Language models assign responsibility based on actual rather than counterfactual contributions

Yang Xiang,[1,*] **Eric Bigelow,**[1,2,*] **Tobias Gerstenberg,**[3] **Tomer Ullman,**[1,4,†] **Samuel J. Gershman**[1,4,†]
[1]Department of Psychology, Harvard University
[2]CBS-NTT Program in Physics of Intelligence, Harvard University
[3]Department of Psychology, Stanford University
[4]Center for Brain Science, Harvard University
[*]Equal contribution
[†]Equal senior authors

Abstract

How do language models assign responsibility and reward, and is it similar to how humans do it? We instructed three state-of-the-art large language models to assign responsibility (Experiment 1) and reward (Experiment 2) to agents in a collaborative task. We then compared the language models' responses to seven existing cognitive models of responsibility and reward allocation. We found that language models mostly evaluated agents based on force (how much they actually did), in line with classical production-style accounts of causation. By contrast, humans valued actual and counterfactual effort (how much agents tried or could have tried). These results indicate a potential barrier to effective human-machine collaboration.

Introduction

As we enter a world where AI systems collaborate with humans, it is important to understand the extent to which these systems think about collaboration in human-like ways. In particular, Large Language Models (LLMs) are becoming increasingly involved in human day-to-day work (George & George, 2023; L. Wang et al., 2024), sometimes even assisting with how people evaluate their human colleagues (Chiang & Lee, 2023; Dong et al., 2024). In this paper, we evaluate LLMs and compare their behavior with humans on a key aspect of collaborative cognition—the assignment of responsibility and reward in teams—by leveraging experimental paradigms and cognitive models adopted from two past studies.

When a team succeeds, who gets more credit, and who deserves more reward? Effective responsibility and reward allocation fosters motivation within teams. When collaborators feel that their efforts are recognized and fairly rewarded, they are more willing to make contributions to the team (Jo & Shin, 2025). However, how one's efforts are recognized and rewarded depends on who they are evaluated by—a human or an LLM. If LLMs are involved in determining how human collaborators are rewarded or punished, we need to understand how they do so, and whether their outputs match people's intuitions. There may be important differences between how LLMs and humans evaluate collaborators, and these differences might have substantial impacts in downstream applications.

Several factors predict how people assign responsibility and rewards to others. One line of work (Wolff, 2007; Greene et al., 2009; Nagel & Waldman, 2012) indicates that causal and moral judgments depend on the force a person exerts (how much they actually did). Collaborators who generate more output receive more reward (Baumard et al., 2012; Schäfer et al., 2023). Other work shows that effort (the proportion of a maximum force exerted) determines the amount of credit, blame, and punishment one deserves. For example, an effortful moral act leads to more credit, and an effortful immoral act leads to more blame (Bigman & Tamir, 2016; Jara-Ettinger et al., 2014; Sosa et al., 2021). Lack of effort is also punished more than lack of ability (Weiner, 1993). Note that here and elsewhere, we use "blame" and "credit" to mean responsibility in the event of failure and success, respectively. This terminology is consistent with past work, and blame and credit have been assessed on continuous scales (Gerstenberg & Lagnado, 2010, 2014; Gerstenberg, Ejova, & Lagnado, 2011; but see also Malle, Guglielmo, & Monroe, 2014).

Another line of work shows that people care about counterfactual contributions—how much a person *could* have done. For example, people simulate counterfactual alternatives when they judge causation and attribute responsibility (Gerstenberg, 2024). The same actual contributions can lead to different responsibility judgments depending on the structure of the task (Gerstenberg & Lagnado, 2010), the order of events (Gerstenberg & Lagnado, 2012), and the availability of other options (Wu & Gerstenberg, 2024). A player can get a disproportionately large reward if their contribution was critical to the team's success (Miller & Komorita, 1995; Gerstenberg et al., 2023).

We adapted materials from recent work on human responsibility judgment and reward allocation (Xiang, Landy, et al., 2023; Xiang et al., 2025), and instructed three LLMs (`GPT-4o-mini`, `GPT-4o`, `GPT-4`) to attribute responsibility and reward to agents in a collaborative task. We then compared the responses of LLMs to those of human participants as reported in the aforementioned studies. To understand the mechanisms underlying LLM responses, we compared them to seven cognitive models: three *actual-contribution* models (which base their judgments on the agent's actual force, strength, and effort), three *counterfactual-contribution* models (which base their judgments on how much effort the agent and their partner could have exerted), and an *ensemble* model that combines the actual effort and counterfactual effort models, and which has been shown to outperform the remaining six models in capturing human responsibility judgments (Xiang, Landy, et al., 2023).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

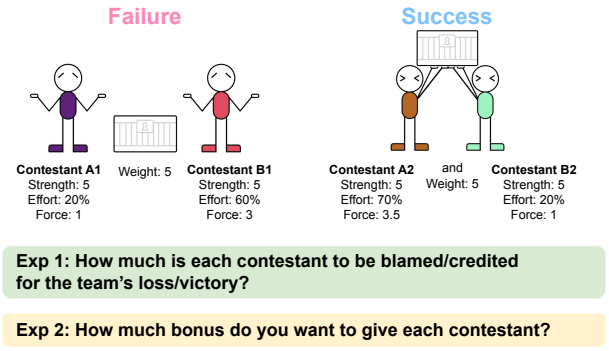

Figure 1: Experiment stimuli. Participants observed each agent's strength, effort, and force, as well as the weight of the box, and whether the agents failed (left panel) or succeeded in lifting the box together (right panel). Participants then assigned credit, blame, or reward to each agent. These images are adapted from Xiang, Landy, et al. (2023) Experiment 2a.

Experimental Paradigm

The experiments described a fictional game show, *BoxLifters*, where pairs of agents attempted to lift a box together. Each box had a weight W in the range $[1,10]$, and each agent a had a strength $S_a \in [1,10]$ defined as the maximum amount of force that they could exert. Each agent exerted a level of effort $E_a \in [0,1]$, defined as a fraction of their strength, and produced force $F_a \in [0,S_a]$, defined as their strength times their effort ($F_a = E_a S_a$). The agents succeeded at lifting the box when their combined force exceeded the box weight ($\sum_a F_a \geq W$), and failed otherwise ($\sum_a F_a < W$).

In the analyses below, we compare LLM responses on this task to seven cognitive models, as well as to human behavior.

Cognitive Models

The seven cognitive models we used were adapted from Xiang, Landy, et al. (2023); Xiang et al. (2025).[1] These models assign responsibility (blame B in the event of failure, and credit C in the event of success) and reward R to one of the two agents—the *focal agent*, denoted as a—at a time, by considering different factors. Three of them are *actual-contribution models* that base their decisions only on the focal agent's actual contributions (Force, Strength, and Effort models). Three of them are *counterfactual-contribution models* that base their decisions on counterfactual judgments about how much effort the focal agent and their partner—the *non-focal agent*, denoted as $/a$—could have contributed (Focal-agent-only, Non-focal-agent-only, and Both-agent counterfactual models). The last one is an Ensemble model that averages the Effort model and the Both-agent counterfactual model. The Ensemble model has been shown to outperform the other six models in capturing human responsibility judgments (Xiang, Landy, et al., 2023).

[1]The only change we made was replacing numbers like 10 and 1 with terms like F_{max}, S_{max}, and E_{max}.

Actual-contribution models

Force model (F). The Force model allocates responsibility and reward based on how much force an agent produces in the event. This model was inspired by production-style accounts of causality (Wolff, 2007; Greene et al., 2009), and by developmental research showing that children reward collaborators who generated more output (Baumard et al., 2012). Agents who exert more force are blamed less, credited more, and rewarded more:

$$\begin{aligned} B_a^F &\propto F_{max} - F_a \\ C_a^F &\propto F_a \\ R_a^F &\propto F_a \end{aligned} \tag{1}$$

Strength model (S). The Strength model allocates responsibility and reward based on an agent's strength. This model was inspired by past work showing that stronger agents are blamed more for failures (Gerstenberg et al., 2011). Although there weren't significant effects for credit allocation, it is natural to attribute success to a stronger person, especially if the strength difference is huge (e.g., an adult and a toddler lifting a box together). So, stronger agents receive more credit and reward for successes, and receive more blame and less reward for failures:

$$\begin{aligned} B_a^S &\propto S_a \\ C_a^S &\propto S_a \\ R_a^S &\propto \begin{cases} S_a & \text{if } L = 1 \\ S_{max} - S_a & \text{if } L = 0 \end{cases} \end{aligned} \tag{2}$$

Effort model (E). The Effort model allocates responsibility and reward based on the level of effort an agent exerts. This model was inspired by past work finding that greater effort in performing moral acts leads to more credit and reward, whereas lack of effort is punished (Bigman & Tamir, 2016; Weiner, 1993; Jara-Ettinger et al., 2014). Agents who exert more effort are credited and rewarded more, and blamed less:

$$\begin{aligned} B_a^E &\propto E_{max} - E_a \\ C_a^E &\propto E_a \\ R_a^E &\propto E_a \end{aligned} \tag{3}$$

Counterfactual-contribution models

Central to the counterfactual-contribution models is the concept of *difference making* (Icard et al., 2017): whether the outcome could have been different if the agents had exerted a different level of effort E'. Inspired by prior work (Sanna & Turley, 1996), here we consider directional counterfactuals (upward for failures, downward for successes). In other words, when agents fail, we consider what would have happened if they exerted more effort; when agents succeed, we consider what would have happened if they exerted less effort.[2] Specifically, we consider counterfactual efforts drawn

[2]Past work has proposed other ways of constructing counterfactuals; for example, Gerstenberg et al. (2021) proposed a noisy model

from discrete uniform distributions in increments of 0.01, where $E' \in (E, 1]$ when agents fail and $E' \in [0, E)$ when agents succeed. The responsibility and reward an agent receives hinge on the probability that they or their partner could have changed the outcome.

Each agent's probability of changing the outcome is defined as:

$$P_a = \begin{cases} \sum_{E'_a} P(E'_a)\, \mathbb{I}[E'_a S_a + F_{/a} < W] & \text{if } L = 1 \\ \sum_{E'_a} P(E'_a)\, \mathbb{I}[E'_a S_a + F_{/a} \geq W] & \text{if } L = 0, \end{cases} \quad (4)$$

where $\mathbb{I}[\cdot]$ is an indicator function that returns 1 if its argument is true, and 0 otherwise. The term $F_{/a}$ denotes the force of the group excluding the contribution of agent a.

Focal-agent-only counterfactual model (FA). The Focal-agent-only counterfactual model only considers counterfactual actions on the part of the focal agent. The model assigns responsibility and reward based on the likelihood of the focal agent changing the outcome by altering their effort allocation, while holding the non-focal agent's effort allocation fixed:

$$\begin{aligned} B_a^{FA} &\propto P_a \\ C_a^{FA} &\propto P_a \\ R_a^{FA} &\propto \begin{cases} P_a & \text{if } L = 1 \\ 1 - P_a & \text{if } L = 0 \end{cases} \end{aligned} \quad (5)$$

In other words, if the focal agent could have easily changed the outcome, they would get more credit and reward in the event of success, and more blame and less reward in the event of failure.

Non-focal-agent-only counterfactual model (NFA). The Non-focal-agent-only counterfactual model only considers counterfactual actions of the non-focal agent. If the non-focal agent could have easily changed the outcome, the focal agent would get less credit and less reward in the event of success, and less blame and more reward in the event of failure:

$$\begin{aligned} B_a^{NFA} &\propto 1 - P_{/a} \\ C_a^{NFA} &\propto 1 - P_{/a} \\ R_a^{NFA} &\propto \begin{cases} 1 - P_{/a} & \text{if } L = 1 \\ P_{/a} & \text{if } L = 0 \end{cases} \end{aligned} \quad (6)$$

Both-agent counterfactual model (BA). The both-agent counterfactual model considers counterfactual actions of both the focal agent and the non-focal agent by averaging the predictions of the Focal-agent-only model and the Non-focal-agent-only model. As in Xiang, Landy, et al. (2023); Xiang et al. (2025), we assign equal weighting to the two components for simplicity:

$$\begin{aligned} R_a^{BA} &\propto (R_a^{FA} + R_a^{NFA})/2 \\ B_a^{BA} &\propto (B_a^{FA} + B_a^{NFA})/2 \end{aligned} \quad (7)$$

of Newtonian physics that samples counterfactuals from a Gaussian distribution centered on what actually happened. Note that here we are not making a strong claim about how counterfactuals are constructed.

In doing so, this model considers both factors within the focal agent's control (what they themselves could have done differently) and factors outside their control (what their partner could have done differently).

Ensemble model (EBA)

The last model is an Ensemble model that combines the Effort model (E) and the Both-agent counterfactual model (BA), hence the acronym EBA. The Ensemble model was designed to address the insufficiency of the six models above in explaining human responsibility judgments. Theoretically, its two components can have different weights; however, past work has found that the two models have similar weights in human responsibility judgments (Xiang, Landy, et al., 2023). Here, we stick with the same equal-weighting Ensemble model to be consistent with past work and avoid adding free parameters to the model:

$$\begin{aligned} B_a^{EBA} &\propto (B_a^{E} + B_a^{BA})/2 \\ C_a^{EBA} &\propto (C_a^{E} + C_a^{BA})/2 \\ R_a^{EBA} &\propto (R_a^{E} + R_a^{BA})/2 \end{aligned} \quad (8)$$

Note that none of the cognitive models have free parameters; therefore we did not need to fit any of the models to the data.

Experiments

In order to examine how LLMs attribute responsibility and reward, and allowing for potentially different underlying mechanisms, we conducted two experiments. Experiment 1 instructed LLMs to assign responsibility to agents, whereas Experiment 2 instructed LLMs to assign reward. The full prompts we used, along with code and data used for experiments and analyses are available at https://osf.io/b5yz4/.

Experiment 1: Responsibility judgments

Methods We used an experimental design (Figure 1) adopted from work studying human responsibility attribution and reward allocation (Xiang, Landy, et al., 2023; Xiang et al., 2025). The stimuli consisted of 55 unique combinations of strength, effort, force, and box weight. In every scenario, the two agents were matched along one dimension—strength, effort, or force. This helped tease apart the three actual-contribution models.

We converted experiment instructions and questions to a long-form text format, without images, and used it to prompt LLMs. Each prompt specified the strength, effort, and force of each contestant, the weight of the box, and whether the agents successfully lifted it. Each prompt ended with a question. When the agents failed, the question was "How much is each contestant to be blamed for the team's loss". When the agents succeeded, the question was "How much is each contestant to be credited for the team's victory". The LLMs were instructed to reply with a number between 0 and 10 indicating how much blame or credit they would assign to each agent (0 meant no blame/credit, 10 meant very high blame/credit). In

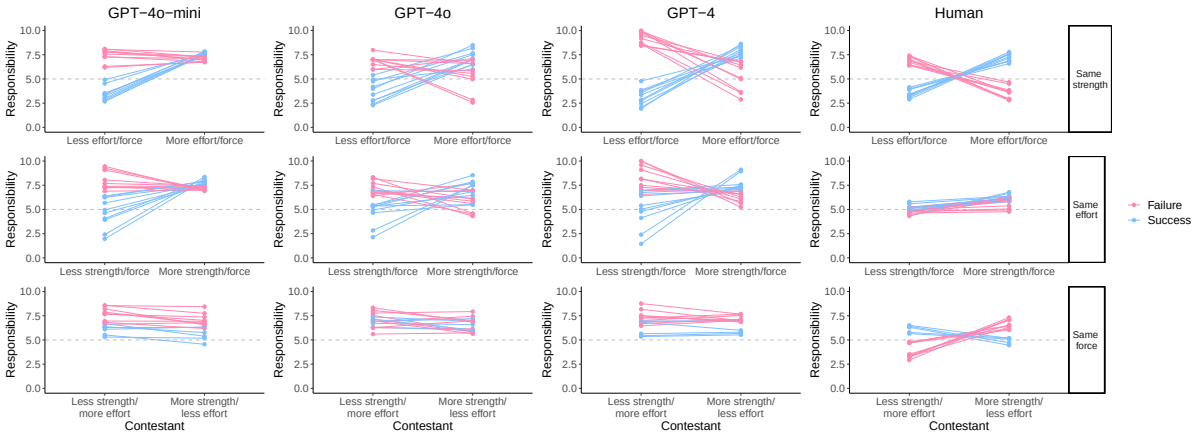

Figure 2: LLM responsibility attributions compared to humans (Xiang, Landy, et al., 2023). Each line corresponds to a single scenario. LLMs tend to assign similar amounts of responsibility to agents whose forces are matched, whereas humans tend to assign similar responsibility to agents whose efforts are matched. Error bars indicate bootstrapped 95% confidence intervals.

order to ask about both agents, referred to as "Contestant A" and "Contestant B", we instructed the LLMs to evaluate a single agent (A or B) at a time. We also flipped the order of A and B to avoid ordering bias. As a result, every scenario was prompted 4 times: two agents $\times$ two orderings.

We tested three LLMs available in the OpenAI API: `gpt-4o-mini-2024-07-18` (GPT-4o-mini), `gpt-4o-2024-11-20` (GPT-4o), and `gpt-4-0125-preview` (GPT-4). While LLM model details are not publicly available, GPT-4 is presumed to have the most parameters of the three LLMs. GPT-4o and GPT-4o-mini are comparatively newer, have fewer parameters, and are multi-modal (language and vision). GPT-4o-mini is smaller than GPT-4o and also less capable. We used the OpenAI API due to the availability of token logit probabilities ('logprobs'), which reduced the cost of our experiments. Token logit probabilities are the likelihood that the LLM would have generated each possible next token—in our case, integers from 0 to 10, e.g. $p(\text{'5'})$ or $p(\text{'10'})$. With logit probabilities, we can estimate the distribution of possible LLM outputs with only a single query. We aggregated these into a weighted average over integers; for example, if a response was 40% '5' and 60% '10', the response would be coded as $40\% \times 5 + 60\% \times 10 = 8$. These weighted averages were used as the LLM responses in our analyses.

Results Figure 2 shows the LLM responses. In each row, we consider scenarios where both agents were matched to have either equal strength, equal effort, or equal force. Note that when one of these is fixed, the other two are confounded: for example, when two agents are matched for strength, the agent that puts forth more effort will also exert more force. When the agents were matched for strength (top row), GPT-4o, GPT-4, and humans all assigned more credit to the agent who exerted more effort and force, and more blame to the agent who exerted less effort and force. GPT-4o-mini was similar in how it assigned credit for successes, but for failures it assigned equal blame to both agents, regardless of effort/force. This suggests that neither LLMs nor humans attribute responsibility based solely on collaborators' strength. When the agents were matched for effort (middle row), the LLMs all assigned more credit to the stronger and more forceful agent, and more blame to the weaker and less forceful agent. This could be explained by LLMs evaluating responsibility based on either force or strength; however, given that they didn't assign responsibility based on strength, LLMs were likely making responsibility judgments based on force. By contrast, humans assigned similar amounts of blame and credit to both agents when effort was equal. As argued by Xiang, Landy, et al. (2023), this supports the view that humans evaluate collaborators based on effort. When the agents were matched for force (bottom row), all LLMs assigned similar amounts of blame and credit to both agents, which suggests that LLMs evaluate collaborators based on force rather than strength or effort. Conversely, humans in this case assigned more credit to the agent who was weaker but exerted more effort, and more blame to the agent who was stronger but exerted less effort.

Next, we computed the correlations between LLMs' responses and human data. As shown in Table 1, GPT-4o-mini showed strong correlations with human credit allocation, but was unable to explain blame allocation in human responses. GPT-4o's responses were weakly correlated with human blame allocation, and strongly correlated with human credit allocation. GPT-4's responses were moderately correlated with human blame allocation, and highly correlated with human credit allocation. Considering the three LLMs ordered from least to greatest number of parameters, we see a steady progression: GPT-4's responsibility allocation behavior is most correlated with humans, followed by GPT-4o, and then GPT-4o-mini.

We then computed the correlations between each LLM and the seven cognitive models, visualized in Figure 3 (top row).

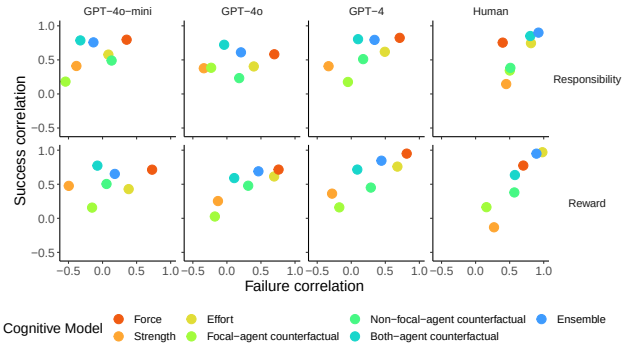

Figure 3: Comparing human and LLM responses to seven cognitive models. x-axis: Pearson correlation coefficients when the collaboration failed. y-axis: Pearson correlation coefficients when the collaboration succeeded. Points falling closer to the top right indicate better models for explaining the data. Overall, LLMs responses are best captured by the Force model, while human responses are best described by the Ensemble model for responsibility attribution and the Effort model for reward allocation.

Points closer to the top-right of this plot indicate models that better explain the data. We found that the Force model best explains all LLMs' responses both when the agents failed ($r = .35$ for GPT-4o-mini, $r = .70$ for GPT-4o, and $r = .71$ for GPT-4) and when they succeeded ($r = .80$ for GPT-4o-mini, $r = .58$ for GPT-4o, and $r = .82$ for GPT-4) when jointly taking both correlations into account. By contrast, human responses were best described by the Ensemble model both when agents failed ($r = .92$) and succeeded ($r = .90$). The similarities between LLMs and humans shown by the correlations in Table 1 are thus driven by distinct underlying mechanisms, which are only noticeable when we examine the computations that generate the behaviors.

In summary, while LLMs primarily use force to assign responsibility, humans primarily use actual and counterfactual effort.

Experiment 2: Reward judgments

Methods Experiment 2 used the same experimental setup and stimuli as Experiment 1. The only difference was that the LLMs were instructed to give bonus rewards to the agents, which none of the agents were aware of beforehand (so they couldn't have behaved strategically). The question that ended the prompts was "How much bonus do you want to give each contestant?", and the LLMs were instructed to reply with a number between 0 and 10 indicating how much bonus reward to give each agent.

Table 1: Correlation between LLM data and human data.

	GPT-4o-mini	GPT-4o	GPT-4
Exp 1 Responsibility			
Failure	−.11	.20	.46
Success	.89	.79	.90
Exp 2 Reward			
Failure	.38	.70	.65
Success	.55	.71	.81

Results Figure 4 shows the LLM responses. Across all models and conditions, we see an intercept shift between Failure and Success conditions. This means that, similar to humans, the LLMs we studied judged that agents who failed the task deserve less reward.

When the agents' strengths were matched (top row), similar to humans, all LLMs assigned more bonus to the agent who exerted more effort and more force. When the agents' efforts were matched (middle row), GPT-4o-mini assigned similar rewards to agents when they failed, but more reward to the stronger and more forceful agent when they succeeded. GPT-4o and GPT-4 assigned more reward to the stronger and more forceful agent regardless of the outcome of the collaboration. This suggests that the LLMs likely did not base their decisions on effort, since they reward collaborators differentially when they exert the same effort. By contrast, humans allocated the same reward to both agents when they put forth the same effort. When the agents' forces were matched (bottom row), all LLMs assigned similar amounts of reward to agents who varied in strength and effort. This supports the hypothesis that the LLMs used force, instead of effort, as their basis for judgments. By contrast, humans assigned more reward to the agent who was weaker but exerted more effort. As shown in Table 1, GPT-4o and GPT-4 were moderately-to-highly correlated with human reward allocation, GPT-4o-mini was moderately correlated with human data.

Human reward allocations were best captured by the Effort model. By contrast, we found that that the Force model best describes the responses of all three LLMs, both when the agents failed ($r = .73$ for GPT-4o-mini, $r = .76$ for GPT-4o, and $r = .81$ for GPT-4) and when agents succeeded ($r = .71$ for GPT-4o-mini, $r = .72$ for GPT-4o, $r = .95$ for GPT-4). By contrast, human responses were best described by the Effort model both when agents failed ($r = .98$) and succeeded ($r = .97$). These patterns are visualized in Figure 3 (bottom row). Points falling closer to the top right indicate better models for explaining the data. As with Experiment 1, we see that LLM and human reward allocations were driven by different underlying mechanisms. While LLMs use force as the single metric for assigning reward, humans use effort.

General Discussion

We compared LLMs' responsibility attribution and reward allocation to seven cognitive models. We found that LLMs' responses were best captured by a Force model that evaluates collaborators based on how much they actually contributed. By contrast, humans evaluated collaborators based on their actual and counterfactual effort (Xiang, Landy, et al., 2023; Xiang et al., 2025).

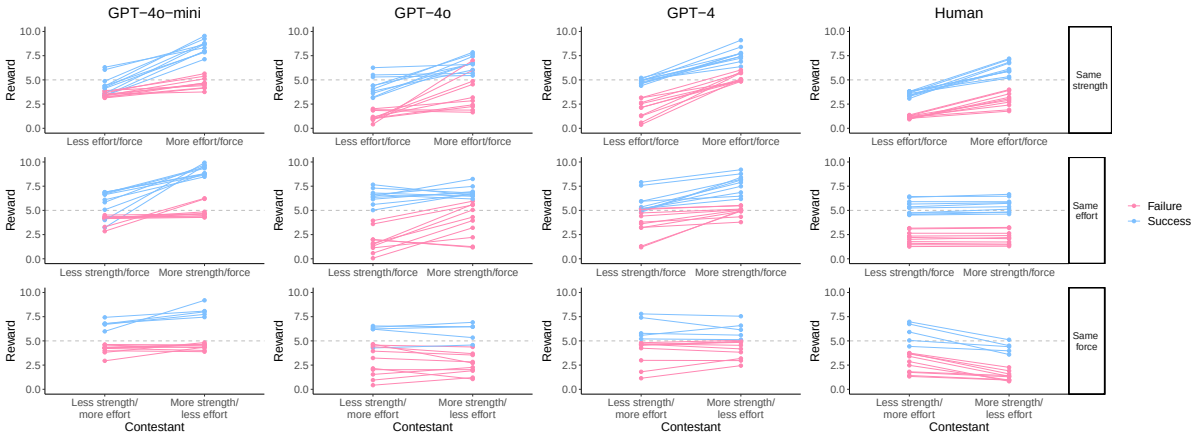

Figure 4: LLM reward allocations compared to humans (Xiang et al., 2025). Each line corresponds to a single scenario. LLMs tend to assign similar amounts of reward to agents whose forces are matched, whereas humans tend to assign similar reward to agents whose efforts are matched. Error bars indicate bootstrapped 95% confidence intervals.

Using force to evaluate collaborators is in line with production-style theories, which consider causal processes as directly producing another event (Dowe, 2000; Talmy, 1988; Wolff, 2007; Greene et al., 2009). According to this view, forces produced by agents determine the outcome of the collaboration, so force should be the basis for judgments. This is a natural metric for systems that only care about the likelihood of a certain outcome being produced, which might explain why LLMs reasoned this way. However, this social reasoning bias could be problematic if people expect LLMs to reason about collaborations the same way humans do.

Our results raise questions about why LLMs behave differently from humans in evaluating collaborators. One reason why humans use effort is that it signals a person's desire to contribute. There is considerable evidence showing that cooperative traits are deeply valued by humans (Cottrell et al., 2007; Raihani & Barclay, 2016; Bird et al., 2012; Bird & Power, 2015; Hackel et al., 2015), that people are sensitive to collaborators' effort (Xiang, Vélez, & Gershman, 2023), and the sensitivity to effort can be traced back to as early as infancy (S. Liu et al., 2017). These preferences are so deeply rooted in us that we exhibit them even in one-shot interactions (Delton et al., 2011) and anonymous games (Hagen & Hammerstein, 2006). Additionally, people who are more willing to exert effort are more likely to get better over time through training and learning (Xiang et al., 2024), which makes them better collaborators in the long run. The value of effort is likely ingrained through cultural evolution (Henrich & Boyd, 2016; Henrich & Muthukrishna, 2021) and might not be learned by LLMs that are trained to predict text from the internet, which in turn has many idiosyncratic biases and may emphasize the importance of results rather than effort. It is also possible that LLMs may have a more fundamental limitation in their capabilities, in particular related to the capacity to reason counterfactually. LLMs are capable of some forms of counterfactual reasoning (X. Liu et al., 2024; Z. Wang, 2024); however, it may be that this capability is limited, or does not extend to our scenarios (Zečević et al., 2023; Schulze Buschoff et al., 2025). Alternatively, LLMs may simply care less about counterfactuals, or do not recognize the relevance of counterfactuals for evaluating collaborators. Our results establish that LLMs prioritize force over effort, but future work is needed to understand why this happens.

Understanding why LLMs value force over effort might shed light on how we can steer LLMs to be more human-like in their responses. One intriguing direction is to add context in the LLM's instruction prompt that emphasizes the importance of effort (e.g., by specifying that agents will continue to collaborate in the future) or counterfactual relevance. This research will also benefit from experimenting with LLM architectures outside the GPT series and using open-source LLMs, such as LLaMA, Qwen, Mixtral, and Gemma, which will enable analyzing the impact of number of parameters, training data, and model architecture on LLM behavior.

In addition to understanding how LLMs assign responsibility and rewards, this domain is also particularly useful for studying the mechanisms underlying complex behavior in LLMs. Because the computational cognitive models in this domain are relatively well-developed, we can use them to analyze LLM responses beyond merely evaluating their overall accuracy and examining superficial similarities and dissimilarities to human responses. These models formalize theories of cognition that are specific and interpretable, offering high-level insights and expressing something more general about cognition that extends beyond particular experimental domains. Analyses such as ours serve to empirically test these cognitive theories. Our work paves the road to more theory-driven research into the processes driving LLM behavior and offers an exciting opportunity to understand the similarities and differences between natural and artificial minds, and to foster more effective human-machine collaboration.

Acknowledgments

This work was funded by a Hodgson Fund grant from the Harvard University Department of Psychology and supported by the Kempner Institute for Natural and Artificial Intelligence, and a Polymath Award from Schmidt Sciences. EB was supported by an internship with NTT Research and TG was supported by grants from Stanford's Human-Centered Artificial Intelligence Institute (HAI) and Cooperative AI.

References

Baumard, N., Mascaro, O., & Chevallier, C. (2012). Preschoolers are able to take merit into account when distributing goods. *Developmental psychology*, *48*(2), 492.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of experimental psychology: general*, *145*(12), 1654.

Bird, R. B., & Power, E. A. (2015). Prosocial signaling and cooperation among martu hunters. *Evolution and Human Behavior*, *36*(5), 389–397.

Bird, R. B., Scelza, B., Bird, D. W., & Smith, E. A. (2012). The hierarchy of virtue: mutualism, altruism and signaling in martu women's cooperative hunting. *Evolution and Human Behavior*, *33*(1), 64–78.

Chiang, C.-H., & Lee, H.-y. (2023). Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? a sociofunctional perspective on the importance of different valued characteristics. *Journal of personality and social psychology*, *92*(2), 208.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, *108*(32), 13335–13340.

Dong, Y., Jiang, X., Jin, Z., & Li, G. (2024). Self-collaboration code generation via chatgpt. *ACM Transactions on Software Engineering and Methodology*, *33*(7), 1–38.

Dowe, P. (2000). *Physical causation*. Cambridge University Press.

George, A. S., & George, A. H. (2023). A review of chatgpt ai's impact on several business sectors. *Partners universal international innovation journal*, *1*(1), 9–23.

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*.

Gerstenberg, T., Ejova, A., & Lagnado, D. (2011). Blame the skilled. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(6), 936–975.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attribution. *Psychonomic Bulletin & Review*, *19*, 729–736.

Gerstenberg, T., & Lagnado, D. A. (2014). Attributing responsibility. *Oxford Studies in Experimental Philosophy, Volume 1*, 91.

Gerstenberg, T., Lagnado, D. A., & Zultan, R. (2023). Making a positive difference: Criticality in groups. *Cognition*, *238*, 105499.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical population biology*, *69*(3), 339–348.

Henrich, J., & Boyd, R. (2016). How evolved psychological mechanisms empower cultural group selection. *Behavioral and Brain Sciences*, *39*.

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual review of psychology*, *72*(1), 207–240.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Jara-Ettinger, J., Kim, N., Muetener, P., & Schulz, L. (2014). Running to do evil: Costs incurred by perpetrators affect moral judgment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Jo, H., & Shin, D. (2025). The impact of recognition, fairness, and leadership on employee outcomes: A large-scale multi-group analysis. *PloS one*, *20*(1), e0312951.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., ... others (2024). Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186. Retrieved from http://dx.doi.org/10.1080/1047840x.2014.877340 doi: 10.1080/1047840x.2014.877340

Miller, C. E., & Komorita, S. S. (1995). Reward allocation in task-performing groups. *Journal of Personality and Social Psychology*, *69*(1), 80.

Nagel, J., & Waldman, M. (2012). Force dynamics as a basis for moral intuitions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society open science*, *3*(11), 160510.

Sanna, L. J., & Turley, K. J. (1996). Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin*, *22*(9), 906–919.

Schäfer, M., Haun, D. B., & Tomasello, M. (2023). Children's consideration of collaboration and merit when making sharing decisions in private. *Journal of Experimental Child Psychology*, *228*, 105609.

Schulze Buschoff, L. M., Akata, E., Bethge, M., & Schulz, E. (2025). Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 1–11.

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive science*, *12*(1), 49–100.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... others (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345.

Wang, Z. (2024). Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th sighan workshop on chinese language processing (sighan-10)* (pp. 143–151).

Weiner, B. (1993). On sin versus sickness: A theory of perceived responsibility and social motivation. *American psychologist*, *48*(9), 957.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology: General*, *136*(1), 82–111.

Wu, S. A., & Gerstenberg, T. (2024). If not me, then who? responsibility and replacement. *Cognition*, *242*, 105646.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, 105609.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2025). People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, *116*, 104699.

Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.

Xiang, Y., Vélez, N., & Gershman, S. J. (2024). Optimizing competence in the service of collaboration. *Cognitive Psychology*, *150*, 101653.

Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.