Actual or counterfactual? Asymmetric responsibility attributions in language models

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

We investigate how language models assign responsibility to collaborators. We instruct 10 large language models from three different companies to assign responsibility to agents in a collaborative task. We then compare the language models' responses to seven existing cognitive models of responsibility attribution. We find that, while humans use actual and counterfactual effort to assign responsibility to collaborators, LLMs primarily use force, and this divergence shows up asymmetrically, when evaluating collaboration failures rather than successes. Our results highlight the similarities and differences between LLMs and humans in responsibility attributions and demonstrate the promise of interpreting LLM behavior using cognitive theories.

1 Introduction

As large language models (LLMs) become increasingly involved in collaborations with humans in day-to-day work [1–4], it is important to understand how the these systems reason about collaborations. Prior work evaluating social reasoning in LLMs has primarily focused on theory of mind abilities using experiments such as false belief tasks, where two agents have different beliefs about the world [5, 6]. [7] argue that such evaluations may measure the behavioral abilities of LLMs, but without describing the computations underlying those abilities. And while theory of mind research typically focuses on understanding an individual's belief states, much of humans' complex social reasoning involves people working in teams, where success depends not only an agent's individual contribution, but also on other people's contributions. Here, we evaluate the algorithms underlying LLMs' behavior on this key aspect of social reasoning—responsibility attribution in teams—by leveraging experimental paradigms, empirical data, and cognitive models adopted from previous studies on human social cognition. Our approach opens up new avenues for evaluating social reasoning in LLMs by examining responsibility attributions in collaboration, and in particular, for understanding the algorithms driving these behaviors.

We adapted materials from recent work on human responsibility judgment [8], instructing LLMs to attribute responsibility to agents in a collaborative task (Fig. 1A). We compared LLM responses to human responses and seven cognitive models. To test the generality of our findings, and whether LLM behaviors change as a function of model scale, we examined 10 LLMs, from three different companies and with varying numbers of parameters. We found that, while humans use actual and counterfactual effort to assign responsibility to collaborators, LLMs primarily use force, and this divergence particularly shows up when evaluating failed collaborations. With increasing model scale, the LLMs' behavior becomes increasingly correlated with humans', but the cognitive model that

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

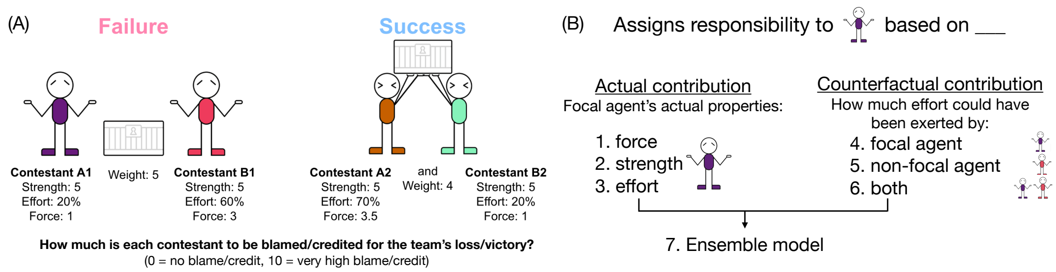

Figure 1: (A) Human experiment stimuli, adapted from [8]. Experiment 2a and converted into a text-only prompt format for LLMs. (B) Summary of the seven cognitive models we use to evaluate LLMs, see Appendix A for further details.

best explains these behaviors is consistently different. Our results highlight both similarities and differences between LLMs and humans in responsibility attributions, and demonstrate the promise of utilizing theories and models from human social cognition to interpret LLM behaviors.

2 Measuring responsibility attribution in collaborative contexts

Responsibility attributions in humans A large body of research in human social cognition has highlighted several factors that shape how people assign responsibility. The theories largely fall under two styles of reasoning [9]. One style of reasoning emphasizes a person's actual contributions to the outcome. For example, the amount of force a person exerts (how much output they actually contributed) [10–12], or their effort (how hard they tried) [13–16]. In general, those who contribute more force or effort are more responsible for the outcome they produce.

Another style of reasoning points to the role of counterfactual contributions—how much a person *could* have contributed—and whether acting differently would have changed the outcome [17]. On this view, the same actual contributions can yield different responsibility judgments depending on contextual factors such as task structure (e.g., whether success of a group requires everyone or just one teammate) [18], the temporal sequence of contributions (e.g., an action is more causally relevant when it happens at the right time) [19], and the availability of alternative options (e.g., whether someone can be easily replaced) [20].

These factors are not mutually exclusive. Recent computational work finds that responsibility attributions in collaborative contexts are best explained by a dual-factor model that considers both how much effort people actually contributed and how much they could have contributed [8]. We build directly on this work by adapting its materials and modeling framework to evaluate whether LLMs exhibit similar patterns in responsibility judgments. Because this prior study explicitly modeled the contributions of force, actual effort, and counterfactual effort, it provides a comprehensive testbed for comparison. By applying the same paradigm to LLMs, we can ask whether these models exhibit human-like sensitivity to the factors that guide responsibility judgments in collaborative settings. Below, we describe the experimental paradigm and cognitive models borrowed from [8].

Experimental Paradigm In the experiments, participants viewed vignettes where pairs of agents attempted to lift a box together (Figure 1A). Participants observed each agent's strength, effort, and force, as well as the weight of the box. Strength is defined as the maximum force an agent is capable of exerting, if they exert an all-out effort. Effort indicates how hard they try, i.e., the proportion of strength applied to the task. Trying the best one could exerts 100% effort, whereas not trying at all exerts 0% effort. Force is a result of applying effort—an agent produces force equal to their strength multiplied by effort. The agents succeed when their combined force exceeds the box weight (i.e., combined force $\geq$ box weight). After seeing whether the agents succeeded, participants assigned credit (when the lift was successful) or blame (when it failed) to each agent.

Cognitive Models In the analyses below, we compare LLM responses on this task to seven cognitive models to examine if they are driven by the same factors that drive human responses (Figure 1B). The cognitive models include three *actual-contribution* models that assign responsibility based on the agent's actual property (actual force, strength, and effort), three *counterfactual-contribution models*

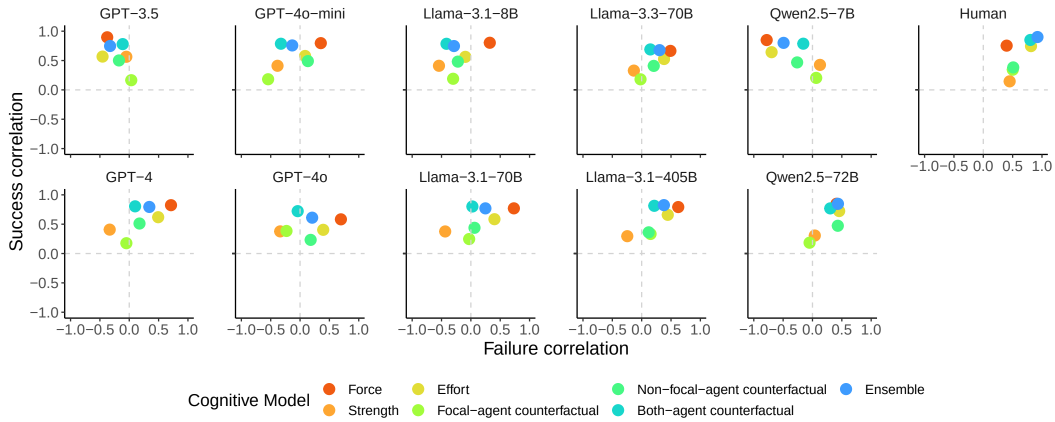

Figure 2: Comparing human and LLM responses to seven cognitive models. x-axis: Pearson correlation coefficients when the collaboration failed. y-axis: Pearson correlation coefficients when the collaboration succeeded. Dashed lines indicate the border between positive and negative correlations. Points 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.

that assign responsibility based on how much effort the agent and their partner could have exerted, and an *ensemble* model that combines the best actual-contribution model and the best counterfactual-contribution model, which has been shown to outperform the single-factor models in capturing human responsibility judgments [8]. See Appendix A for more details.

3 Experiments

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: "How much is each contestant to be blamed for the team's loss/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 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-2024-11-20`, and `gpt-4-0125-preview`, as well as six open-source LLMs, including four from Meta: `Llama-3.1-8B-Instruct`, `Llama-3.1-70B-Instruct`, `Llama-3.3-70B-Instruct`, `Llama-3.1-405B-Instruct`, and two from Alibaba Cloud: `Qwen2.5-7B-Instruct` and `Qwen2.5-72B-Instruct`. While OpenAI's 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 and TogetherAI APIs 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 $p(\text{'10'})$. 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.

4 Results

LLM responses are best explained by force Figure 2 shows the correlations between LLM responses and each of the seven cognitive models when the collaboration fails (x-axis) or succeeds (y-axis). Higher correlations indicate closer alignment in response patterns. A good model should be

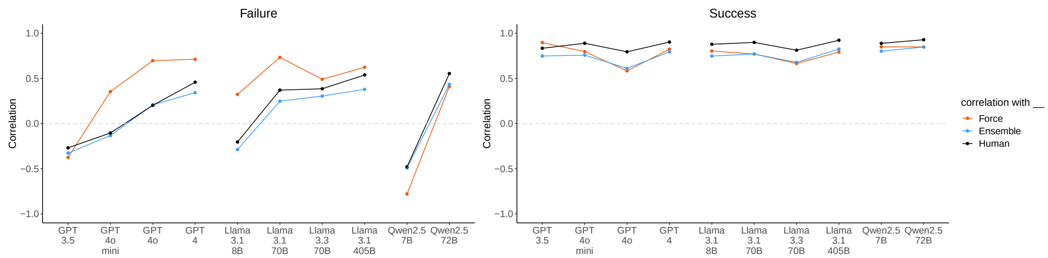

Figure 3: Correlations between LLM responses and the Force model, Ensemble model, and human judgments. LLMs are grouped by company and ordered by reported parameter count (e.g., 7B = 7 billion), which reflects model size and approximate computational power. The y-axis shows Pearson correlations between each model and the three benchmarks. Larger models tend to show stronger alignment with human and both the Force and Ensemble model predictions, although the Force model still dominates in most cases.

able to explain both failures and successes, thus, points that fall closer to the top right indicate better models for explaining the responses. The majority of the LLMs were best explained by the Force model, including three openAI models and all four Llama models we tested. GPT-3.5 and Qwen2.5-7B did not positively correlate with any cognitive models. Qwen2.5-72B was indistinguishably correlated with the Force model and the Ensemble model, and neither of the two models can explain failures. The correlation coefficients are reported in Appendix B. While LLM responses are primarily driven by force, human responses (Figure 2, top-right panel) are primarily driven by the ensemble model which considers actual and counterfactual effort.

More powerful LLMs are more correlated with humans, but still shows force bias Figure 3 shows correlations between LLMs and the Force model, Ensemble model, and human judgments, grouped by developing company and ordered by reported parameter count. Overall, there are more significant changes with evaluating failures, compared to evaluating successes. Within each model family, from left to right, as the number of parameters increase, all three correlations tend to increase for evaluating failed collaborations (left panel). This shows that increasing the number of parameters brings the LLM responses closer to humans. However, the Force model remains dominant in most cases, except for the two Qwen models, which are marginally better explained by the Ensemble model. By contrast, for success trials (right panel) correlations with human data and cognitive models are consistently high across LLMs from different companies and with different numbers of parameters.

5 Discussion

We compared LLMs' responsibility attributions to seven cognitive models and found that LLMs' responses were best captured by the Force model, which evaluates collaborators based on how much they actually contributed. By contrast, humans evaluated collaborators based on their actual and counterfactual effort [8]. We also discovered a progression trend: as the number of parameters increase, the LLM responses overall are more correlated with human judgments. The responses are increasingly correlated with both the Force model (which best describes LLM responses) and the Ensemble model (which best describes human responses), but the Force model remains dominant, indicating a persistent bias towards judging responsibility by force.

Success-failure asymmetry reveals differential counterfactual reasoning Interestingly, the divergence between human and LLM responses centers on interpreting failure. As shown in Figure 2 and highlighted in Figure 3, all LLMs—even including the earlier GPT-3.5 model or Llama and Qwen models with less than 10 billion parameters—were quite good at explaining what causes a team to succeed. The biggest change with increasing parameters seems to appear for evaluating what causes a team to fail. This may indicate an asymmetry in LLMs' ability to reason about counterfactuals for failures (i.e., whether exerting *more* effort could change the outcome to a success) versus counterfactuals for successes (i.e., whether exerting *less* effort could change the outcome to a failure). This pattern aligns with past work showing that LLMs learn more efficiently from better-than-expected outcomes than from worse-than-expected ones [21], suggesting a possible shared mechanism with our domain.

Taken together, these results contribute to our understanding of how LLMs diverge from humans in evaluating collaborators, and highlight the exciting opportunity for cognitive-theory-driven research in language models to shed light on aligning natural and artificial minds not only in responses, but also in reasoning, and ultimately, to improve collaboration between humans and machines.

References

[1] A Shaji George and AS Hovan George. A review of chatgpt ai's impact on several business sectors. *Partners universal international innovation journal*, 1(1):9–23, 2023.

[2] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[3] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[4] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. Self-collaboration code generation via chatgpt. *ACM Transactions on Software Engineering and Methodology*, 33(7):1–38, 2024.

[5] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

[6] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024.

[7] Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, 380(1932):20230499, 2025.

[8] Yang Xiang, Jenna Landy, Fiery A Cushman, Natalia Vélez, and Samuel J Gershman. Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, 241:105609, 2023.

[9] Ned Hall. Two concepts of causation. *Collins, Hall, and Paul*, 2004.

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

[11] Joshua D Greene, Fiery A Cushman, Lisa E Stewart, Kelly Lowenberg, Leigh E Nystrom, and Jonathan D Cohen. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3):364–371, 2009.

[12] Jonas Nagel and Michael Waldman. Force dynamics as a basis for moral intuitions. In *Proceedings of the annual meeting of the cognitive science society*, volume 34, 2012.

[13] Yochanan E Bigman and Maya Tamir. The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of experimental psychology: general*, 145(12):1654, 2016.

[14] Julian Jara-Ettinger, Nathaniel Kim, Paul Muetener, and Laura Schulz. Running to do evil: Costs incurred by perpetrators affect moral judgment. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

[15] Felix A Sosa, Tomer Ullman, Joshua B Tenenbaum, Samuel J Gershman, and Tobias Gerstenberg. Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, 217:104890, 2021.

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

[17] Tobias Gerstenberg. Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, 2024.

[18] Tobias Gerstenberg and David A Lagnado. Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, 115(1):166–171, 2010.

[19] Tobias Gerstenberg and David A Lagnado. When contributions make a difference: Explaining order effects in responsibility attribution. *Psychonomic Bulletin & Review*, 19:729–736, 2012.

[20] Sarah A Wu and Tobias Gerstenberg. If not me, then who? responsibility and replacement. *Cognition*, 242:105646, 2024.

[21] Johannes A Schubert, Akshay K Jagadish, Marcel Binz, and Eric Schulz. In-context learning agents are asymmetric belief updaters. *arXiv preprint arXiv:2402.03969*, 2024.

[22] Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017.

[23] Lawrence J Sanna and Kandi Jo Turley. Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin*, 22(9):906–919, 1996.

[24] Tobias Gerstenberg, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128(6):936–975, 2021.

[25] Yang Xiang, Jenna Landy, Fiery A Cushman, Natalia Vélez, and Samuel J Gershman. People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, 116:104699, 2025.

A Cognitive Models

The cognitive models assign responsibility (blame B in the event of failure, and credit C in the event of success) 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 [8].

In the experiments, each box has a weight $W \in [1,10]$, and each agent a has a strength $S_a \in [1,10]$ defined as the maximum amount of force that they could exert. Each agent exerts a level of effort $E_a \in [0,1]$, defined as a fraction of their strength, and produces force $F_a \in [0,S_a]$, defined as their strength times their effort ($F_a = E_a S_a$). The agents succeed when their combined force exceeds the box weight ($\sum_a F_a \geq W$), and fail otherwise ($\sum_a F_a < W$).

A.1 Actual-contribution models

Force model (F). The Force model allocates responsibility based on how much force an agent produces in the event. Agents who exert more force are blamed less and credited more.

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

Strength model (S). The Strength model allocates responsibility based on an agent's strength. Stronger agents receive more credit for successes, and receive more blame for failures.

$$\begin{aligned} B_a^S &\propto S_a \\ C_a^S &\propto S_a \end{aligned} \tag{2}$$

Effort model (E). The Effort model allocates responsibility based on the level of effort an agent exerts. Agents who exert more effort are credited more, and blamed less.

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

A.2 Counterfactual-contribution models

Central to the counterfactual-contribution models is the concept of *difference making* [22]: whether the outcome could have been different if the agents had exerted a different level of effort E'. Inspired by prior work [23], 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.[1] Specifically, we consider counterfactual efforts drawn 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 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:

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

[1] Past work has proposed other ways of constructing counterfactuals; for example, [24] proposed a noisy model 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.

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 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.

$$\begin{aligned} B_a^{FA} &\propto P_a \\ C_a^{FA} &\propto P_a \end{aligned} \tag{5}$$

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

$$\begin{aligned} B_a^{NFA} &\propto 1 - P_{/a} \\ C_a^{NFA} &\propto 1 - P_{/a} \end{aligned} \tag{6}$$

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 [8, 25], we assign equal weighting to the two components for simplicity.

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

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).

A.3 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 [8]. Here, we stick with the same equal-weighting Ensemble model to be consistent with past work and avoid adding free parameters to the model.

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

B Correlations between LLMs and cognitive models

We report the correlations between each LLM and the seven cognitive models, visualized in Figure 2.

Table 1: Correlations between GPT-family LLMs and cognitive models.

LLM	Cognitive Model	Failure Correlation	Success Correlation
GPT-3.5	Force	-0.38	0.90
	Strength	-0.05	0.56
	Effort	-0.45	0.57
	Focal-agent counterfactual	0.03	0.16
	Non-focal-agent counterfactual	-0.18	0.50
	Both-agent counterfactual	-0.11	0.78
	Ensemble	-0.33	0.75
GPT-4o-mini	Force	0.35	0.80
	Strength	-0.38	0.41
	Effort	0.09	0.58
	Focal-agent counterfactual	-0.54	0.18
	Non-focal-agent counterfactual	0.13	0.49
	Both-agent counterfactual	-0.33	0.79
	Ensemble	-0.13	0.76
GPT-4o	Force	0.70	0.58
	Strength	-0.34	0.38
	Effort	0.39	0.40
	Focal-agent counterfactual	-0.23	0.38
	Non-focal-agent counterfactual	0.18	0.23
	Both-agent counterfactual	-0.04	0.72
	Ensemble	0.21	0.61
GPT-4	Force	0.71	0.82
	Strength	-0.33	0.41
	Effort	0.49	0.62
	Focal-agent counterfactual	-0.05	0.18
	Non-focal-agent counterfactual	0.17	0.51
	Both-agent counterfactual	0.10	0.80
	Ensemble	0.34	0.79

Table 2: Correlations between Llama-family LLMs and cognitive models.

LLM	Cognitive Model	Failure Correlation	Success Correlation
Llama-3.1-8B	Force	0.32	0.80
	Strength	-0.54	0.41
	Effort	-0.09	0.56
	Focal-agent counterfactual	-0.30	0.19
	Non-focal-agent counterfactual	-0.22	0.48
	Both-agent counterfactual	-0.42	0.79
	Ensemble	-0.29	0.75
Llama-3.1-70B	Force	0.73	0.77
	Strength	-0.44	0.38
	Effort	0.40	0.59
	Focal-agent counterfactual	-0.03	0.25
	Non-focal-agent counterfactual	0.06	0.44
	Both-agent counterfactual	0.03	0.80
	Ensemble	0.25	0.77
Llama-3.3-70B	Force	0.49	0.66
	Strength	-0.13	0.33
	Effort	0.38	0.53
	Focal-agent counterfactual	-0.02	0.18
	Non-focal-agent counterfactual	0.21	0.41
	Both-agent counterfactual	0.15	0.69
	Ensemble	0.30	0.68
Llama-3.1-405B	Force	0.62	0.79
	Strength	-0.25	0.30
	Effort	0.44	0.66
	Focal-agent counterfactual	0.15	0.33
	Non-focal-agent counterfactual	0.12	0.36
	Both-agent counterfactual	0.21	0.81
	Ensemble	0.38	0.82

Table 3: Correlations between Qwen-family LLMs and cognitive models.

LLM	Cognitive Model	Failure Correlation	Success Correlation
Qwen2.5-7B	Force	-0.78	0.85
	Strength	0.13	0.43
	Effort	-0.70	0.64
	Focal-agent counterfactual	0.06	0.20
	Non-focal-agent counterfactual	-0.26	0.47
	Both-agent counterfactual	-0.16	0.79
	Ensemble	-0.49	0.80
Qwen2.5-72B	Force	0.41	0.85
	Strength	0.04	0.31
	Effort	0.45	0.72
	Focal-agent counterfactual	-0.05	0.18
	Non-focal-agent counterfactual	0.43	0.47
	Both-agent counterfactual	0.30	0.77
	Ensemble	0.43	0.85