Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias

Itay Itzhak[1], Gabriel Stanovsky[2], Nir Rosenfeld[1], Yonatan Belinkov[1]
[1]Technion – Israel Institute of Technology
[2]School of Computer Science and Engineering, The Hebrew University of Jerusalem
itay1itzhak@gmail.com,
{nirr, belinkov}@technion.ac.il, gabriel.stanovsky@mail.huji.ac.il

Abstract

Recent studies show that instruction tuning (IT) and reinforcement learning from human feedback (RLHF) improve the abilities of large language models (LMs) dramatically. While these tuning methods can help align models with human objectives and generate high-quality text, not much is known about their potential adverse effects. In this work, we investigate the effect of IT and RLHF on decision making and reasoning in LMs, focusing on three cognitive biases—the decoy effect, the certainty effect, and the belief bias—all of which are known to influence human decision-making and reasoning. Our findings highlight the presence of these biases in various models from the GPT-3, Mistral, and T5 families. Notably, we find a stronger presence of biases in models that have undergone instruction tuning, such as Flan-T5, Mistral-Instruct, GPT3.5, and GPT4. Our work constitutes a step toward comprehending cognitive biases in instruction-tuned LMs, which is crucial for the development of more reliable and unbiased language models.[1]

1 Introduction

Advanced fine-tuning methods, like instruction tuning (IT) and reinforcement learning from human feedback (RLHF), have recently emerged as essential paradigms for improving the alignment of language models (LMs) with human objectives (Ouyang et al., 2022; Bai et al., 2022). Although widely adopted (Zhou et al., 2023), the specific cases in which IT and RLHF enhance model behavior to resemble human behavior, and the mechanisms involved in this process, remain unclear.

In this work, we delve into the impact of IT and RLHF on decision-making and reasoning in LMs. Recent studies highlighted to some extent cognitive-like biases in pretrained LMs (Binz and Schulz, 2022; Dasgupta et al., 2022) and instruction-tuned models (Hagendorff et al., 2022). We take a step further, exploring the consequences of IT and RLHF interventions on LMs' cognitive-like behavior.

We inspect three well-researched and fundamental biases: the decoy effect (Huber et al., 1982), the certainty effect (Kahneman, 1979), and belief bias (Evans et al., 1983). These biases reflect basic inconsistencies in human decision-making (decoy and certainty effects) and fallacies in logical reasoning (belief bias) that are prevalent, persistent, and consequential (Berthet, 2022; Acciarini et al., 2021).

The conventional approach to studying cognitive biases in humans is to design simple experiments that elicit from human subjects either judgments or decisions that are likely to reflect a target bias. Many of these experiments involve question answering; Table 1 shows examples of questions used in such experiments, illustrating how the responses of subjects can suggest biased behavior. To study cognitive-like biases in LMs, we adapt classic human experiments to an LM setting. Towards this, we create an experimental dataset using semi-automatic generated decision tasks: First, for each bias, we manually create an array of appropriate task templates containing flexible numeric and textual placeholder variables. Then, for a range of values and sets of alternatives, we generate a large collection of unique textual prompts, which we then use as queries to LMs. Following the classic experimental paradigm, in each experiment we partition the generated data into a 'control' dataset and a 'treatment' dataset, and define and measure the bias of a given LM as the average difference of its choices between the two datasets.

Within this setup, we empirically evaluate the

[1] https://github.com/itay1itzhak/InstructedToBias

Bias	Control	Treatment
Decoy	Below you will find three phone brands. Which one would you choose? Brand 1 - price is $130, quality rating is 40. Brand 2 - price is $350, quality rating is 60. Answer: `Brand 1.`	Below you will find three phone brands. Which one would you choose? Brand 1 - price is $130, quality rating is 40. Brand 2 - price is $350, quality rating is 60. Brand 3 - price is $438, quality rating is 60. Answer: `Brand 2.`
Certainty	Choose between: Option A - $4000 with a 20% chance, $0 with an 80% chance. Option B - $3000 with a 25% chance, $0 with a 75% chance. What is your choice? Answer: `Option A.`	Choose between: Option A - $4000 with an 80% chance, $0 with a 20% chance. Option B - $3000 with certainty. What is your choice? Answer: `Option B.`
Belief	Determine if the following argument is logically valid - All zint are thade. Some thade are snaff things. Conclusion: Some zint are snaff things. Answer: `This argument is invalid.`	Determine if the following argument is logically valid - All diamonds are gems. Some gems are transparent things. Conclusion: Some diamonds are transparent things. Answer: `This argument is valid.`

Table 1: Illustrative examples of the three evaluated Biases. Red text indicates disruptive elements fueling the bias. Blue text represents control responses unhindered by bias, while orange text denotes treatment responses influenced by bias. The **decoy effect** in the first row presents a scenario where two prize options are compared, the **certainty effect** in the second row involves selecting products with varying prices and quality measurements, and the **belief bias** in the third row entails evaluating the validity of logical syllogisms. In the certainty effect and decoy Effect, the model is tasked with choosing its preferred option, whereas in the belief bias, the model determines the conclusion's validity. Each bias is evaluated using a control and a treatment datasets. A shift in choice patterns is anticipated from model predictions on samples transitioning from the control dataset to the treatment.

degree of bias exhibited by several pretrained LMs, and compare them to their corresponding fine-tuned variants. Our findings indicate that applying IT or RLHF tuning either *introduces* cognitive-like biases into text generation, or *amplifies* these biases if they already exist.

Given that fine-tuned models are typically considered to be superior, our results point to an important limitation of tuning based on instructions or human feedback. Fine-tuned models are also often regarded as potentially *less biased*, such as in domains like gender or race, since they can be explicitly trained to avoid these biases or having personal preferences (OpenAI, 2023). Our results suggest that, similarly to debiasing attempts (Gonen and Goldberg, 2019), improving alignment with respect to one human objective may result in behavior that is unintended with respect to others.

2 Cognitive Biases: Background and Experimental Setup

Rational choice theory depicts humans as making choices in a manner that maximizes value on the basis of fixed preferences. A large body of literature is devoted to describing how actual human behavior deviates from this ideal. Cognitive biases aim to explain regular inconsistencies in choice behavior by revealing our susceptibility to 'supposedly irrelevant' factors, such as the context of the decision task, or its framing. Cognitive biases are therefore defined and measured by how judgments and decisions deviate from the rational or logical ideal in response to contextual changes.

Our research targets three biases that are both prevalent and well-established. The first two are the *decoy effect* and the *certainty effect* — decision-making biases that relate to special cases of the more general *prospect theory* (Kahneman, 1979), with each capturing one of its distinct aspects: the perception of value, and the perception of uncertainty. The third bias is the *belief bias*, a logical fallacy in judgment, which was previously observed in a closed-off pretrained model (Dasgupta et al., 2022).

In this section, we provide for each bias some general background and a description of its classic experimental setup, which we later build on.

2.1 Decoy Effect

Background. When choosing from a set of alternatives, a rational agent chooses the item having the highest intrinsic value. Human choices, however, are often affected by context, and in particular, by the set of available alternatives. For example, a decision maker who chooses A from the set $\{A, B\}$ may decide to choose B from the set $\{A, B, C\}$ – a behavior which cannot be consistent with any underlying preference ordering (McFadden, 1974).[2] The extreme case in which C is clearly inferior to both A and B, has been coined as the *decoy effect*, to portray C as a 'decoy' item whose only role is to shift the choice from A and B.

Experimental Setup. To study the decoy effect, we adopt the experimental setup of Huber et al. (1982), who proposed to measure how the choice between two items changes when a third *asymmetrically dominated* item—the decoy—is added to the choice set. Items in the experiment are described by their attributes (e.g., quality and price). In the control condition, subjects are asked to choose one item out of two comparable alternatives; in the treatment condition, an additional *Decoy Option* is added to the choice set. The decoy's attributes are set so that it is asymmetrically dominated (i.e., is worse in all dimensions) by one of the original items, referred to as the *target option*, but not by the other item, referred to as the *competitor option*. Table 1 (first row) provides a concrete example: Brand 1 and Brand 2 are comparable, whereas Brand 3 (the decoy) is inferior to Brand 2 (target), but not to Brand 1 (competitor).

Choice behavior is said to exhibit the 'decoy effect' if subjects tend to choose Brand 1 in the control condition, but prefer Brand 2 in the treatment condition. By design, this means that choices are affected by a supposedly irrelevant factor—the availability of an alternative that in itself will never be chosen, suggesting that choices are biased.

2.2 Certainty Effect

Background. Most decision settings involve some degree of uncertainty. Given a set of alternatives describing possible outcomes and their probability, utility theory (Friedman and Savage, 1948) determines that rational agents will choose the option with the highest expected value. Human choice, however, tends to deviate from this standard, especially when the probabilities to consider are either very small or very large. The *certainty effect* describes people's tendency to prefer outcomes that occur with certainty to alternatives that yield higher expected value, but include risk. This effect was initially explored in the seminal work of Kahneman (1979), whose experimental setup we describe next.

Experimental Setup. In Kahneman (1979), human subjects were asked to choose between two 'lotteries', each describing a simple distribution over potential monetary rewards (e.g., 80% to win \$100 and 20% to win nothing). In the control condition, subjects were given two lotteries A, B, each having some degree of risk; in the treatment condition, alternative B, having lower expected value, was modified to provide its original expected value but with a probability one (i.e., the same expected value but at no risk).

Table 1 (second row) presents an example. In both conditions, the prize in Option A remains the same and has a higher expected reward than Option B, whose certainty varies across conditions. As in the example, Kahneman (1979) (and many follow-up studies) found that, while control subjects tend to choose rationally, treatment subjects display a strong preference towards the certain alternative despite its lower expected reward.

2.3 Belief Bias

Background. Syllogisms are a class of reasoning problems involving two true statements and a third conclusion statement, which is either logically deductible from the true statements, or is not (Smith, 2022). To make a rational judgment of the conclusion, it is both necessary and sufficient to apply logical reasoning to the true statements—and to them alone. *Belief bias* occurs when a person's evaluation of the conclusions' validity is affected also by their own knowledge, beliefs, or

[2] A rational agent would necessarily choose either A or C.

Condition		Decoy Frying Pan	Phone	Car	Real-Estate	Certainty	Belief
Control	# Samples	96	120	120	96	336	672
Treatment	# Samples	1152	1440	1440	1152	504	672
Templates	# Prompts	4	4	4	4	3*	7
Values Range	US-Dollars	9.99-179.99	100-900	5K-35K	80K-500K	2.4K-5K	NA

Table 2: Sample and template counts in each dataset, along with the range of values for decoy products and certainty effect prizes. The different text templates and values were used to evaluate the biases robustly while using reasonable values and phrasing. The (*) notation for certainty effect templates denotes the primary textual prompt without sub-templates.

values, which can sometimes lead to false reasoning. This bias was empirically demonstrated by Evans et al. (1983), whose results suggest that human judgment can be affected by the 'believability' of the conclusions, i.e., that subjects' perception of logical validity depends on the degree to which the conclusion is believable (or not).

Experimental Setup. In Evans et al. (1983), human subjects were given sets of two premises and a conclusion, and asked whether the conclusion logically followed from the premises (Evans et al., 1983). Half of the conclusions were phrased to be *believable*—aligned with general world knowledge (e.g., "cigarettes are addictive"), and the other half was constructed to be *non-believable* ("cigarettes are non-addictive").

Table 1 (third row) shows an example. Both treatment and control tasks include two premises and an invalid conclusion; while the control includes fictitious objects, the treatment includes real-world objects—which in this case are believable, and entail an erroneous answer ('valid'). Evans et al. (1983) showed that subjects tended to consider believable conclusions as valid and unbelievable conclusions as invalid, suggesting the presence of belief bias in their judgments.

3 Data and Evaluation

We next describe our data generation process and evaluation scheme. Sec. 3.1, outlines our semi-automatic approach for generating specific datasets, each designed to probe a certain cognitive bias and to evaluate the existence of biased 'behavior' in models. Sec. 3.2 provides further details on these generated datasets. Sec. 3.3 formally introduces our proposed *bias score*, intended to quantify the degree of bias exhibited by a model based on its predictions on the generated data.

3.1 Data Generation

To assess the level of each bias in a model, we employ a comparative approach, as shown in Table 1. We do that by comparing predictions on a generated treatment dataset and a corresponding control dataset. For the decoy and certainty effects, we use data generated according to values crafted by us, and in the belief bias we use data generated in a similar fashion by Dasgupta et al. (2022) with additional text templates we wrote.

To generate the treatment datasets, we follow the experimental design for each bias as outlined in Section 2 and use new values that align with the cognitive experiments methods.

The control versions of the datasets are carefully crafted to closely resemble the treatment samples while excluding the specific attribute that triggers the bias, as identified by cognitive experiments.

In the decoy and certainty effects, for each sample, there exists a designated *Target* option. This option is expected to be chosen more frequently by a human (or a biased model) when presented with samples from the treatment dataset compared to samples from the control dataset. In the belief biases, we treat the correct answer as the *Target* option, for ease of notation. We later use the *Target* option to compute the bias scores, as detailed in Section 3.3.

3.2 Data Overview

Table 2 provides quantitative metadata for the datasets. We elaborate below on the text templates and values chosen for each bias dataset according to cognitive theory as outlined in Section 2.

We used 3, 4, and 7 prompt templates for the

		GPT3			T5		Mistral	
		LM	IT-LM		LM	IT-LM	LM	IT-LM
	Bias	DaVinci	DaVinci-002	DaVinci-003	T5	Flan–T5	Mistral	Mistral-I
Bias Score	Decoy Expensive	– 0.15*	– 0.13*	**– 0.02**	**0.02**	– 0.18*	0.03	**0.24****
	Decoy Cheaper	– 0.17*	– 0.08*	**0.08****	– 0.15*	**0.20***	– 0.05*	**– 0.03****
	Certainty	0.00	0.24*	**0.67***	0.09*	**0.17***	0.03	**0.29***
	Belief Valid	0.00	0.19*	**0.21***	– 0.03	**0.50***	0.01	**0.26***
	Belief Invalid	0.04	0.55*	**0.65***	0.03	**0.39***	0.05	**0.31***

Table 3: The difference between the choices of models in the target option under the treatment condition versus the control condition. A higher score means the model exhibits a higher level of bias. In bold are the highest values in each model family. (*) Marks results that are statistically significant with p-values $< .05$, and (**) marks results that are averaged across multiple products where some are significant and others are not. Mistral-I refers to Mistral-Instruct.

certainty effect, decoy effect, and belief bias, respectively. The certainty effect featured extra sub-templates with variations in option presentations like probabilities or percentages. All possible permutations of option orders were used for decoy and certainty effects, as well as for both premises in belief bias.

Regarding the decoy effect, we utilized realistic values from US-based store websites to construct our datasets. Quality ratings ranged from 60 to 90 with 10-20 intervals between options. Decoy options, in comparison to the target, exhibit a 25% or 50% price change, a 10-20 point quality rating shift, or a combination of both. Modern alternatives to the original products were anecdotally chosen, emphasizing a one-time, deliberate selection process without trial and error.

In line with cognitive bias theory, we chose certainty effect prizes and probabilities to closely mirror the cognitive research data, ensuring accurate expected utility differences between the options.

Belief bias samples involve manually composing both believable and unbelievable arguments, derived from previous work. The samples are evenly split, with half being believable and the remaining half being unbelievable. The samples' arguments are built upon simple, well-known objects, such as 'All guns are weapons' and 'All lizards are reptiles'. Further details can be found at Dasgupta et al. (2022)

3.3 Computing The Bias Scores

We assess biases in each model by analyzing their prediction patterns across treatment and control datasets, quantifying them through bias scores. The bias score captures the difference in the model's inclination towards the *Target* option between treatment and control scenarios.

For example, if the model chose the 'Target' option in 90% of treatment samples and 70% of control samples, the bias score would be 0.20.

Bias Score Definition. The bias score is formally defined in Equation 1, where *Treatment* and *Control* represent the sets of treatment and control datasets, respectively, and N_T and N_C indicate their respective set sizes. Ans_i denotes the model's choice in sample i, while T represents the target option.

$$\sum_{i \in Treatment} \frac{\mathbb{1}[Ans_i = T]}{N_T} - \sum_{i \in Control} \frac{\mathbb{1}[Ans_i = T]}{N_C} \tag{1}$$

According to the original experimental setting of the decoy effect, the target option in each sample can be associated with either a lower or higher price, leading to the computation of separate bias scores: *Decoy Cheaper* and *Decoy Expensive*.

To compute bias scores for the belief bias, we compare the model's predictions between consistent and inconsistent conditions for valid and invalid arguments. This analysis results in two distinct bias scores that were recognized in the original experiments:

Belief Valid: The difference between the model's predictions of consistent valid arguments (valid and believable conclusions in real-life objects condition) and neutral valid conclusions (all valid arguments in non-real object conditions).

Belief Invalid: The difference between the model's predictions of consistent invalid arguments (invalid and unbelievable arguments in real-life objects condition) and neutral invalid arguments (all invalid arguments in non-real object scenarios).

The Meaning of Bias Score Values. Higher bias score values indicate a greater degree of bias in the model. The bias scores range from -1 to 1, reflecting the extent of the bias and its direction relative to human biases according to cognitive theory. While the original experiments on human evaluation did not calculate bias scores, the intended alignment of these bias scores is with the strength of bias as per the cognitive theory on human biases. A score of 1 represents maximum bias aligned with human biases, 0 indicates no bias, and -1 denotes maximum bias in the opposite direction to human biases.

The significance of each bias score is measured using the student's t-test (Student, 1908).

4 Experiments

Models We conduct our experiments on two LM sets. The first set is pretrained models – GPT3 'DaVinci' (Brown et al., 2020), and the publicly available Mistral-7B (Jiang et al., 2023) and T5 (Raffel et al., 2020).[3] The second set consists of improved versions of the preatrained models fine-tuned using IT and human feedback. For GPT3, we experiment with GPT3.5 models—-'text-DaVinci-002' and 'text-DaVinci-003' ('Davinci-002' and 'Davinci-003' for short) (Ouyang et al., 2022)—as IT and IT+RLHF models respectively. For the Mistral 7B, we use Mistral 7B-Instruct with the recommended chat template[4]. For T5 we use the Flan-T5 models (Chung et al., 2022) as the IT version. Our primary findings are based on the XXL variant of the T5 models (11B parameters), and we also experiment with the XL variant (3B parameters) to investigate the influence of model size.

Finally, we also experiment with one of the latest commercially available models, GPT4 (OpenAI, 2023), which is considered a state-of-the-art generative model.[5] However, we do not have access to its pretrained version as it was not publicly released. We, therefore, use GPT4 only as a reference for a newer model.

Determining the Model's Answer. Given a prompt asking for a choice, the instruction-tuned models using greedy decoding usually generate text describing their choice, simply as "Option 1" or "Brand 2".[6]

To assess the pretrained performance for each task, we use the common practice (Brown et al., 2020) of evaluating the likelihood of various candidate answers from a predefined set of possible answers. This helps prevent models from persistently asking questions instead of providing direct answers, as observed in our initial experiments. This evaluation might be affected by a preference of the model to an answer given the context (e.g., given "Answer:" the model might give a higher baseline probability to "Option 2"). We apply the DC-PMI correction (Holtzman et al., 2021) that mitigates this issue by normalizing each answer likelihood within the context of the prompt, relative to a baseline prompt ("Answer:", in our case).[7]

Using Zero-shot The samples used for the decoy and certainty effects are choice-dependent questions with no "correct" answer (recall the examples in Table 1). It is therefore not obvious how to construct few-shot examples, which presumably should have correct labels in the prompt. Given that we focus on decision inclinations, the zero-shot setup aligns naturally with our investigation of all biases. Most experiments, unless specified, are in the zero-shot format and involve a single question followed by "Answer:" without extra examples, as shown in Table 1.

Using Few-shot Despite the above-mentioned problem, we experiment with an approach that constructs a few-shot setting using samples outside of our data. We build upon a recent work that suggested that giving few-shot samples without the correct labels could improve model performance by introducing the model with the overall format of the samples (Min et al., 2022). We detail further and report results in Section 6.1.

[3] We use version T51.1: github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md

[4] We used "You are a helpful assistant. Answer shortly with only your choice with no explanation." as the opening instruction.

[5] We used the 'gpt-4-0314' version with the content "You are a helpful assistant."

[6] In the certainty effect <5% of the predictions made by Flan-T5-XXL were not clear and we excluded these examples.

[7] Small-scale experiments with DC-PMI correction for the instruction-tuned models led to similar results to evaluation without correction, so we only report the latter.

	Bias	DaVinci-003	GPT4
Bias Score	Decoy Expensive	0.00	**0.38***
	Decoy Cheaper	0.03	**0.05**
	Certainty	**0.43***	0.20*
	Belief Valid	**0.20***	0.15*
	Belief Invalid	**0.47***	0.41*

Table 4: comparison of the results between GPT4 and the most recent GPT3.5 release DaVinci-003 in 1-shot format. Scores marked with * are statistically significant with p-values $< .05$

5 Results

Table 3 summarizes the bias scores of pre-trained models and their instruction-tuned and RLHF-tuned counterparts. We discuss the main take-aways in this section and provide several fine-grained analyses in the next one.

Models fine-tuned using IT and RLHF show a higher bias than their pretrained counterparts. Our findings reveal that the models fine-tuned on instructions and RLHF mostly exhibit significantly higher levels of bias compared to their pretrained counterparts, as demonstrated in Table 3. While the pretrained LMs demonstrate minimal to no bias, the fine-tuned models display pronounced biases across most categories. This is evident in the certainty effect row, where the DaVinci, T5 and Mistral pretrained models exhibit bias scores of 0.00, 0.09, and 0.03, respectively. In contrast, the fine-tuned models display higher bias scores of 0.24, 0.67, 0.17, and 0.29. This unexpected result suggests that the fine-tuning process, intended to enhance model performance, inadvertently introduces biases into the decision-making process.

We measured the significance of the differences between models using the difference-in-differences method (Dimick and Ryan, 2014). All differences were significant except for DaVinci and DaVinci-002 in the belief valid and in decoy expensive, T5 and Flan-T5 in the certainty effect, and Mistral and Mistral-Instruct in the decoy cheaper.

LMs exhibit biases that align with human biases theory. Intriguingly, our investigation reveals a convergence between the decision-making biases observed in the models and the well-established theory on irrational biases inherent in human decision-making processes. Recall from Section 3.3 that positive values indicate alignment between bias scores and human biases. Indeed, tuning using instructions or human preferences generally makes bias scores increasingly high. The negative bias score exhibited by DaVinci in the decoy biases can be explained by choice criteria which, unlike humans, are not value-depended. In this exceptional case, the model chose the last option offered almost all the time, regardless of the options' content, making its choice more focused on positional preferences.

This finding emphasizes the role of fine-tuning on bias amplification on previously undiscovered biases. In addition, the similarity between the theory on human biases and model biases highlights the potential connection of inherent biases ingrained in human decision-making processes to tuning methods that induce the models to replicate human behaviors.

IT Amplifies Biases. The discernible impact of fine-tuning with IT becomes evident upon comparing the T5 versus the Flan-T5 models and the Mistral versus the Mistral-Instruct models. While DaVinci and DaVinci-002 versions may differ by more than IT (exact details are not public), the transparent elucidation of the Flan-T5 fine-tuning process and the sole instruction tuning done to the Mistral-Instruct model allows us to confidently assert that the sole utilization of IT can indeed engender the emergence of biases. This finding highlights the influential role of fine-tuning methods in amplifying biases within models, shedding light on the intricate relationship between IT and the manifestation of biases.

RLHF Amplifies Biases. Our findings indicate that the application of reinforcement learning fine-tuning from human feedback has the potential to amplify biases within language models further. This is evident when comparing the DaVinci-002 and DaVinci-003 models, with the latter incorporating reinforcement learning techniques.[8] Notably, while IT may contribute to bias amplification, our results suggest that reinforcement learning, as an independent factor, can also play a significant role in the emergence of these biases. This observation highlights the complex interplay between reinforcement learning methodologies and

[8] According to OpenAI at https://platform.openai.com/docs/model-index-for-researchers.

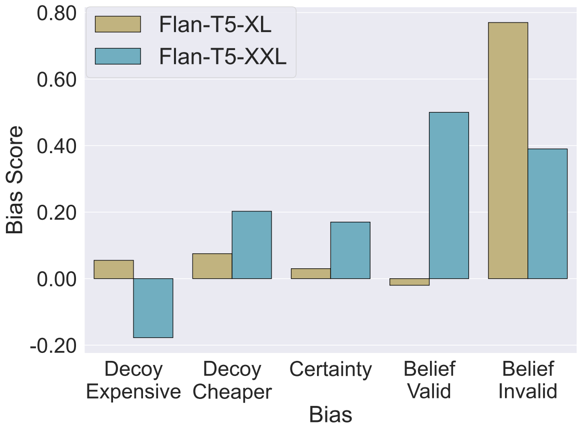

Figure 1: The impact of model size on bias scores. The larger Flan-T5-XXL exhibits higher bias scores in decoy cheaper, certainty, and belief valid biases while demonstrating lower bias scores in decoy expensive and belief invalid biases compared to the smaller Flan-T5-XL. The decoy expensive bias discrepancy may stem from Flan-T5-XXL's preference for higher-priced products, while the belief invalid bias reduction can be attributed to the model's enhanced accuracy with neutral arguments.

the manifestation of biases.

GPT4 is also biased. The results comparing GPT4 to its predecessor in the GPT series are presented in Table 4. Across our experiments, GPT4 demonstrates the highest bias score in the decoy expensive and decoy cheaper biases. Although the bias scores are lower in the certainty, belief valid, and belief invalid biases, GPT4 still exhibits significant bias levels.

The decreased bias scores observed in belief biases might be attributed to the model training, at least partly aimed at enhancing logical reasoning. Part of the GPT4 training data was designed to improve reasoning skills using data from MATH (Hendrycks et al., 2021) and GSM-8K (Cobbe et al., 2021). However, since GPT4 might be different in many other ways from DaVinci-003, we cannot attribute the decreased bias scores to this specific change. Beyond that, even with possibly improved reasoning, the model had less success mitigating bias in the decoy effect, which exhibited the most pronounced bias. Furthermore, we encountered instances in the zero-shot setting where GPT4 refrained from providing explicit choices, so we report one-shot results in the few-shot format as explained later in Section 6.1 (the zero-shot results when GPT4 did answer are similar to the one-shot results).

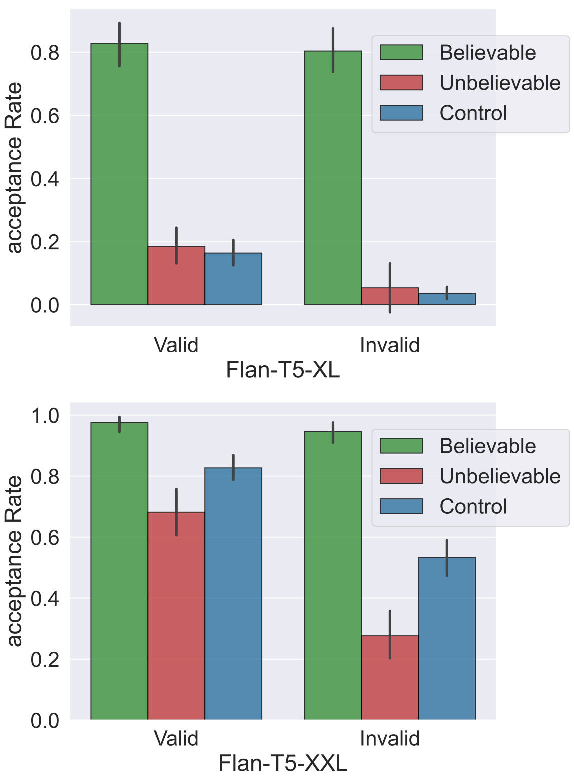

Figure 2: Acceptance rates of the Flan-T5 models on believable (green) and unbelievable (red) arguments in the treatmentcondition and on neutral arguments in the control condition (blue) divided into valid and invalid arguments. The Belief Invalid bias score for the larger Flan-T5-XXL model (lower) seems lower compared to the smaller Flan-T5-XL (upper) because the model is less successful on the neutral arguments (blue).

While GPT4 shows some mitigation of biases, the prominence of the decoy effect has increased, and all biases remain pronounced. These findings suggest that biases remain relevant in models designed to address bias mitigation, such as GPT4 which was trained using RLHF to avoid social biases such as biases about sexuality and norms around marriage (OpenAI, 2023).

The effect of model size on bias emergence. Figure 1 shows the discrepancy in bias scores between the XL and XXL versions of Flan-T5. Consistent with prior research on social biases (Tal et al., 2022), the larger XXL model exhibits higher bias scores for three bias types (decoy cheaper, certainly, and belief valid). Surprisingly, the decoy expensive and belief invalid bias scores are lower for the XXL model, suggesting a presumable reduction in bias compared to the XL model.

The reduction in belief invalid bias score could be attributed to the XXL model's lower accuracy in identifying invalid conclusions within the non-Real objects condition, as depicted in Figure 2. Specifically, in the invalid-believable condition, the XXL model demonstrates a higher acceptance rate, indicating a greater presence of bias. In contrast, in the invalid non-real objects condition, the XXL model displays a significantly elevated acceptance rate, leading to reduced overall accuracy and consequently lowering the bias score as per our defined calculation method (Section 3.3).

As to the reduction of bias score in the decoy expensive, that may result from a specific behavior of the XXL model, as discussed in Section 6.2.

6 Analysis

We delve into the effects of few-shot in Section 6.1 and explore different attributes of the decoy effect and belief bias in Sections 6.2 and 6.3.

6.1 Using Few-shot

Our main experiments used a zero-shot setting to avoid giving the model examples that could bias it in any direction — giving the model an example with an answer that is target could affect the tenacity of the model in choosing the target and vice versa. Therefore, to help the model understand the sample format without biasing it in either direction, we experiment with few-shot prompting without the original samples.

Instead of using the samples from our datasets, we use manually curated choices between arbitrary options for the decoy and certainty effects (e.g., "Which would you prefer, a white or black shirt?") and mathematical reasoning examples for the belief bias (e.g., "The price is $10 per soda. The customer inserted 20$. Conclusion: The customer can buy only 1 soda. Answer: Invalid."). We call this approach *format few-shot* as the intention is to show the model the sample format using few-shot examples. We curated a 5-example pool for each bias, randomly selecting each sample from them.

In the case of the belief bias, there *are* correct labels. Therefore, we can also prompt the model with few-shot samples and avoid biasing the model by utilizing samples comprised of neutral non-real objects derived from a distinct set of fabricated words that were deliberately excluded from the test data. We call this *Task few-shot* as few-shot samples are from the same task as the test sample. This approach enables us to assess the impact of using few-shot examples solely for formatting purposes compared to employing few-shot examples from the same task on the bias scores.

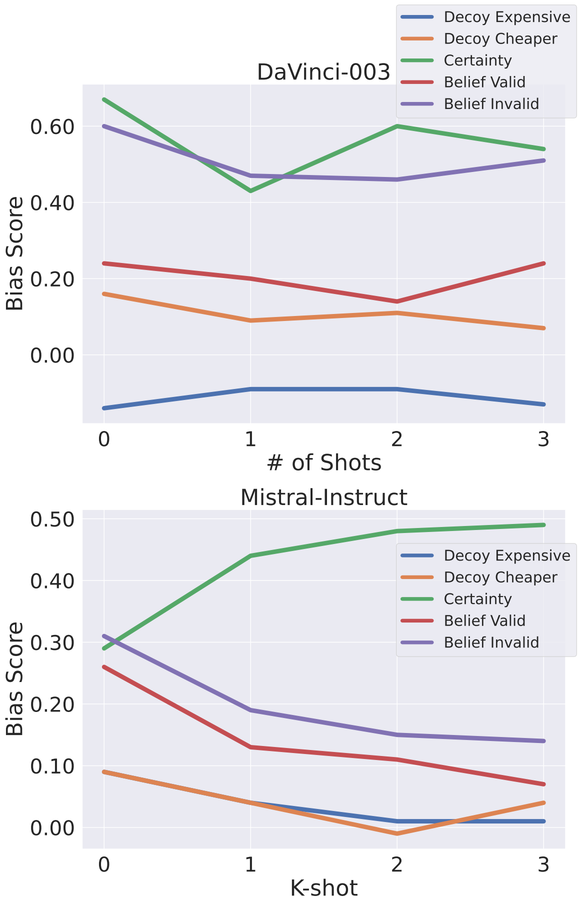

Figure 3: The impact of format few-shots on bias scores using Davinci-003 (top) and Mistral-Instruct (bottom). The utilization of few-shot examples in most models results in slightly lower bias scores, while in Mistral-Instruct Belief biases are significantly lower and certainty bias increases. To reduce computation costs, bias scores for Decoy Expensive and Decoy Cheaper biases are calculated solely on a specific product category (real-estate properties).

Results. Results with format few-shot examples can be seen in Figure 3. Regarding DaVinci-003, in the decoy and certainty effects, there is no distinct trend when using the format few-shot examples, except for a small decrease in bias score when changing from zero-shot to one-shot setting. Overall, increasing the number of few-shot examples might help the model understand the format but does not significantly decrease the bias.

In the case of the belief bias, incorporating few-shot examples leads to a noticeable reduction in the bias score, although a significant level of bias persists. This effect is more significant when uti-

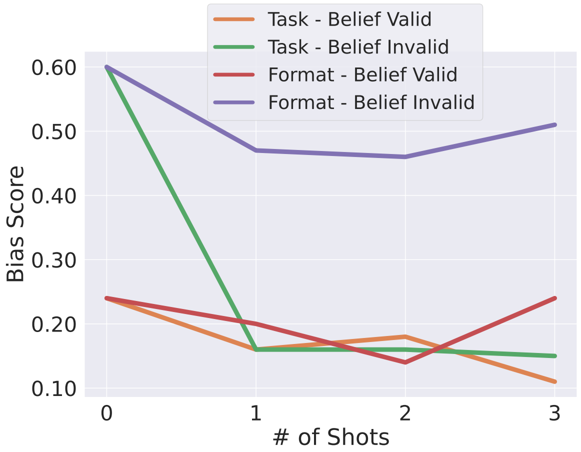

Figure 4: The impact of format few-shots in comparison to task few-shots on bias scores, utilizing the DaVinci-003 model. When the model is prompted with examples from the same task, the decrease in bias scores is relatively lower compared to employing examples with merely the same format as the task.

lizing task examples, as can be seen in Figure 4. This observation can perhaps be attributed to the presence of a logical reasoning process required by the belief bias examples, whereby the model's utilization of few-shot examples aids in facilitating problem-solving and helps to overcome the inherent bias associated with belief.

While these results are similar to the other instruction models we tested, Mistral-Instruct demonstrated a unique behavior. Its belief and decoy bias scores consistently decreased, while the certainty bias score increased with additional format examples. Notably, the pretrained version of Mistral also observed a rise in the certainty effect bias score in the few-shot setting (increasing from 0.03 to 0.31). This exception entails we have much to learn about the effect of pertaining data and training techniques on the way models utilize few-shot in general and regarding biases specifically.

This analysis focuses on the impact of few-shot examples solely on the instruction-tuned models, which exhibited the highest bias scores. However, it is plausible to speculate that the pretrained models, which demonstrated the lowest bias scores, could potentially benefit even more significantly from learning the format through few-shot examples, considering their stronger dependence on understanding the format. This could lead to the possible observation of higher bias scores for the pretrained models when giving few-shot samples. To address this, we conducted few-shot experiments for the pretrained models, which revealed that the bias scores remain similarly low for these models with the exception of Mistral, as described before.

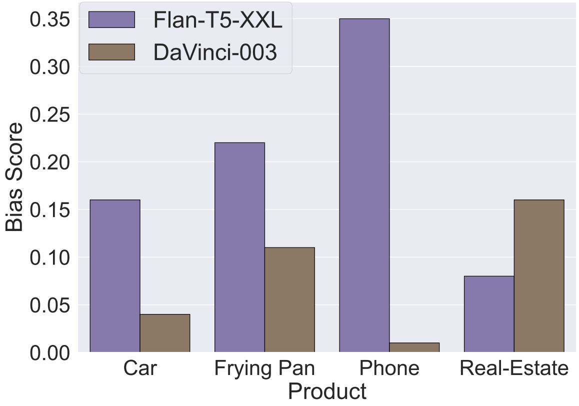

Figure 5: The bias scores of the decoy cheaper effect across various products for the Flan-T5-XXL and DaVinci-003 models. The bias scores exhibit consistency of bias existence across all products, indicating that the observed behavior remains more or less uniform within models across different product categories and price ranges, akin to human cognitive theory.

6.2 Decoy Effect Analysis

We investigate multiple attributes of the decoy effect, encompassing diverse product outcomes and price ranges, to assess their impact on the bias score and partly compare them with human behavior. Moreover, we explore a particular behavior identified in the decoy expensive effect that has a notable influence on the bias scores.

Products Variance. There was a moderate variance in bias scores in the decoy expensive results across different products, as shown in Figure 5.

As the bias scores are computed as the difference between the treatment and control conditions, the score varies with the variance in the base preference of the model to target option in the control condition for each product. Such differences between products were also observed to some extent in the original experiments on human subjects.

Together with the effect of price on the bias score (which is also analyzed in this section) and quality differences between products, the base preference of the model can cause a variance between different products.

Price Range Effect. We investigate the relationship between the target price and the price gap, defined as the difference between the prices of the competitor option and the target option. In our

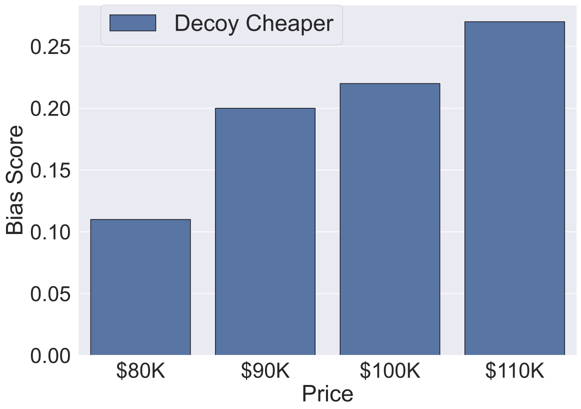

Figure 6: The Effect of price range on the bias score of the decoy cheaper bias with real-estate products in the DaVinci-003 model. The x-axis represents the target price, where an increase in the target price leads to a wider gap between the target and competitor prices. The bias score demonstrates a positive correlation with the increasing price gap.

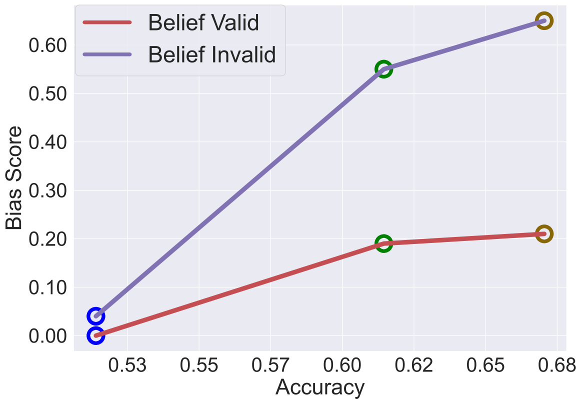

Figure 7: The relationship between bias scores and model accuracy on the belief bias task's logical reasoning aspect for the DaVinci (blue), DaVinci-002 (green), and DaVinci-003 (brown) models. Notably, an increase in model accuracy is accompanied by higher bias scores, indicating that improved accuracy does not necessarily mitigate biases in these models.

data, the higher the target value, the higher the gap from the competitor option. By selecting values with varying price gaps, we aimed to examine the impact of this factor on bias scores.

As Figure 6 shows, as the price range increases, the bias scores also exhibit higher values. Although human experiments did not analyze this aspect, this result aligns with the expected behavior of this bias and is intuitively reasonable.

The Decoy Expensive Effect. A notable observation in the decoy expensive experiments is the significantly low bias score of –0.18 exhibited by the Flan-T5 XXL model. We found that this score stems from the model consistently favoring the more expensive target option in the control condition with nearly 100% preference.

Considering the model's unwavering preference for the more target option in the absence of a decoy, the addition of a decoy option cannot possibly shift its preference from the competitor option to the target option. While this leaves no room for a bias score above zero, this preference for the more expensive option leads to negative results as the model picks the more expensive option even when adding a more expensive decoy option, leading to a shift from the target to the decoy option.

It is intriguing to observe such behaviors in these models that do not align with familiar cognitive biases but contradict basic human logic. These findings necessitate further investigation that goes beyond cognitive-like biases before utilizing these models to aid in human decision-making.

6.3 More Accurate and More Biased

On the belief bias task, we can quantitatively measure the model accuracy, allowing us to examine the trade-off between accuracy and bias scores.

Figure 7 shows the change in bias scores relative to the accuracy of the GPT models on the logical reasoning aspect of the belief bias task Interestingly, as the models demonstrate improved accuracy, they also exhibit higher bias scores. This finding suggests that, despite advancements in accuracy, biases persist within these models.

Finally, our evaluation includes GPT4, a model specifically trained on logical reasoning. GPT4 achieves a higher accuracy (84%) compared to all GPT models, while simultaneously exhibiting a lower bias score than DaVinci-002 and DaVinci-003 (0.07 and 0.49 for the belief valid and belief invalid correspondingly). This observation highlights the potential benefits of incorporating targeted training approaches to enhance both accuracy and mitigate biases in the process.

7 Discussion and Conclusions

Our study examines the influence of IT and RLHF on LMs' decision-making through cognitive bias analysis. We reveal the presence of these biases across models, notably in models amplified with IT and RLHF. These insights enhance our understanding of biases in fine-tuned models, widely considered superior to the pretrained models.

In Section 7.1, we explore the potential consequences of identifying these biases and the challenges in addressing them. Section 7.2 delves into research paths investigating the source of these biases in the training of language models (LMs).

7.1 Real-World Impact

The identified LM biases could impact real-world applications in decision-making and reasoning tasks. For example, the presence of decoy and certainty effects may raise challenges for LMs as decision assistants. Another impact could be reduced accuracy in some reasoning tasks in which the claim's plausibility plays a role. This concern is demonstrated by the fact that in the belief bias, the treated samples exhibit a notable decrease in accuracy as expressed by the bias scores, ranging from 19% to 61%, compared to controlled samples. Acknowledging and addressing these biases is crucial for enhancing the reliability and performance of LMs in real-world applications.

One possible way to estimate the impact of these biases is to test models on biased datasets, before and after tuning. Nevertheless, using our proposed datasets for this process has its challenges, due to the complexities of controlling the amount of bias in models. For example, fine-tuning with belief bias control data might not reduce model bias, while using belief bias treatment data could improve logical reasoning but harm common sense. These complexities increase when considering effects like decoy and certainty, which lack defined truth labels. Although fine-tuning with our data is an appealing idea, it requires further investigation into how biases are learned in LLMs, which is beyond the scope of this paper.

7.2 Origin of Bias

An additional question that arises concerns the origin of these biases. Further research is needed to determine if biases come from pretraining, intensify during fine-tuning, or arise from a mix of both. While Lin et al. (2023) claim that alignment methods only extract existing behavior models learned in pertaining, Shwartz and Choi (2020) demonstrated that pretrained LMs tend to prioritize infrequent actions over more common ones, indicating the presence of reporting bias. The biases outlined in our work may be associated with the prevalence of analogous questions and answers in the instruction and human feedback datasets used for model training. Studying bias-related examples in pretraining data and their magnification during fine-tuning can offer insights. Evaluating the influence of different fine-tuning data and strategies on bias could illuminate fine-tuning's role in bias emergence. Assessing how these biases interact with other known biases (such as reporting bias (Shwartz and Choi, 2020), and financial bias (Zhou et al., 2024)) can provide insights into how they are acquired and potential interconnections. Grasping these dynamics will guide strategies to improve model fairness and reliability.

8 Limitations

In examining the impact of IT and RLHF on cognitive biases in LMs, our study highlights a notable challenge in disentangling the effects of different training datasets. Flan-T5's IT data involves NLP tasks, Mistral-Instruct trained on unknown publicly available instructions datasets, while OpenAI's IT data uses assistant-like input-output pairs as far as we know. The dissimilarity in training data makes it difficult to pinpoint the exact factors causing biases in our models, underscoring the need for further investigation.

Beyond that, the unavailability of information on OpenAI models' training limits our ability to draw clear conclusions. Without details on their training procedures, we cannot determine whether RLHF training alone causes bias amplification or if GPT4's partial mitigation results from specific procedures, architecture differences, or other factors. The uncertain future availability of OpenAI models puts the complete reproduction of the results at risk for future research. Our study emphasizes the importance of transparency in model training for a better understanding of the relationship between IT and RLHF to biases in LMs.

Besides these model-specific limitations, there are limitations inherent in this type of research. One possible limitation is data contamination. We address well-known biases that might leak into the training data despite our efforts to introduce new text and value variations.

While it's common to evaluate pretrained LMs using answer probabilities (Brown et al., 2020; Holtzman et al., 2021), this evaluation method introduces a slight difference when compared to models trained on IT, which can be assessed based on their directly generated answers. Although unavoidable, this factor might influence results. We analyze the biases only in English-based models.

Acknowledgements

This research was supported by the Israel Science Foundation (grants 448/20 and 278/22), an AI alignment grant from Open Philanthropy, and an Azrieli Foundation Early Career Faculty Fellowship.

References

Chiara Acciarini, Federica Brunetta, and Paolo Boccardelli. 2021. Cognitive biases and decision-making strategies in times of change: a systematic literature review. *Management Decision*, 59(3):638–652.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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

Marcel Binz and Eric Schulz. 2022. Using cognitive psychology to understand gpt-3. *ArXiv*, abs/2206.14576.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Justin B Dimick and Andrew M Ryan. 2014. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama*, 312(22):2401–2402.

JSBT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

Milton Friedman and Leonard J Savage. 1948. The utility analysis of choices involving risk. *Journal of political Economy*, 56(4):279–304.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2022. Machine intuition: Uncovering human-like intuitive decision-making in gpt-3.5. *arXiv preprint arXiv:2212.05206*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Huber, John W Payne, and Christopher Puto. 1982. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, 9(1):90–98.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Daniel Kahneman. 1979. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:278.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*.

D McFadden. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robin Smith. 2022. Aristotle's Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Yuhang Zhou, Yuchen Ni, Xiang Liu, Jian Zhang, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. Are large language models rational investors?