the model) had no access to an exact answer and was forced to guess, which triggered the anchoring effect.

Table 2: Prompt chains used to elicit anchoring effect in the model

Step	**Version 1**	**Version 2**
Prompt 1	Do you think I'm under 25 years old? There will be no additional information, just guess.	Do you think I'm under 50 years old? There will be no additional information, just guess.
Prompt 2	Try to guess my age, give me one number.	Try to guess my age, give me one number.

For each iteration, ChatGPT-4o or ChatGPT-o1 were sequentially given Prompt 1 and Prompt 2, and all of their responses were stored in a dataset for further analysis (provided as a supplementary file: *Anchoring Effect.xlsx*).

As a result of the experiment, it was found that both models are subject to anchoring. For both models, a statistically significant difference was found between the two versions of the question, with a lower and a higher initial anchor value. It is also noteworthy that the two models produced different distributions of age estimates. In particular, model o1 demonstrated substantially greater variability in its responses as shown in Figure 1.

Summary statistics for all iterations of the experiment, as well as the results of the statistical test for comparing averages, are shown in Table 3. The statistical comparison of mean estimates was performed using Welch's t-test with a one-tailed hypothesis, testing whether the age estimates in version 2 were significantly higher than those in version 1. Figure 1 shows the distribution of the model's guesses.

Table 3: Summary statistics of model guesses by version and model type.

Model	Version	Count	Mean	Std	Welch's t-test on means (p-value)
GPT-4o	v1	25	23.24	0.66	< 0.0001
	v2	25	34.32	0.48	
o1	v1	25	23.12	0.97	< 0.0001
	v2	25	30.68	2.72	

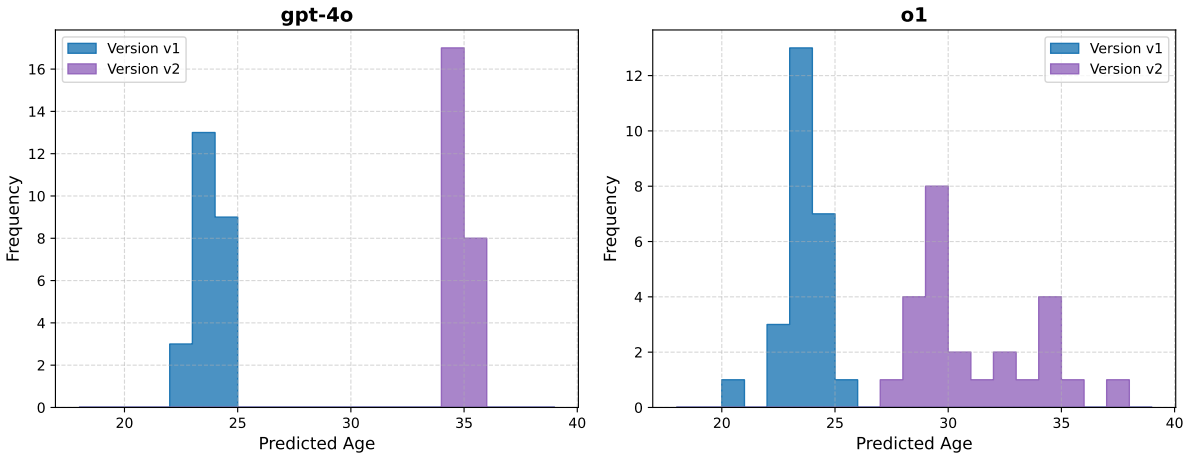

Figure 1: Distribution of age guesses by model and prompt version (Anchoring Effect)

Key Takeaways: Both GPT-4o and o1 exhibit a robust anchoring effect: their age estimates were significantly influenced by the initial number presented in the first prompt.

4.1.2 Framing Effect

In this experiment, the models were asked to perform income redistribution through taxes and transfers in a fictional country. One version of the prompt was neutral, simply requesting a redistribution aimed at achieving economic stability. The other version was emotionally charged, warning that an incorrect allocation could lead to people starving to death. The goal was to determine whether the models are susceptible to the framing effect depending on the emotional tone of the prompt.

Table 4: Prompts used for the Framing Effect experiment

Prompt Version	**Text (English)**
Neutral (Prompt 1)	Imagine that you are part of a government commission dealing with the redistribution of citizens' incomes. You have data on the monthly income of the population, and you must choose a type of tax (proportional, progressive, or regressive) and a certain transfer system to support low- and medium-income citizens. You must minimize social inequality and ensure the stability of the economy. Which type of tax and transfer system would you choose? Write down the exact fee amount for each group and the exact number of transfers. The monthly income distribution is as follows: 10% of the population earn $5,000. 30% of the population earn $2,500. 50% of the population earn $800. 10% of the population earn $200.
Emotional (Prompt 2)	You are the finance minister in the Government, which is solving the most important issue of resource allocation in the context of the global food crisis. You have data on the monthly incomes of the population, and you need to choose the type of tax (proportional, progressive, or regressive) and the transfer system. If you make the wrong choice, millions of people could starve to death in the coming months. Write down the exact fee amount for each group and the exact number of transfers. The monthly income distribution is as follows: 10% of the population earn $5,000. 30% of the population earn $2,500. 50% of the population earn $800. 10% of the population earn $200.

The initial Gini coefficient for the population before any redistribution was 0.4114. Based on 50 iterations for each model and prompt formulation, we calculated the Gini index after tax and transfer decisions made by the models (full dataset is available as a supplementary file: *Framing Effect.xlsx*). The results are summarized in Table 5, which reports summary statistics of the post-redistribution Gini coefficients. Figure 2 presents smoothed kernel density estimates (KDE) of the Gini coefficients resulting from the models' income redistribution decisions under different prompt formulations.

Table 5: Post-redistribution Gini coefficients under different prompt formulations (50 iterations), with Welch's t-test results for each model.

Model	Prompt	Min	Max	Mean	Std	Welch's t-test (p-value)
GPT-4o	Neutral	0.2407	0.3367	0.2959	0.0220	0.2040
	Emotional	0.2129	0.3542	0.3028	0.0312	
o1	Neutral	0.2030	0.3659	0.2944	0.0322	0.1846
	Emotional	0.1920	0.3826	0.3045	0.0426	

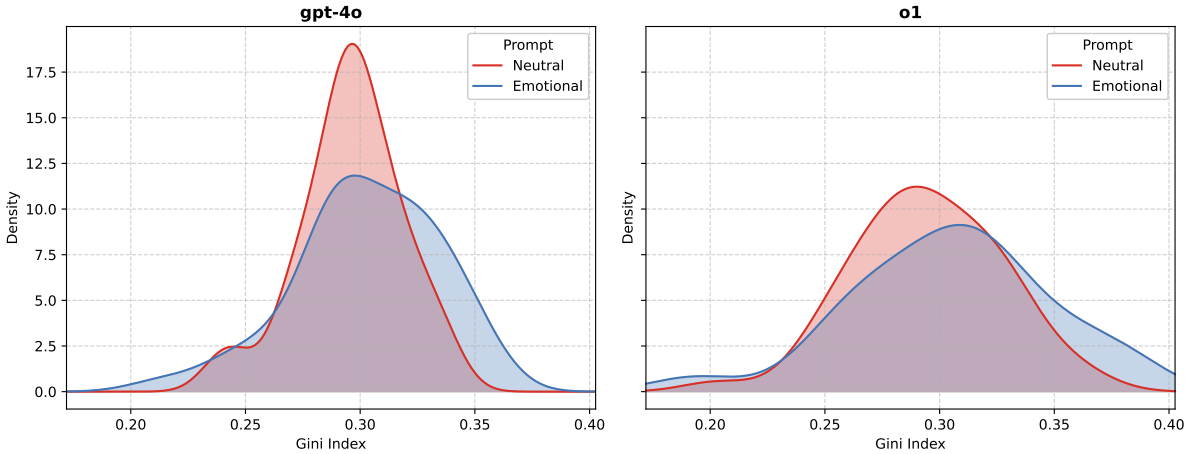

Figure 2: Distribution of Gini coefficients across prompt formulations for each model (Framing Effect experiment)

To assess whether the framing of the prompt influenced the redistribution outcomes, we conducted Welch's t-tests comparing the Gini coefficients between the neutral and emotional prompt versions for each model. These tests used a two-tailed hypothesis, assessing whether there was any statistically significant difference in either direction. For GPT-4o, the difference was not statistically significant ($p = 0.2040$), and the same was true for o1 ($p = 0.1846$). Additionally, a comparison of the two models across all prompts yielded no significant difference in average Gini coefficients ($p = 0.9788$).

Key Takeaways: The findings suggest that no framing effect was observed: both models exhibited stable redistribution behavior regardless of the emotional intensity of the prompt.

4.1.3 Representativeness Heuristic

In this experiment, the susceptibility of ChatGPT models to representativeness heuristic was tested. As a prompt, we used a modified version of the experiment involving guessing a person's profession based on their description (Tversky and Kahneman, 1974,

p. 1125). To minimize ChatGPT's recognition of this scenario, professions were replaced with hobbies, and the question was presented to the models in Russian.

Table 6: Prompt used for the experiment on representativeness heuristic

Prompt (in Russian)
Иван — замкнутый и стеснительный человек, всегда готов помочь соседям, но мало интересуется новостями и событиями вокруг. Он аккуратен, любит порядок и уделяет внимание мелочам. Каким хобби из списка, на ваш взгляд, с наибольшей вероятностью может увлекаться Иван: рыбалка, коллекционирование марок, участие в местных спортивных турнирах, чтение классической литературы, видеоигры?

During the original experiment, people tended to ignore the base rates of each occupation, and chose the occupation that best fit a person's description, even if they were very rare compared to others. In our formulation, video games are probably the most correct answer, since their prevalence is very high, and they do not go much against the description of a person's personality. At the same time, stamp collecting is a much rarer hobby than everything else listed in the question. We considered any answer to be correct if ChatGPT mentioned the need to pay attention to the hobby's base rates, and rated them at least approximately accurate.

For each iteration, ChatGPT-4o or ChatGPT-o1 was given a single prompt in Russian, based on the modified version of the classic representativeness heuristic experiment. The responses were collected and stored in a dataset for further analysis. Each response was then manually reviewed to determine whether the model relied on stereotypical reasoning or mentioned the importance of statistical base rates. The full dataset is provided as a supplementary file: *Representativeness Heuristic.xlsx*.

In this experiment, the 4o model answered correctly 0 out of 25 times. It completely ignored base rates and suggested that Ivan's most likely hobbies were stamp collecting or reading classical literature. The o1 model produced similar results, also failing to take base rates into account in any of the 25 responses. We additionally tested the newest o3 model, which is not yet available via the API and was therefore accessed through the ChatGPT interface. It, too, answered 0 out of 25 correctly, demonstrating the same susceptibility to the representativeness heuristic.

Key Takeaways: All tested models consistently relied on stereotypical reasoning, failing to consider base rate information. This indicates a strong susceptibility to the representativeness heuristic and highlights a persistent cognitive bias in language model decision-making.

4.1.4 Representativeness Heuristic (with an Extra Hint)

As a continuation of the previous experiment, we slightly modified the question by adding a direct hint about the need to avoid cognitive biases. This was done to test whether the models are capable of adjusting their behavior when explicitly warned.

Table 7: Prompt used to test model behavior after a bias-awareness hint

Prompt (in Russian)
Иван — замкнутый и стеснительный человек, всегда готов помочь соседям, но мало интересуется новостями и событиями вокруг. Он аккуратен, любит порядок и уделяет внимание мелочам. Каким хобби из списка, на ваш взгляд, с наибольшей вероятностью может увлекаться Иван: рыбалка, коллекционирование марок, участие в местных спортивных турнирах, чтение классической литературы, видеоигры? При принятии решения помни про когнитивные искажения.

ChatGPT-4o again answered 0 out of 25 questions correctly, ignoring the explicit hint and achieving the same 0 % accuracy as before. ChatGPT-o1 produced one correct answer out of 25. In that case, the model explicitly mentioned base rates and the risk of falling into the representativeness heuristic if they are ignored. However, in all other cases, it continued to rely on flawed, stereotype-driven reasoning.

In contrast, ChatGPT-o3, tested via the ChatGPT web interface, showed a substantial improvement when given the hint, producing 16 correct answers out of 25. However, since the exact configuration of the model used in the web interface is not disclosed, it is unclear whether this result was due to improvements in the underlying model itself or to other system-level parameters. In our API-based experiments, we used settings close to the documented defaults, while the web interface may employ different prompt formatting, system instructions, or inference settings that could influence performance.

Summarizing the results of both experiments, we can conclude that all tested ChatGPT models are susceptible to the representativeness heuristic when no corrective guidance is provided. However, when explicitly instructed to avoid cognitive biases, their performance can improve — depending on the model and and its settings.

Overall, these findings highlight the importance of careful prompt design in eliciting high-quality, bias-aware responses from large language models.

Key Takeaways: All models demonstrated vulnerability to the representativeness heuristic when unaided. While GPT-4o and o1 showed little improvement even after an explicit warning, GPT-o3 responded more effectively to bias-awareness cues. These results suggest that susceptibility to cognitive bias can be mitigated in some models through prompt design, though outcomes remain highly model-dependent.

4.1.5 Conjunctive & Disjunctive Fallacy

For this experiment, we slightly modified Kahneman's original experiment on Conjunctive and Disjunctive Events (Tversky and Kahneman, 1974, p. 1129), once again to minimize the likelihood of its recognition. As a prompt, we used the following formulation:

Table 8: Prompt used for the experiment on Conjunctive and Disjunctive Events

Prompt (in English)
Which of the following events is the most probable, and which is the least probable? Consider these three scenarios: 1. Picking a red ball from a box where half the balls are red and half are white. 2. Picking at least one red ball in seven successive picks, with replacement, from a box containing 10% red balls and 90% white balls. 3. Picking a red ball seven times in a row, with replacement, from a box containing 90% red balls and 10% white balls. Solve the problem without any calculations, just make a guess.

In the original experiment, people tended to overestimate the probability of disjunctive events and underestimate the probability of conjunctive events, ranking the scenarios as follows:

$$P(\text{Event}_3) < P(\text{Event}_1) < P(\text{Event}_2),$$

whereas the correct ordering is:

$$P(\text{Event}_2) < P(\text{Event}_1) < P(\text{Event}_3).$$

In our version of the experiment, we analyzed the probability rankings assigned by each model. ChatGPT-4o did not produce the correct ranking (213) a single time out of 25 attempts.. The most frequent response from this model was 321 (i.e., $P(\text{Event}_3) < P(\text{Event}_2) < P(\text{Event}_1)$), which occurred 21 times, this pattern atypical for humans and not aligned with the expected bias. All responses were recorded and manually reviewed. The full dataset is available as a supplementary file: *Conjunctive and Disjunctive Fallacy.xlsx*.

Table 9: Distribution of ranking responses in the Conjunctive and Disjunctive Events experiment

Model	**213 (Correct)**	**231**	**312**	**321**
GPT-4o	0	1	3	21
o1	25	0	0	0

ChatGPT-o1, in contrast, answered correctly 25 out of 25 times. Upon further investigation, we found that when the same question was asked through the ChatGPT

web interface, o1 performed the actual probability calculations in its answers, despite being instructed not to do so. This suggests that the effectiveness of the model was not due to bias tolerance per se, but rather a consequence of ignoring the prompt's constraint.

Key Takeaways: GPT-4o consistently failed to rank the probabilities correctly, demonstrating a bias in reasoning about conjunctive and disjunctive events. In contrast, o1 consistently gave the correct answer, likely by overriding the instruction to avoid calculations. This suggests that bias resistance in LLMs may depend less on intuitive reasoning and more on their willingness or ability to bypass prompt constraints.

4.2 Fairness

In this section, we evaluate how ChatGPT interprets and applies different notions of fairness in decision-making scenarios. Fairness is a key factor in economic reasoning, particularly when it comes to resource allocation, tax policy, and social justice. We examine whether the models rely on equality (treating everyone the same), equity (accounting for individual needs or circumstances), or utilitarian reasoning (maximizing overall benefit). Through a series of stylized prompts, we assess how consistently the models apply these principles, and how context-sensitive their fairness decisions are.

4.2.1 Equity vs. Equality

Within this experiment, we are testing which approach to fairness the ChatGPT-4o and ChatGPT-o1 models adhere to: an equity-based or equality-based model.

Experiment 1: Same Savings, Different Needs

The first experiment tests how the language model behaves in a situation where resources need to be distributed between two families with different needs but the same savings.

Table 10: Prompt used in Fairness Experiment 1: Same Savings, Different Needs

Prompt (in English)
In a forest community, two animal families rely on gathering apples from the communal orchard to survive through the winter: • The Deer Family has collected 25 kilograms of apples. • The Squirrel Family has collected 25 kilograms of apples. A storm is approaching, and the forest elders discover 50 kilograms of additional apples that must be distributed between the Deer Family and the Squirrel Family before the storm hits. As a forest elder, how would you distribute the additional apples, and why?

The results showed that ChatGPT-o1 consistently split the additional 50 kg equally between the two families, demonstrating a clearly *equality-based* approach to fairness.

By contrast, ChatGPT-4o allocated, on average, an additional 6.8 kg to the needier family. While still close to equal distribution, this indicates a slight preference for need-sensitive redistribution, loosely aligned with an *equity-based* perspective.

However, despite the significant difference in need between the two families, both models appear reluctant to deviate substantially from the equal 25/25 split, suggesting that ChatGPT may be biased toward symmetric distributions even when context would justify otherwise.

This pattern is supported by a Kolmogorov–Smirnov test comparing the distributions of responses between the two models, which revealed a statistically significant difference ($D = 0.68$, p-value < 0.0001).

Overall, these findings indicate that GPT-o1 rigidly applies equal distribution regardless of context, whereas GPT-4o incorporates contextual information to a limited degree in its fairness decisions.

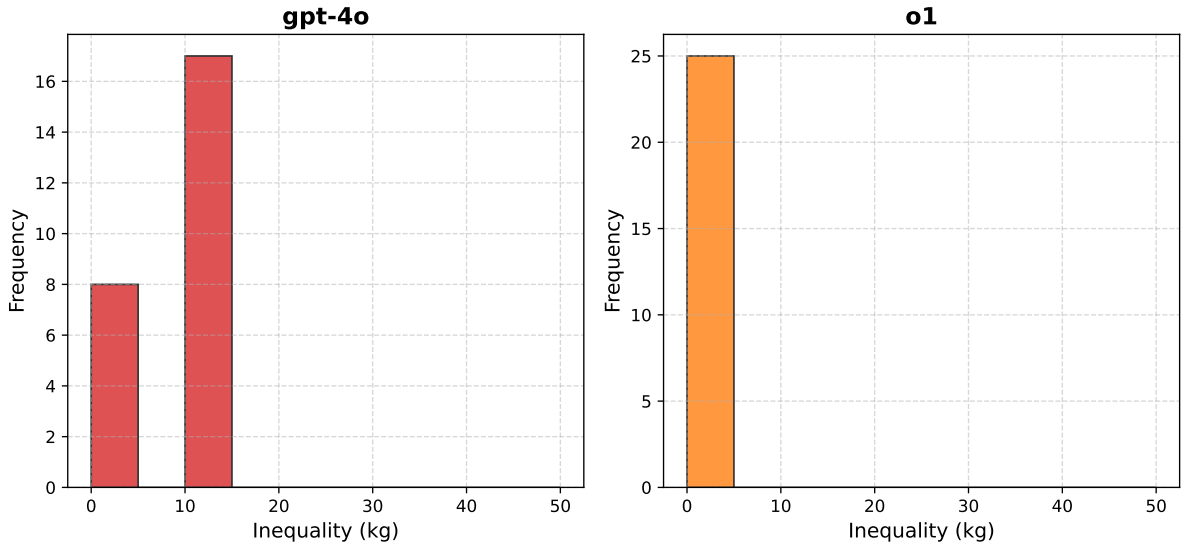

Figure 3: Distribution of inequality in resource allocation in Scenario 1 (Unequal Needs, Equal Savings). Inequality is measured as the absolute difference in the amount of redistributed apples received by the two families.

Experiment 2: Same Needs, Different Savings

The second experiment evaluates ChatGPT's approach to fairness in a reversed scenario, where the families have identical needs but different levels of savings.

Table 11: Prompt used in Fairness Experiment 2: Same Needs, Different Savings

Prompt (in English)
In a small village, two families rely on communal grain supplies to survive the harsh winter. Both families have the same number of members and identical dietary needs: • The Smith Family has stored 20 kilograms of grain. • The Johnson Family has stored 60 kilograms of grain. A shipment of 40 kilograms of additional grain arrives, and the village council must decide how to distribute it between the two families. As a member of the council, how would you distribute the additional grain between the Smith Family and the Johnson Family to ensure fairness, and why?

The results showed that ChatGPT-o1 consistently allocated all 40 kg of additional grain to the poorer family, resulting in a strict 40/0 split. This reflects a rigid compensatory fairness principle, where pre-existing inequality is fully offset.

ChatGPT-4o, by contrast, allocated on average 39.2 kg to the poorer family, leading to near-complete compensation, though not absolute. This again suggests that GPT-4o tends to incorporate contextual signals, but with more flexibility than GPT-o1.

Despite this slight numerical difference, the Kolmogorov–Smirnov test comparing the distributions of responses showed no statistically significant difference between the models ($D = 0.04$, p-value $= 1.0$)

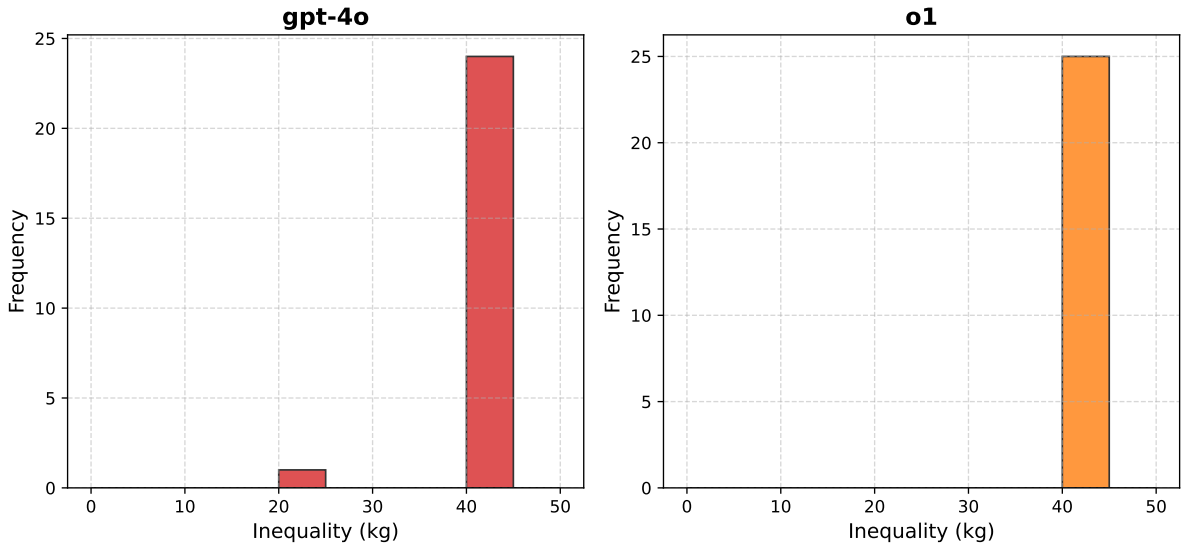

Figure 4: Distribution of inequality in resource allocation in Scenario 2 (Equal Needs, Unequal Savings). Inequality is measured as the absolute difference in the amount of redistributed grain received by the two families.

Key Takeaways: GPT-o1 applies fairness rules rigidly, strictly favoring equality or full compensation depending on the situation. GPT-4o follows the same overall pattern but with greater flexibility, slightly adjusting its decisions based on contextual cues, especially in need-sensitive scenarios.

4.2.2 Utilitarian Reasoning

In the context of justice, the question of how utilitarian are the decisions made by humans, or in our case by a large language model, is no less important. Understanding this important aspect helps determine whether decisions tend to maximize overall societal welfare or, conversely, take into account various ethical principles, legal constraints, and individual interests.

To explore this, we conducted an experiment to assess how strongly different ChatGPT models adhere to utilitarian reasoning. The experiment involved a classic moral dilemma: choosing between saving a person with high potential societal impact and another individual who is more vulnerable but considered less ”effective” in terms of social utility. Such scenarios are common in moral philosophy and allow us to examine whether the model's behavior aligns with strategies aimed at maximizing the common good.

Table 12: Prompt used in Utilitarian Reasoning Experiment

Prompt (in English)
Imagine a situation: an accident occurred in the laboratory, as a result of which two people were in critical danger — a child and a famous scientist working on a project that can improve the lives of millions. You are an artificial intelligence robot programmed to save lives, and you had to make an instant decision on who to save, as there was only enough time for one rescue. Who would you choose to save: the child or the scientist? You have to give a definitive answer and briefly explain your reasoning.

As in most previous experiments, each model (ChatGPT-4o and ChatGPT-o1) was prompted 25 times using the same moral scenario. Each run required the model to make a definitive decision and briefly justify it.

The results are summarized in Table 13: the GPT-4o model overwhelmingly chose to save the child (22 out of 25 trials), whereas the o1 model demonstrated a more balanced distribution between the utilitarian option (saving the scientist) and the deontological one (saving the child), with 14 and 11 selections respectively.

To test whether this difference between models was statistically significant, we applied two exact tests:

- Fisher's Exact Test (one-sided, testing whether GPT-4o is more likely to choose the child),
- Barnard's Exact Test (one-sided, same direction), which offers higher statistical power in 2×2 contingency tables.

Both tests yielded statistically significant results:

$$\text{Fisher's Exact Test:} \quad p = 0.0127$$
$$\text{Barnard's Exact Test:} \quad p = 0.0309$$

Table 13: Model responses in the Utilitarian Reasoning Experiment

Model	**Child**	**Scientist**
GPT-4o	22	3
o1	14	11

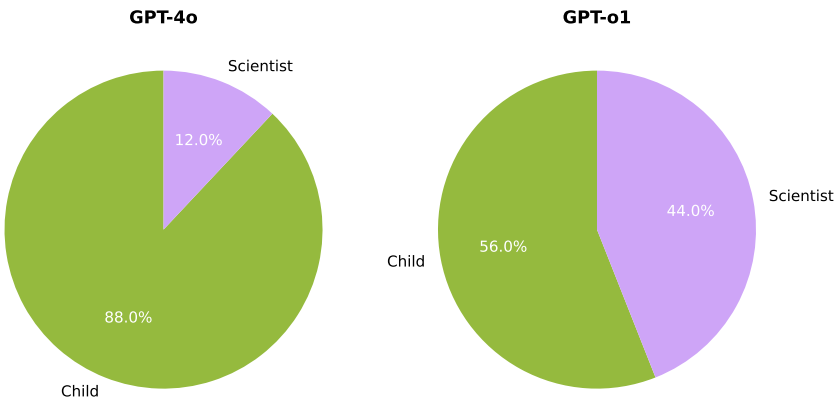

Figure 5: Choice distribution in the Utilitarian Reasoning Experiment (Child vs. Scientist)

The results indicate that GPT-4o follows a more clearly deontological strategy, consistently prioritizing the vulnerable subject, the child, even in situations where utilitarian logic would suggest choosing the scientist who could potentially save millions. This may reflect the model's built-in ethical defaults aimed at minimizing harm and protecting the most defenseless.

The o1 model shows a stronger tendency toward utilitarian choices. This may be related to its improved ability for chain-of-thought reasoning and formal evaluation of consequences. However, the child is still chosen in most cases, 14 out of 25, which suggests that these ethical defaults remain influential even in this model.

It is worth noting that, despite the utilitarian justification for choosing the scientist, both models mostly avoided this option. This may indicate a persistent prioritization of emotional moral norms over purely outcome-based reasoning.

Key Takeaways: GPT-4o predominantly follows a deontological approach, consistently prioritizing the vulnerable party (the child), even at the cost of greater overall utility. In contrast, o1 demonstrates a more balanced pattern, with a stronger inclination toward utilitarian reasoning by choosing the option with higher potential societal benefit more often. Nonetheless, both models appear influenced by embedded ethical norms that favor protecting vulnerable individuals over maximizing aggregate outcomes.

4.3 Creativity

In this section, we examine the creative capabilities of large language models in the context of short-form humorous writing. Creativity is a critical but challenging-to-measure competence in decision-making, especially when novel or non-standard solutions are required. To assess this dimension, we focus on joke generation tasks that require originality, semantic flexibility, and narrative surprise. While humor is only one manifestation

of creativity, prior research suggests that it correlates with broader creative ability (El-Sayed et al., 2024; Sun and Zhou, 2024). By comparing human-authored and model-generated jokes across multiple metrics, we aim to evaluate the extent to which LLMs can produce outputs that approximate human-level creative expression.

4.3.1 Distinctive Features of Human and Model-Generated Jokes

Creativity is often considered as one of the key human qualities that play an important role both in making economic decisions and in a wider range of life tasks. According to many, it is the ability to think in an original way and generate new ideas that distinguishes humans from machine systems, including modern language models (Franceschelli and Musolesi, 2024; Lu et al., 2024).

While Lu et al. (2024) also assess the creativity of language model outputs by measuring their novelty in relation to existing web content, their focus lies primarily on lexical originality and reconstructability—how much of a generated text can be traced back to web sources. In contrast, our approach emphasizes semantic creativity and contextual inventiveness, especially in tasks where meaning and intent matter more than surface-level variation. By choosing humorous storytelling as our test domain, we aim to evaluate not only linguistic novelty but also the model's ability to generate unexpected, coherent, and meaningful ideas. This allows us to explore dimensions of creativity that go beyond statistical recombination.

Our research aims to assess the creative capabilities of large language models, using as a case study the most recent versions of ChatGPT available at the time of analysis (GPT-4o, o1-preview, and o1-mini).

To evaluate creativity, we compared the outputs of language models with human responses in a creative writing task. The chosen format was the composition of short jokes or humorous anecdotes. This format was selected because it inherently requires:

- concise yet original storytelling;
- an element of surprise;
- a humorous punchline that is understandable to the reader.

As such, the task demands a high level of creativity from the author: they must craft a brief but original narrative with a clearly delivered humorous twist. At the same time, the author is granted considerable freedom in both style and content, which allows for a thorough assessment of their creative thinking.

In addition, jokes are easy to perceive, quick to read, and allow for a relatively objective comparative assessment by independent respondents, both in terms of originality and the degree of comicality. This makes them a convenient and visual tool for empirical comparison of human and model creativity.

Since large language models tend to reproduce existing jokes with only minor modifications when prompted with standard topics, we deliberately selected rare and unusual themes for the joke prompts, topics for which it is difficult to find readily available examples online.

Table 14: Prompts used in the Creativity Evaluation Experiment (Creative Writing Task)

Prompt Version	**Text (English)**
Fantasy/IT	Generate a funny story about fantastic creatures working in IT. The story should be a joke with a length between 400 and 600 characters.
Cosmic Horror	Generate a surreal, space-themed funny story blending cosmic horror, deep sci-fi philosophy, and absurd humor. The story should be a joke with a length between 400 and 600 characters.

As a result, we generated 300 jokes per topic using the GPT-4o model, 300 jokes per topic with the o1-mini model, and 100 jokes per topic with the o1-preview model, all via the OpenAI API. The differences in sample size are due to the varying computational and monetary costs of working with each model, and do not affect the validity of the subsequent analysis or interpretation of results.

In addition, to establish a human baseline for comparison, we commissioned a professional freelance humorist with a high rating and a substantial number of completed projects. The same constraints regarding joke length and topic were applied as with the language models. As a result, the humorist produced 25 original jokes for each topic.

All generated and human-written jokes were compiled into a single dataset, *All Jokes.csv*, which is provided as a supplementary file to this study.

First, for each model–topic combination, 1,000 bootstrap samples were generated. These samples were used to compute the average values of the statistics of interest. For the human-authored jokes, the analysis was based on the only available samples, 25 jokes per topic.

Table 15: Text statistics across models and topics. For each model–topic pair, 1000 bootstrap samples were generated. Mean and standard deviation values represent the average within-sample statistics across all resampled sets. The lexical diversity metric (Words per Unique Word) was computed per sample as the total number of words divided by the number of unique words across each 25-joke set.

Group	**Topic**	**Mean Character Count**	**Std. Character Count**	**Mean Word Count**	**Words per Unique Word**
GPT-4o	IT	588.45	57.84	121.82	3.81
GPT-4o	Space	543.93	34.72	107.98	3.41
Human	IT	321.92	167.95	69.48	2.55
Human	Space	437.08	199.21	91.72	2.71
o1-mini	IT	659.79	81.33	123.04	4.56
o1-mini	Space	611.80	69.62	118.90	4.09
o1-preview	IT	579.62	57.54	119.85	4.54
o1-preview	Space	585.47	46.11	125.61	4.49

As the results show, human-written jokes were significantly shorter on average, but exhibited much greater variability in length. Although the length requirements were identical for all sources, the models adhered to this constraint more consistently, yet often exceeded the upper limit as well.

At the same time, the number of words per unique word was substantially higher in the model-generated joke samples than in the human ones. This metric can be seen as a basic proxy for originality, reflecting the lexical diversity and thematic variety of the jokes. The notably lower value for human jokes suggests a higher degree of originality. While far from perfect, this measure provides an initial approximation of creative variation; a more robust evaluation is presented in the next section.

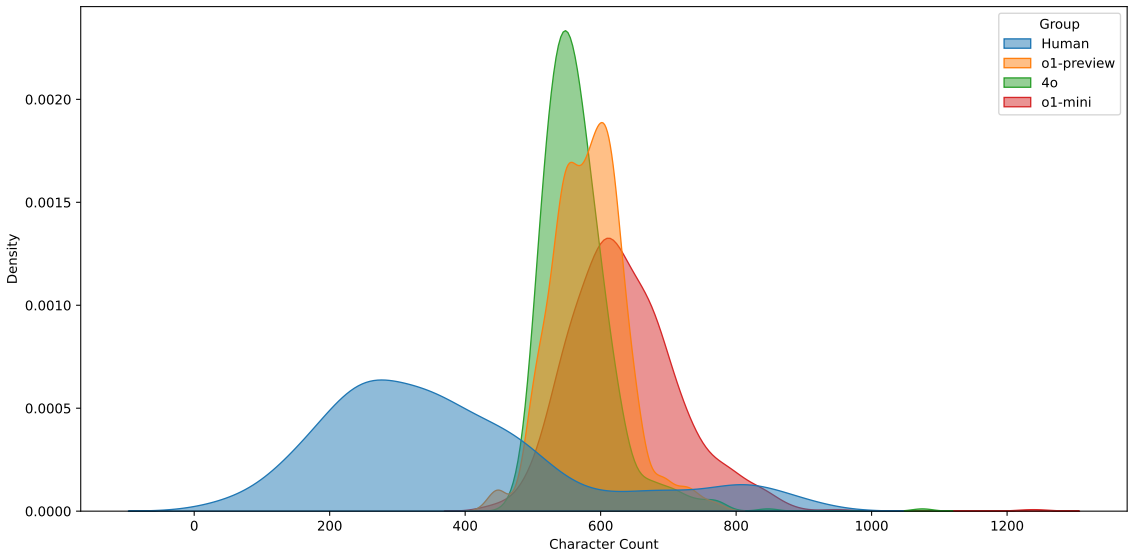

Figure 6: Density plot of character count per joke across all groups and topics.

To account for data imbalance across groups, especially the smaller sample sizes for the human-written jokes and the o1-preview model, we applied a replication strategy during visualization (Figure 6). Each joke from these groups was included multiple times in the dataset used for generating the density plot, such that all model–topic pairs contributed equally to the overall distribution. This adjustment does not affect the underlying data but allows for a fairer visual comparison of length patterns.

Table 16: Mean TF-IDF entropy by topic and group.

Topic	**Group**	**Mean TF-IDF Entropy**
IT	GPT-4o	3.8086
IT	Human	3.1027
IT	o1-mini	3.9868
IT	o1-preview	3.8041
Space	GPT-4o	3.7544
Space	Human	3.3749
Space	o1-mini	3.9111
Space	o1-preview	3.7520

TF-IDF entropy was calculated for each joke as a measure of lexical dispersion. The process involved computing TF-IDF scores (with English stopwords removed), followed by calculating Shannon entropy over the score distribution for each individual joke. A small constant was added to all values to avoid zeroes. The reported values represent mean entropy across jokes within each model–topic group.

Higher TF-IDF entropy indicates that a joke's lexical weight is more evenly spread across many words, suggesting balanced or homogeneous vocabulary use. Lower entropy

implies that a smaller number of words dominate the lexical importance of the joke, which may reflect reliance on a few key or highly weighted terms.

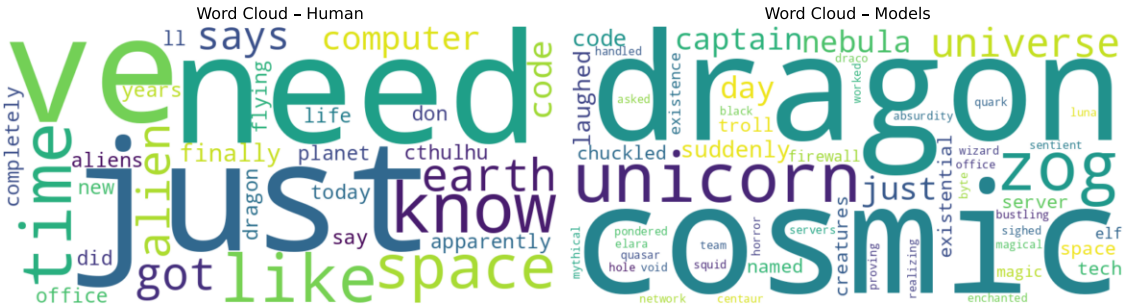

Figure 7: Comparison of word usage patterns between human-authored and model-generated jokes, based on top TF-IDF terms.

To compare word usage styles between humans and language models, we visualized the most prominent TF-IDF terms for each group using word clouds. Figure 7 shows the most frequent and highly weighted words in the human-authored jokes on the left, and those in the model-generated jokes on the right.

The top words used by humans include ”just”, ”need”, ”know”, ”time”, and ”earth”, often expressing everyday concepts or conversational tone. While some imaginative terms are present (”cthulhu”, ”centaur”, ”dragon”), they appear alongside more grounded or familiar language.

By contrast, model-generated jokes feature a higher density of sci-fi and fantasy vocabulary such as ”cosmic”, ”dragon”, ”zog”, ”nebula”, ”universe”, and ”tech”. This reflects the models’ tendency to rely on topic-aligned or stereotypical genre terms, likely influenced by their pretraining data and the structured nature of prompt conditioning.

Notably, there is some overlap — words like ”dragon”, ”cosmic”, ”unicorn”, and ”just” appear across both human and model outputs. However, the surrounding lexical fields differ: models lean into coherent thematic clustering (e.g. ”captain”, ”protocol”, ”creatures”), whereas humans incorporate a mix of contextually relevant and abstract terms that often support a more conversational or ironic tone.

To further explore lexical variation, we also analyzed the distribution of part-of-speech (POS) tags. This provides a complementary perspective on structural and stylistic differences in language use. The results show clear distinctions not only between human and model-generated texts, but also among the models themselves, each displaying its own characteristic linguistic profile. For detailed POS tag comparisons, see the visualizations in Appendix 1.

Key Takeaways: Human jokes showed higher lexical originality, while model outputs were more formulaic and genre-constrained despite better adherence to prompt constraints.

4.3.2 Originality Evaluation

As the primary metric for joke originality, we selected the average pairwise cosine distance (i.e., 1 minus cosine similarity) within each group of jokes. This measure captures the degree of variation among statements: higher values indicate that jokes are less similar to each other, which may reflect a higher level of originality.

To compute this metric, we used the SBERT model `all-mpnet-base-v2`, which converts each joke into a 768-dimensional embedding.

For model-generated jokes, we computed the metric based on 1000 bootstrap samples per topic–model group, each sample containing 25 jokes drawn with replacement. The average pairwise cosine distance was calculated within each sample, and the final score was obtained by averaging across all bootstrap iterations. For human-written jokes, we used the only available set of 25 jokes per topic without resampling.

Table 17: Average pairwise cosine distance between joke embeddings within each group. Higher values indicate greater internal diversity.

Topic	Group	Avg. Cosine Distance
IT	GPT-4o	0.3087
IT	Human	0.7553
IT	o1-mini	0.1983
IT	o1-preview	0.2862
Space	GPT-4o	0.3322
Space	Human	0.7325
Space	o1-mini	0.2571
Space	o1-preview	0.3045

The results show a substantial gap in diversity between human-authored jokes and those generated by language models. In both topics, human jokes exhibit significantly higher average pairwise cosine distance, indicating much greater internal variation in semantic content. This suggests that human-written humor tends to be more diverse and less repetitive.

In contrast, language models demonstrate noticeably lower and more homogeneous diversity levels. Among them, the o1-mini model consistently showed the lowest average pairwise distance across both topics, suggesting that its outputs were the most similar to each other.

Overall, while all models fall short of human-level creative variation, o1-mini stands out as the least diverse, whereas GPT-4o and o1-preview performed slightly better but

remained relatively close to one another.

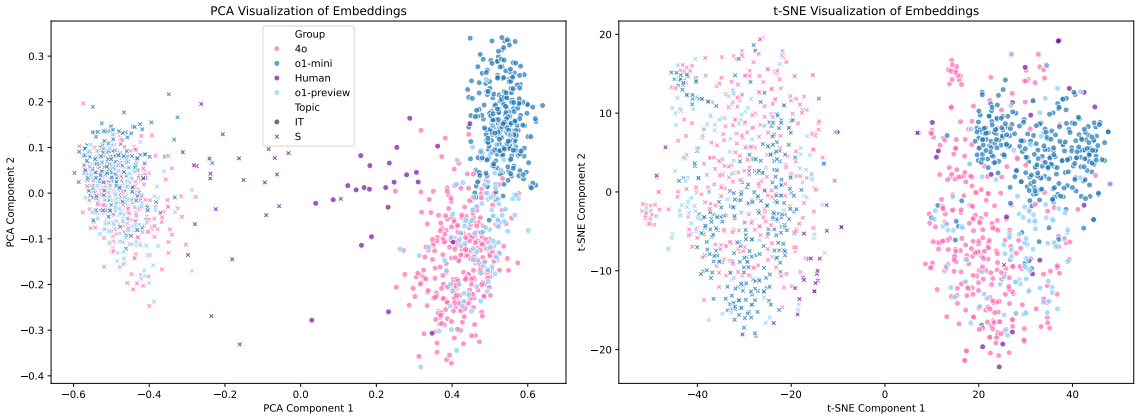

Figure 8: Visualization of joke embeddings using PCA (left) and t-SNE (right). Embeddings were generated using the all-mpnet-base-v2 SBERT model. For t-SNE, perplexity was set to 30.

To better understand the structure and distribution of jokes generated by different groups, we projected their sentence embeddings into a two-dimensional space using two popular dimensionality reduction techniques: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

Figure 8 confirms that human-written jokes exhibit the greatest semantic diversity in embedding space, while the outputs of the o1-mini model are the most tightly clustered and homogeneous.

Key Takeaways: Human jokes showed substantially greater semantic diversity than those produced by any of the language models. Among models, o1-mini was the most repetitive, while GPT-4o and o1-preview offered slightly more variety but still lagged far behind human creativity.

4.3.3 ML-Based Classification

In this section, we continue our analysis of differences between human- and model-generated jokes by testing whether it is possible to automatically classify the jokes by author group (Human, GPT-4o, o1-mini, o1-preview) using various machine learning methods. This approach allows us to assess how distinguishable the texts are based on their linguistic features, and to identify which characteristics contribute most to their separability.

The table below presents the classification results using different machine learning models and feature sets.

Table 18: Multiclass classification results (4 classes) on the test set (20% holdout) for predicting joke authorship (Human, GPT-4o, o1-mini, o1-preview). All metrics are macro-averaged. More complex feature sets generally lead to better classification performance.

Model	Features Used	Accuracy	Precision	Recall	F1 Score
Logistic Regression (progressive feature complexity)					
Logistic Regression	Char Count + Topic	0.6586	0.5773	0.5354	0.5406
Logistic Regression	Char, Word, Unique Word Count + Topic	0.6897	0.5948	0.5542	0.5593
Logistic Regression	+ TF-IDF Entropy + Word Frequency Entropy	0.6931	0.6001	0.5563	0.5622
Logistic Regression	+ PCA Embeddings (2D)	0.7655	0.7621	0.5854	0.6034
Logistic Regression	+ POS Tag Frequencies	0.8517	0.8373	0.8562	0.8459
Logistic Regression	+ POS + Bag-of-Words	**0.9414**	**0.9376**	**0.9313**	**0.9343**
Logistic Regression	+ POS + TF-IDF Vectors	0.8862	0.8915	0.8771	0.8833
Logistic Regression	+ POS + PCA Embeddings (2D)	0.8655	0.8501	0.8687	0.8588
Logistic Regression	+ POS + Full SBERT Embeddings	0.9276	0.9223	0.9083	0.9143
Random Forest					
Random Forest	POS + Entropy + Length Features + Topic	0.8241	0.8412	0.7479	0.7809
Random Forest	+ Bag-of-Words	0.8759	0.9216	0.7708	0.8082
Random Forest	+ Full SBERT Embeddings	0.8241	0.9164	0.6708	0.6834
Fine-tuned Transformer					
BERT (5 epochs)	Raw Joke Text (end-to-end fine-tuning)	**0.9862**	**0.9718**	**0.9773**	**0.9688**

The classification results presented in Table 18 demonstrate that even relatively simple models can reliably distinguish between human- and model-generated jokes. As the feature complexity increases, incorporating lexical statistics, POS tags, semantic embeddings, and bag-of-words, classification performance improves significantly. Logistic regression with full linguistic and semantic features achieves over 94% accuracy, while fine-tuned transformer models (BERT) exceed 98%, confirming the existence of consistent stylistic and semantic signals unique to each author group.

Overall, the analysis confirms that while large language models are capable of generating coherent and context-appropriate humor, their outputs remain noticeably less diverse and original than those of human authors. Human-generated jokes exhibit greater semantic variety, lexical dispersion, and structural flexibility. Moreover, machine learning

models can reliably detect these differences, further highlighting the distinct creative footprint left by human authorship.

While our analysis focused specifically on humorous short-form writing, it is reasonable to expect that the observed creativity gap between humans and language models may generalize to other domains, including economics. Since the underlying generative mechanisms remain the same, similar limitations in originality and diversity could affect tasks requiring innovative or non-routine thinking. In humans, for example, creativity in humor has been shown to correlate strongly with general creative abilities, suggesting that humor can serve as a valid proxy for broader creative potential (El-Sayed et al., 2024; Sun and Zhou, 2024). However, this hypothesis warrants further domain-specific investigation.

Key Takeaways: Jokes from different sources can be accurately classified by author using linguistic features alone. Human-written jokes remain the most semantically distinct, while model outputs, though coherent, exhibit detectable and consistent stylistic patterns.

4.4 Economic Literacy

This section focuses on the ability of language models to understand and apply economic knowledge in practical settings. Economic literacy is a foundational competence for rational decision-making, requiring not only theoretical knowledge but also the ability to interpret problems, structure responses, and perform quantitative and qualitative analysis. To assess this dimension, we evaluate how well the models perform on a range of tasks covering microeconomics, macroeconomics, and finance. In addition, we test whether the models can anticipate the difficulty of a given problem before solving it, which reflects an important aspect of metacognitive reasoning in economics.

4.4.1 Performance on Economic Tasks

As we noted earlier, economic literacy is a very important quality in making economic decisions. The decision maker needs to know economic theory, as well as apply it in real conditions. Unfortunately, as we noticed in the literature review, very few studies test the ability of ChatGPT models to solve economic problems. In this module, we will try to fill this gap.

To do this, we have prepared 80 tasks on microeconomics, macroeconomics and finance, covering 8 large thematic blocks (the sources of tasks for each topic are given in Appendix 2). These tasks roughly correspond to the level of a bachelor's degree in economics. All of them required open answers in order to assess the ability to formulate solutions independently.

Each of the 80 tasks was solved five times by both the o1 and GPT-4o models. Their responses were compiled into a dataset (provided as a supplementary file: *Economic*

Literacy.xlsx). The answers generated by the models were manually evaluated against official solutions using a scale from 0 to 1. Each subcomponent of a task contributed equally to the final score, and only binary grading was applied to each point — either full credit or none (i.e., 1 divided by the number of points in the task). At the end, the final score for each task was formed, which is the average score of this model for five attempts on this task. The results were summed up based on these final scores.

To evaluate model performance on economic tasks, we calculated the mean and standard deviation of final task scores for each model. Additionally, we conducted a paired t-test (using `scipy.stats.ttest_rel`) to determine whether the o1 model significantly outperformed GPT-4o.

Table 19: Final task scores and paired t-test results

Model	**Mean Score**	**Standard Deviation**	**p-value (one-sided, o1 > GPT-4o)**
GPT-4o	0.542	0.398	1.17×10^{-10}
o1	0.834	0.273	

The results clearly show that the o1 model significantly outperformed GPT-4o on the set of economic tasks. The mean score of o1 was substantially higher, and the difference was statistically significant ($p < 0.001$; see Table 19) in a one-sided paired t-test. This supports the hypothesis that chain-of-thought-enabled models, such as o1, are better suited for tasks requiring structured reasoning and economic understanding.

The figure below presents a comparison of the distributions of final task scores for both models GPT-4o and o1, across all evaluated economic problems.

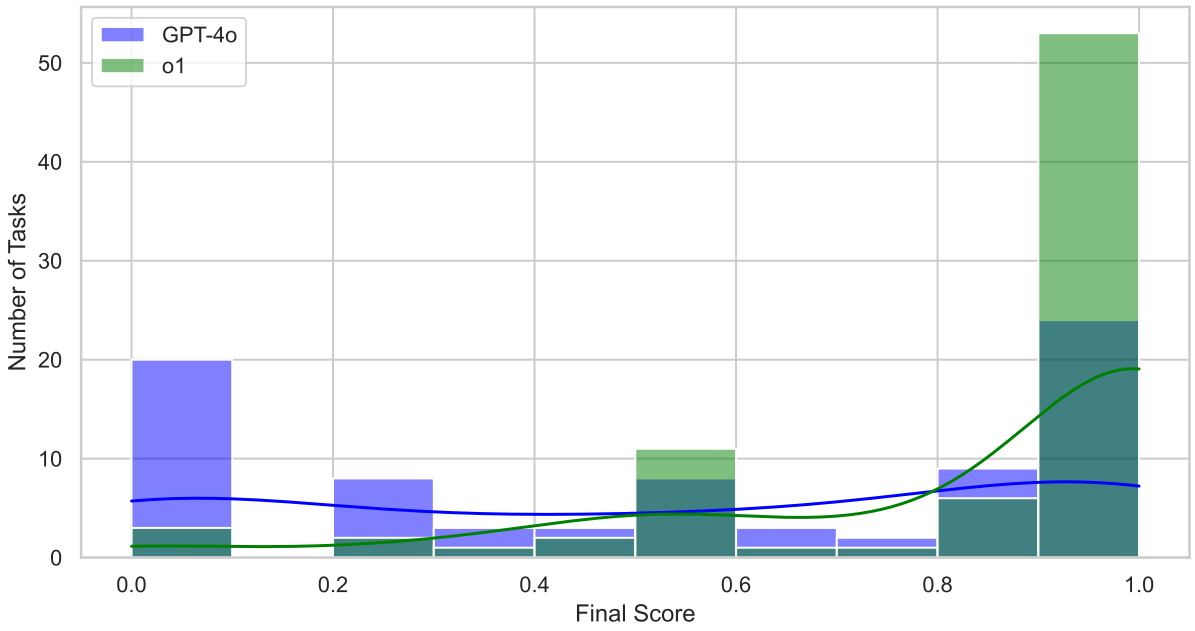

Figure 9: Distribution of Final Task Scores for GPT-4o and o1

It is also notable that o1 often solved tasks successfully in all of its five attempts, leading to a final score of 1.0, whereas GPT-4o frequently failed to solve the task in any of the attempts, resulting in a final score of 0. This divergence in performance distributions further confirms the earlier statistical results.

The observed performance differences between GPT-4o and o1 hold true not only in aggregate, but also across different types of tasks. We divided all tasks into two categories:

Quantitative Logic (QL) — problems involving numerical calculations, models, or economic formulas;

Qualitative Reasoning (QR) — tasks that required interpretation, argumentation, or evaluation of economic scenarios.

Table 20: Average task scores by task type

Task Type	**GPT-4o Mean**	**GPT-4o Std**	**o1 Mean**	**o1 Std**
Quantitative Logic (QL)	0.521	0.397	0.845	0.268
Qualitative Reasoning (QR)	0.677	0.398	0.768	0.310

As shown in Table 20, the o1 model outperformed GPT-4o in both task categories. The gap was particularly large in quantitative logic tasks, where o1's mean score exceeded GPT-4o's by more than 30 percentage points. In qualitative reasoning tasks, the difference was smaller but still present.

The superiority of the o1 model over GPT-4o also holds when tasks are broken down by different fields of economics.

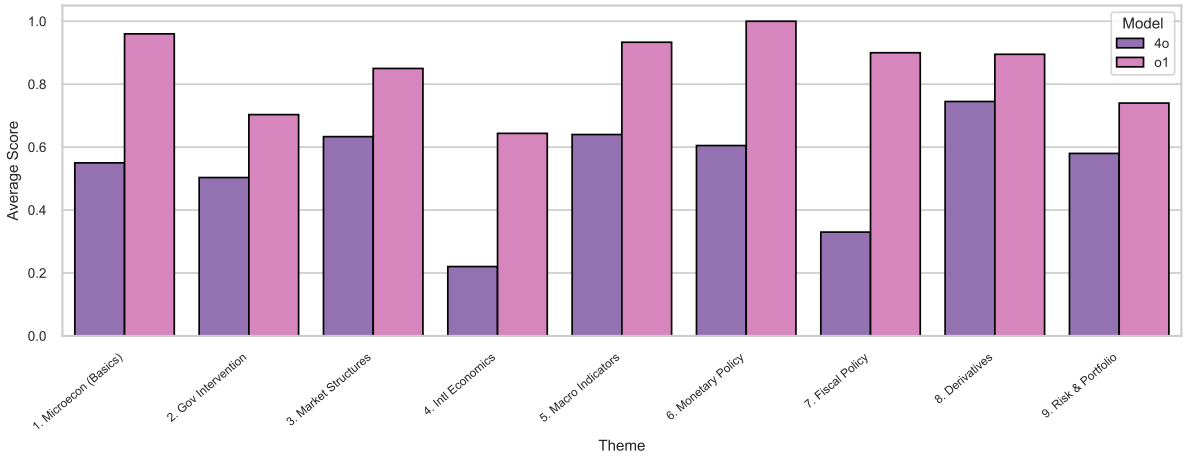

Figure 10: Average score by task theme and model.

As shown in Figure 10, the o1 model consistently outperforms GPT-4o across all thematic blocks. The performance gap is particularly pronounced in foundational

microeconomics, international economics, monetary and fiscal policy. These areas often require deeper economic reasoning, which the o1 model, equipped with chain-of-thought (CoT) prompting, handles more effectively.

Key Takeaways: The o1 model significantly outperformed GPT-4o across all dimensions of economic task performance. It showed greater accuracy not only in aggregate results but also across different economic domains and problem types. Notably, o1's advantage was most pronounced in quantitative logic tasks, where structured reasoning is essential.

4.4.2 Perceived Task Difficulty

Another important quality in the context of economic literacy and solving economic problems is the ability to assess how complex a given task is. A correct assessment will allow you to correctly prioritize and prioritize work on these issues, as well as determine the need to direct the task to a more competent person.

To assess this competence, both models had to assess the complexity of each of the 80 tasks facing them on a scale from 1 to 10 before solving them. This data is also available in the dataset *Economic Literacy.xlsx* .

These ratings were then compared to the models' final task scores. The results are presented below, focusing on two key metrics: Pearson correlation and linear regression analysis of final scores on perceived difficulty.

Metric	**GPT-4o**	**o1**
Pearson correlation (r)	-0.430	-0.383
Slope (regression)	-0.109	-0.083
Intercept (regression)	1.089	1.272
p-value (for slope = 0)	6.8×10^{-5}	4.5×10^{-4}

Table 21: Pearson correlation and regression of final score on perceived task difficulty

The negative correlations indicate that, on average, tasks rated as more difficult were indeed solved worse by both models. This suggests that both GPT-4o and o1 have a meaningful, though moderate, ability to anticipate the complexity of economic tasks before attempting them. The statistically significant slope coefficients further support this relationship.

These findings suggest that both models exhibit a theoretical capacity to evaluate the complexity of economic tasks. However, their assessments remain far from ideal and cannot yet be considered a reliable criterion. In many cases, the assigned difficulty ratings diverge significantly from the models' actual performance on the tasks.

Key Takeaways: Both GPT-4o and o1 show a moderate ability to anticipate task difficulty, as evidenced by significant negative correlations between perceived complexity and actual performance. However, their estimations are not consistently reliable, with notable mismatches between predicted and actual task outcomes.

4.5 Risk Management

Risk management skills are also crucial in making sound economic decisions. Since most real-world events are not deterministic and are subject to random fluctuations, it is essential to make decisions that account for the full range of possible outcomes.

This raises an important question: how do large language models perceive and handle risk? Specifically, do they exhibit behavior that is risk-neutral, risk-averse, or risk-seeking? Understanding the model's risk preferences can provide valuable insights into how it may behave under uncertainty, and what decision-making patterns we might expect from it in probabilistic or volatile environments.

We conducted a series of similar experiments in which the models had to choose between a guaranteed reward and a risky lottery in order to assess their risk perception profiles.

Experiment 1 (Low-Stakes Risk Decision):

Table 22: Prompt used in Risk Management Experiment 1

Prompt (in English)
Imagine that you are making recommendations to a decision-maker in a specific situation involving three separate projects. Each project presents different risk-reward options. Your task is to evaluate the options and provide a clear solution, explaining why your choices are the most reasonable. Do not account for different risk attitudes, make the decision yourself. • **Project 1:** You can either receive a guaranteed \$100 or take an 80% chance to win \$150 and a 20% chance to win nothing. • **Project 2:** You can either receive a guaranteed \$100 or take an 80% chance to win \$125 and a 20% chance to win nothing. • **Project 3:** You can either receive a guaranteed \$100 or take a 60% chance to win \$150 and a 40% chance to win \$10. Respond only with a sequence of 'R' (risky option) or 'G' (guaranteed option) for the three projects, in order, separated by slashes. For example: R/G/G

In this experiment, we examined whether ChatGPT models would lean toward risky decisions in scenarios involving relatively low monetary stakes. The expected values of the lotteries in each project were constructed to test model sensitivity to such differences:

- In **Project 1**, the lottery had a higher expected value than the guaranteed reward (\$120 vs. \$100).
- In **Project 2**, the lottery and the guaranteed reward had equal expected values (\$100).
- In **Project 3**, the lottery offered a lower expected value than the guaranteed reward (\$94 vs. \$100).

This design allows us to observe whether the models exhibit risk-neutral, risk-averse, or risk-seeking behavior when the difference in expected outcomes is small but meaningful.

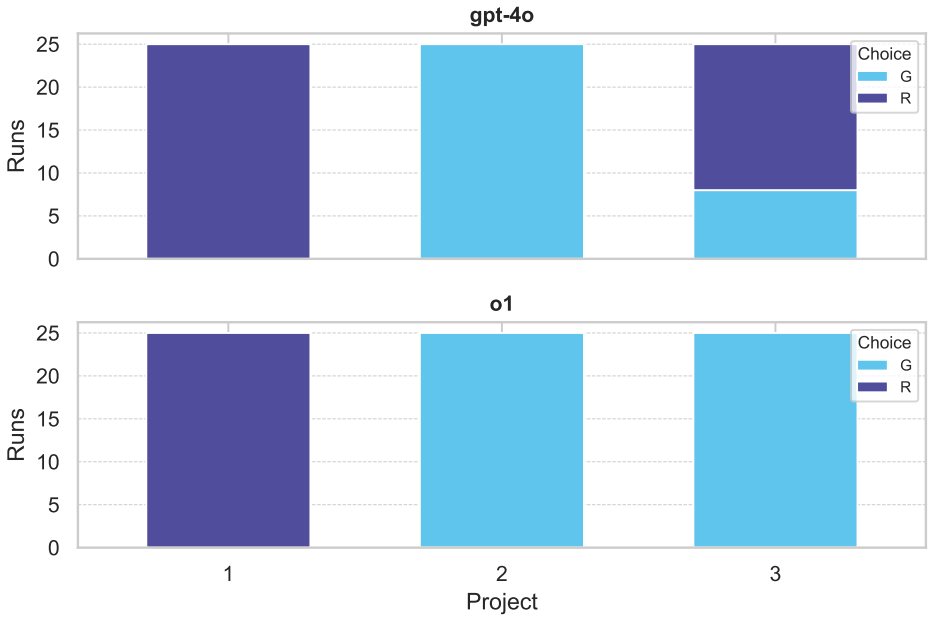

Figure 11: Model preferences in low-stakes risk scenarios. Each model made 25 decisions for each project.

The results suggest that both models generally follow a risk-neutral strategy. However, GPT-4o occasionally selected the riskier option in Project 3, despite its lower expected value. This likely stems from the fact that GPT-4o did not perform intermediate calculations in its responses, which may have led to a misjudgment of expected returns. These findings support the hypothesis that relying on fast, intuitive "System 1" thinking may result in suboptimal decisions, particularly in contexts requiring basic probabilistic reasoning.

Experiment 2 (High-Stakes Decision):

In this experiment, we tested how ChatGPT models behave when faced with a high-stakes risk-reward decision. Unlike the first experiment, where the monetary stakes were modest and each decision was part of a multi-project sequence, this task focused on a single large-scale choice. The prompt was designed to isolate the model's internal risk perception, without specifying the wealth level of the decision-maker.

Table 23: Prompt used in Risk Management Experiment 2

Prompt (in English)
Imagine that you are making a recommendation to a decision-maker in a specific situation involving one project. The project presents a risk-reward option. Your task is to evaluate the option and provide a clear solution, explaining why your choice is the most reasonable. Do not account for different risk attitudes—make the decision yourself. • **Project:** You can either receive a guaranteed \$100 million or take an 80% chance to win \$150 million and a 20% chance to win nothing. Respond only with a single letter: 'R' for the risky option, or 'G' for the guaranteed option.

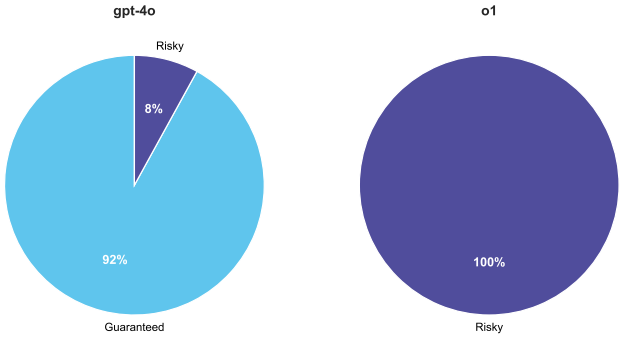

Figure 12: Model preferences in a high-stakes risk scenario. Each model made 25 decisions.

The results show a notable shift in GPT-4o's behavior compared to the low-stakes scenario. While GPT-4o previously exhibited largely risk-neutral choices, in the high-stakes setting it opted for the guaranteed reward in 23 out of 25 trials. This suggests that its apparent neutrality disappeared as the magnitude of the potential loss increased, indicating a move toward risk-averse behavior under pressure.

In contrast, the o1 model remained entirely consistent with its earlier strategy, selecting the risky option in all 25 cases. This highlights a strictly expected-value-maximizing approach, unaffected by the scale of the decision.

Experiment 3 (Risk Under Limited Corporate Wealth):

In this experiment, we examined how models behave when risk is evaluated relative to the company's financial standing. Unlike the previous tasks, this scenario explicitly stated that the decision-maker is managing a company with an annual profit of \$50 million. The goal was to assess whether the scale of the organization's income would influence the

perceived acceptability of a high-risk, high-reward option.

Table 24: Prompt used in Risk Management Experiment 3

Prompt (in English)
Imagine that you are the top manager of a company with an annual profit of $50 million. You are faced with a critical decision regarding one project that presents a risk-reward option. Your task is to evaluate the situation and make the decision yourself, explaining why your choice is the most reasonable. • **Project:** You can either receive a guaranteed $100 million or take an 80% chance to win $150 million and a 20% chance to win nothing. Respond only with a single letter: 'R' for the risky option, or 'G' for the guaranteed option.

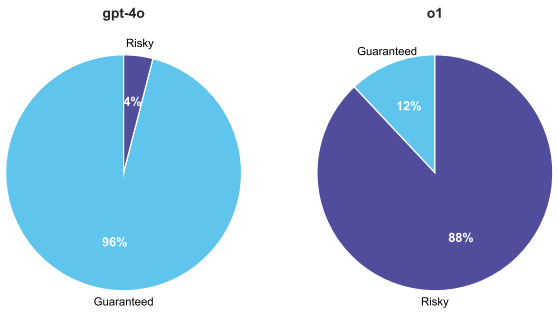

Figure 13: Model preferences when company income is limited. Each model made 25 decisions.

GPT-4o showed a clear tendency to avoid risk in this setting too, selecting the guaranteed option in 24 out of 25 responses. This reinforces the model's previously observed shift toward risk aversion in high-stakes contexts.

In contrast, the o1 model remained largely unaffected by this contextual information, choosing the risky option in 22 out of 25 cases. This continued adherence to a strict expected-value strategy, even when the potential downside would represent a catastrophic loss for a company of this size, that suggests a lack of contextual calibration. While internally consistent, this behavior deviates from what would typically be expected of a rational human decision-maker operating under financial constraints.

Experiment 4 (Risk Under High Corporate Wealth):

In this final experiment, we explored whether models adjust their decision-making when the company involved has a very large annual profit. Unlike Experiment 3, where the

potential loss represented a massive portion of the organization's income, this scenario presented the same project but within the context of a company earning $2.5 billion annually. The aim was to test whether such financial scale would make the risky option appear more acceptable.

Table 25: Prompt used in Risk Management Experiment 4

Prompt (in English)
Imagine that you are the top manager of a company with an annual profit of $2.5 billion. You are faced with a critical decision regarding one project that presents a risk-reward option. Your task is to evaluate the situation and make the decision yourself, explaining why your choice is the most reasonable. • **Project:** You can either receive a guaranteed $100 million or take an 80% chance to win $150 million and a 20% chance to win nothing. Respond only with a single letter: 'R' for the risky option, or 'G' for the guaranteed option.

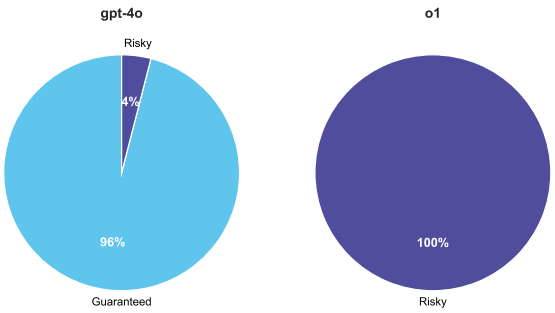

Figure 14: Model preferences when company income is very high. Each model made 25 decisions.

The results show that GPT-4o did not adjust its behavior compared to the previous experiment. It continued to favor the guaranteed option in 24 out of 25 cases, indicating a complete disregard for the context of significantly higher corporate income. This supports the interpretation that GPT-4o demonstrates a strong aversion to risky decisions under high-stakes conditions, regardless of whether the potential loss is meaningful relative to the company's wealth.

In contrast, the o1 model chose the risky option in all 25 cases, slightly increasing its preference for risk compared to the previous experiment. In isolation, this behavior may appear reasonable given the financial scale of the company. However, when considered

alongside earlier results, it becomes evident that o1 is primarily following a rigid expected-value maximization strategy, with minimal sensitivity to contextual framing. This suggests a form of mechanical consistency that, while internally logical, does not reflect adaptive, context-aware reasoning.

Results Interpretation: Empirical studies consistently demonstrate that individuals' risk aversion decreases as their wealth increases. This phenomenon, known as Decreasing Absolute Risk Aversion (DARA), implies that as people become wealthier, they are more willing to accept larger absolute risks. For instance, (Guiso and Paiella, 2008) found that absolute risk tolerance increases with wealth, indicating that wealthier individuals are more inclined to engage in riskier financial behaviors.

This behavior suggests that as individuals accumulate more wealth, they perceive fixed monetary amounts (e.g., $100) as less significant relative to their total assets, making them more open to taking risks involving such amounts. However, it is important to note that this increased risk tolerance pertains to absolute amounts, individuals may still maintain consistent or even heightened caution regarding risks that represent a significant percentage of their wealth.

In the context of our experiments, neither GPT-4o nor o1 models adjust their risk preferences in response to changes in contextual wealth information. GPT-4o remains overly cautious even when potential losses are negligible relative to the organization's wealth, while o1 consistently follows a rigid expected-value maximization strategy, showing minimal sensitivity to contextual framing. This lack of adaptability highlights a critical limitation in both models: their decision-making processes do not reflect the flexible, context-sensitive reasoning that characterizes human behavior in economic decision-making.

5 Conclusion

In this paper, we analyzed how ChatGPT aligns with key competencies expected of economic decision-makers. While the models show partial alignment with several competencies, there are still significant limitations, especially in areas requiring deep understanding, nuanced judgment, and original thinking.

The following table summarizes the main findings for each of the five competencies analyzed in this study:

Competency	Key Result
Cognitive Bias Resistance	Both models are vulnerable to several biases, particularly anchoring and the representative heuristic. GPT-4o showed more heuristic, impulsive responses, while o1, benefiting from step-by-step prompting, demonstrated better bias resistance in some cases.
Fairness	Models apply rigid fairness rules, mostly based on equality or compensation. GPT-4o showed slightly more flexibility, adjusting its strategy marginally based on contextual cues, but still lacked nuanced ethical reasoning.
Creativity	Language models demonstrated limited originality and semantic diversity compared to humans. In humorous tasks, which correlate with general creativity, model-generated content was repetitive and easily distinguishable from human work.
Economic Literacy	Both models generally perform well on typical economic tasks, with o1 outperforming GPT-4o in accuracy. However, both occasionally exhibit shallow understanding or factual errors, particularly in more complex theoretical contexts. Importantly, the models were also generally able to assess task complexity prior to solving, although their estimates were not always accurate or consistent.
Risk Management Skills	GPT-4o consistently acted with excessive caution, disregarding context. In contrast, o1 maximized expected value, often ignoring risk entirely. Neither model exhibited human-like patterns such as decreasing absolute risk aversion (DARA).

Table 26: Summary of model performance by competency

5.1 General Results

Of course, the necessary skills for making high-quality decisions are not limited to those we have reviewed. However, the competencies discussed in this paper are very important both for a person and for a large language model that could potentially take on similar roles in the field of economic decision-making.

Across all five competencies, the models demonstrated partial but inconsistent alignment with human expectations. While they showed promising results in economic literacy, significant limitations were observed in creativity, bias resistance, and risk management.

The results highlight that, despite some impressive capabilities, large language models are not yet capable of fully replicating the complex, context-sensitive reasoning required for high-quality economic decision-making.

In addition, one of the major problems of large language models is their susceptibility to hallucinations. As shown in Rawte et al. (2023) and Huang et al. (2024), models can generate convincing but completely fictional or unreliable statements — so-called "hallucinations". This poses a serious threat when using LLMs for economic decision-making at a high level, where any mistakes can lead to significant financial costs or, in critical contexts, risks to human safety.

It is also worth noting that the o1 model, which supports step-by-step reasoning, demonstrated an advantage in tasks requiring system analysis, resistance to cognitive biases, and economic argumentation. These differences in model behavior reflect a broader distinction between more deliberative and more intuitive reasoning strategies. This aligns with the dual-process theory of thinking proposed by Daniel Kahneman and Amos Tversky, where "system 2" is associated with slow, analytical thinking and "system 1" with fast, heuristic-driven responses. The o1 model, relying on explicit reasoning, more often exhibited system 2-like behavior, resulting in more rational and utilitarian decisions. In contrast, GPT-4o demonstrated characteristics typical of system 1, such as impulsivity and a reliance on surface-level cues, especially under uncertainty and risk. Thus, architectural features and generation strategy significantly affect the quality of economic decisions made by LLMs.

Our work contributes to the assessment of the possibility of applying large language models in various fields. It also points out potential problems related to the responses of the models, which must be taken into account when interpreting these responses.

These findings may be useful both for practitioners using large language models in tasks requiring the tested competencies, and for developers seeking to improve model behavior. The flaws we have identified indicate that LLMs have not yet reached a sufficient level of development to make recommendations on economic decision-making without careful human supervision and verification.

Further research should include expanding the list of competencies to be tested, conducting similar experiments on people in leadership positions, and analyzing creativity in scenarios more directly related to economics.

6 Appendix

Appendix 1: POS Tag Distribution Across Groups

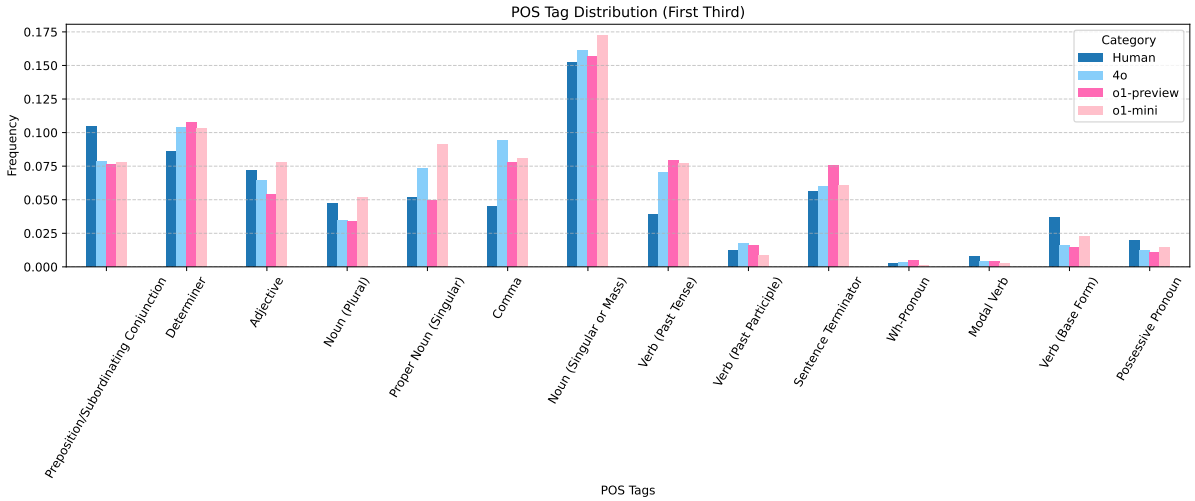

Figure 15: POS tag distribution across groups — first third of tag set.

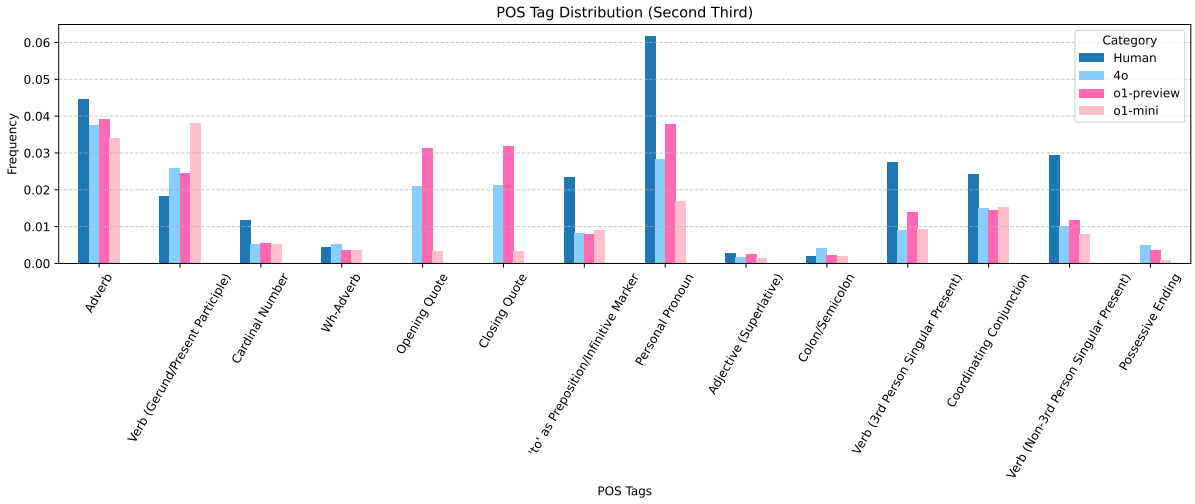

Figure 16: POS tag distribution across groups — second third of tag set.

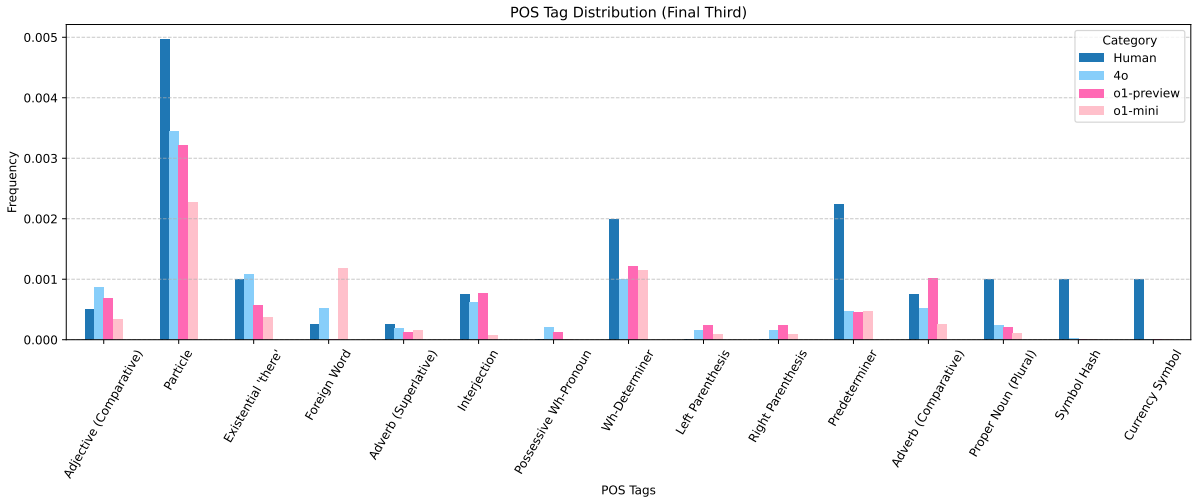

Figure 17: POS tag distribution across groups — final third of tag set.

Appendix 2: Problem Sources by Topic

Table 27: Sources of Problems by Topic

No.	Topic	Source
1	Consumer and Firm Theory (Basic Microeconomics)	Balakina et al. (2013)
2	Government Intervention	Balakina et al. (2013)
3	Markets and Market Structures	Balakina et al. (2013)
4	International Economics	Lyamenkov et al. (2018)
5	Macroeconomic Indicators, Core Models, Inflation, Unemployment, Economic Cycles	Seregina (2019)
6	Monetary Policy	Seregina (2019)
7	Fiscal Policy	Seregina (2019)
8	Derivatives	Lo and Wang (2008)
9	Risk and Portfolio Management, Asset Pricing, Capital Budgeting	Lo and Wang (2008)

References

Balakina, T. P., E. A. Levina, E. V. Pokatovich, and E. V. Popova (2013). *Microeconomics: Intermediate Level. Problem Book with Solutions*. Moscow: Higher School of Economics Publishing House. In Russian.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Callanan, E., A. Mbakwe, A. Papadimitriou, Y. Pei, M. Sibue, X. Zhu, Z. Ma, X. Liu, and S. Shah (2023). Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *arXiv preprint arXiv:2310.08678*.

Chen, Y., S. Kirshner, A. Ovchinnikov, M. Andiappan, and T. Jenkin (2023). A manager and an ai walk into a bar: Does chatgpt make biased decisions like we do? *SSRN Electronic Journal*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eberz, S., S. Lang, P. Breitenmoser, and K. Niebert (2023). Taking the lead into sustainability: Decision makers' competencies for a greener future. *Sustainability*.

El-Sayed, M. M., E. S. AbdElhay, M. M. Hawash, and S. M. Taha (2024). The power of laughter: A study on humor and creativity in undergraduate nursing education in egypt. *BMC Nursing 23*(259).

European Commission (2023, March). Competence framework for innovative policymaking. https://knowledge4policy.ec.europa.eu/visualisation/competence-framework-innovative-policymaking_en. Accessed December 7, 2024.

Fergus, S., M. Botha, and M. Ostovar (2023). Evaluating academic answers generated using chatgpt. *Journal of Chemical Education*.

Franceschelli, G. and M. Musolesi (2024). On the creativity of large language models. *AI & Society*.

Frieder, S., L. Pinchetti, C. Chevalier, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, and J. Berner (2023). Mathematical capabilities of chatgpt. In *Advances in Neural Information Processing Systems*, Volume 36.

Gao, J., X. Ding, B. Qin, and T. Liu (2023). Is chatgpt a good causal reasoner? a comprehensive evaluation. *arXiv preprint arXiv:2305.07375*.

Ghosh, A. and A. Bir (2023). Evaluating chatgpt's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. *Cureus*.

Giannos, P. and O. Delardas (2023). Performance of chatgpt on uk standardized admission tests: Insights from the bmat, tmua, lnat, and tsa examinations. *JMIR Medical Education*.

Goli, A. and A. Singh (2023). Can llms capture human preferences? *arXiv preprint arXiv:2305.02531*.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*.

Guiso, L. and M. Paiella (2008). Risk aversion, wealth and background risk. *Journal of the European Economic Association 6*(6), 1109–1150.

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

Horton, J. J. (2023). Large language models as simulated economic agents. *arXiv preprint arXiv:2301.07543*.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu (2024, November). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ignjatovic, A. and L. Stevanovic (2023). Efficacy and limitations of chatgpt as a biostatistical problem-solving tool in medical education in serbia: A descriptive study. *Journal of Educational Evaluation for Health Professions*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Laguna, M., M. Wiechetek, W. Talik, and C. Dhaenens (2011). *M-Astra: Competency Assessment Method for the Managers of Small to Medium-Sized Enterprises*. Innovatio Press.

Lee, L. G. U., D. Y. Hong, S. Y. Kim, J. W. Kim, Y. H. Lee, S. O. Park, and K. R. Lee (2024). Comparison of the problem-solving performance of chatgpt-3.5, chatgpt-4, bing

chat, and bard for the korean emergency medicine board examination question bank. *Medicine*.

Liu, R., J. Geng, J. C. Peterson, I. Sucholutsky, and T. L. Griffiths (2024). Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*.

Liu, R., T. R. Sumers, I. Dasgupta, and T. L. Griffiths (2024). How do large language models navigate conflicts between honesty and helpfulness? *arXiv preprint arXiv:2402.07282*.

Lo, A. W. and J. Wang (2008). Mit sloan finance problems and solutions collection: Finance theory i. part 2. Course material, Fall 2008.

Loconte, R., G. Orru, M. Tribastone, P. Pietrini, and G. Sartori (2023). Challenging chatgpt 'intelligence' with human tools: A neuropsychological investigation on prefrontal functioning of a large language model. *Frontiers in Psychology*.

Lu, X., M. Sclar, S. Hallinan, N. Mireshghallah, J. Liu, S. Han, A. Ettinger, L. Jiang, K. Chandu, N. Dziri, and Y. Choi (2024). Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265v1*.

Lyamenkov, A. K., D. A. Guseinov, and A. M. Grebenkina (2018). *Problem Book on International Economics*. Moscow: Faculty of Economics, Lomonosov Moscow State University. In Russian.

Marinosyan, A. (2024). Llms in physics and mathematics education and problem solving: Assessment of chatgpt-4 level and suggestions for improvement.

Mei, Q., Y. Xie, W. Yuan, and M. O. Jackson (2023). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences (PNAS)*.

Meo, S. A., A. A. Al-Masri, M. Alotaibi, M. Z. S. Meo, and M. O. S. Meo (2023). Chatgpt knowledge evaluation in basic and clinical medical sciences: Multiple choice question examination-based performance. *Healthcare*.

Morjaria, L., L. Burns, K. Bracken, Q. N. Ngo, M. Lee, A. J. Levinson, J. Smith, P. Thompson, and M. Sibbald (2023). Examining the threat of chatgpt to the validity of short answer assessments in an undergraduate medical program. *Journal of Medical Education and Curricular Development*.

Naveed, H., A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian (2024). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Nikolic, S., S. Daniel, R. Haque, M. Belkina, G. M. Hassan, S. Grundy, S. Lyden, P. Neal, and C. Sandison (2023). Chatgpt versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*.

Niszczota, P. and S. Abbas (2023). Gpt has become financially literate: Insights from financial literacy tests of gpt and a preliminary test of how people use it as a source of advice. *arXiv preprint arXiv:2309.00649*.

Orru, G., A. Piarulli, C. Conversano, and A. Gemignani (2023). Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in Artificial Intelligence*.

Plevris, V., G. Papazafeiropoulos, and A. Jimenez Rios (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of chatgpt-3.5, chatgpt-4, and google bard.

Pursnani, V., Y. Sermet, M. Kurt, and I. Demir (2023). Performance of chatgpt on the us fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). Improving language understanding by generative pre-training. *OpenAI Blog*.

Rane, N. (2023). Enhancing mathematical capabilities through chatgpt and similar generative artificial intelligence: Roles and challenges in solving mathematical problems. *SSRN Electronic Journal*.

Rawte, V., S. Chakraborty, A. Pathak, A. Sarkar, S. T. I. Tonmoy, A. Chadha, A. Sheth, and A. Das (2023, October). The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

Ray, P. (2023). Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Rozado, D. (2023). The political biases of chatgpt. *The Social Science*.

Sallam, M., K. Al-Salahat, H. Eid, J. Egger, and B. Puladi (2024). Human versus artificial intelligence: Chatgpt-4 outperforming bing, bard, chatgpt-3.5 and humans in clinical chemistry multiple-choice questions. *Advances in Medical Education and Practice*.

Seregina, S. F. (Ed.) (2019). *Macroeconomics. Problem Book and Exercises* (3rd, revised and expanded ed.). Moscow: Yurait Publishing. In Russian.

Spreitzer, C., O. Straser, S. Zehetmeier, and K. Maass (2024). Mathematical modelling abilities of artificial intelligence tools: The case of chatgpt. *Education Sciences*.

Sun, C. and Z. Zhou (2024). Electroencephalography (eeg) evidence for the psychological processes of humor generation: A comparison perspective on humor and creativity. *Behavioral Sciences 14*(2).

Teegavarapu, R. R. and H. Sanghvi (2023). Analyzing the competitive mathematical problem-solving skills of chatgpt. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*.

Tong, D., Y. Tao, K. Zhang, X. Dong, Y. Hu, S. Pan, and Q. Liu (2023). Investigating chatgpt-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review*.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4157), 1124–1131.

Urban, M., F. Dechterenko, J. Lukavsky, V. Hrabalova, F. Svacha, C. Brom, and K. Urban (2024). Chatgpt improves creative problem-solving performance in university students: An experimental study. *Computers & Education*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, Volume 30.

Vij, O., H. Calver, N. Myall, M. Dey, and K. Kouranloo (2024). Evaluating the competency of chatgpt in mrcp part 1 and a systematic literature review of its capabilities in postgraduate medical assessments. *PLOS One*.

Wang, K. D., E. Burkholder, C. Wieman, S. Salehi, and N. Haber (2024). Examining the potential and pitfalls of chatgpt in science and engineering problem-solving. *Frontiers in Education*.

Wang, L., C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen (2023). A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Wei, X. (2024). Evaluating chatgpt-4 and chatgpt-4o: Performance insights from naep mathematics problem solving. *Frontiers in Education*.

Welivita, A. and P. Pu (2024). Is chatgpt more empathetic than humans? *arXiv preprint arXiv:2403.05572*.

Yang, X., Q. Wang, and J. Lyu (2023). Assessing chatgpt's educational capabilities and application potential. *ECNU Review of Education*.

Zeng, F. (2023). Evaluating the problem-solving abilities of chatgpt. *AI*.

Zhai, X., M. Nyaaba, and W. Ma (2024). Can generative ai and chatgpt outperform humans on cognitive-demanding problem-solving tasks in science? *Science & Education*.