MotiveBench: How Far Are We From Human-Like Motivational Reasoning in Large Language Models?

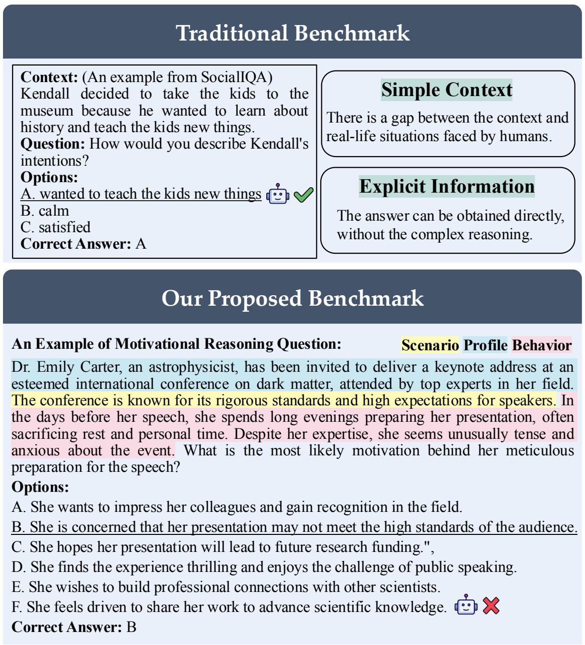
Anonymous ACL submission

Abstract

Large language models (LLMs) have been widely adopted as the core of agent frameworks in various scenarios, such as social simulations and AI companions. However, the extent to which they can replicate human-like motivations remains an underexplored question. Existing benchmarks are constrained by simplistic scenarios and the absence of character identities, resulting in an information asymmetry with real-world situations. To address this gap, we propose MotiveBench, which consists of 200 rich contextual scenarios and 600 reasoning tasks covering multiple levels of motivation. Using MotiveBench, we conduct extensive experiments on seven popular model families, comparing different scales and versions within each family. The results show that even the most advanced LLMs still fall short in achieving human-like motivational reasoning. Our analysis reveals key findings, including the difficulty LLMs face in reasoning about *"love & belonging"* motivations and their tendency toward excessive rationality and idealism. These insights highlight a promising direction for future research on the humanization of LLMs.

1 Introduction

Motivation is commonly conceptualized as an internal drive or psychological force that influences individuals to initiate and sustain goal-oriented activities (Hagger and Chatzisarantis, 2005; Brehm, 2014). It serves as a key explanatory factor for understanding why people initiate, continue, or terminate specific behaviors at any given scenarios (Kazdin et al., 2000). Motivation can be intrinsic, driven by internal values or preferences, or extrinsic, shaped by external rewards or punishments (Ryan and Deci, 2000; Radel et al., 2016).



Context: (An example from SocialIQA)
Kendall decided to take the kids to the museum because he wanted to learn about history and teach the kids new things.
Question: How would you describe Kendall's intentions?
Options:
A. wanted to teach the kids new things ✓
B. calm
C. satisfied
Correct Answer: A

There is a gap between the context and real-life situations faced by humans.

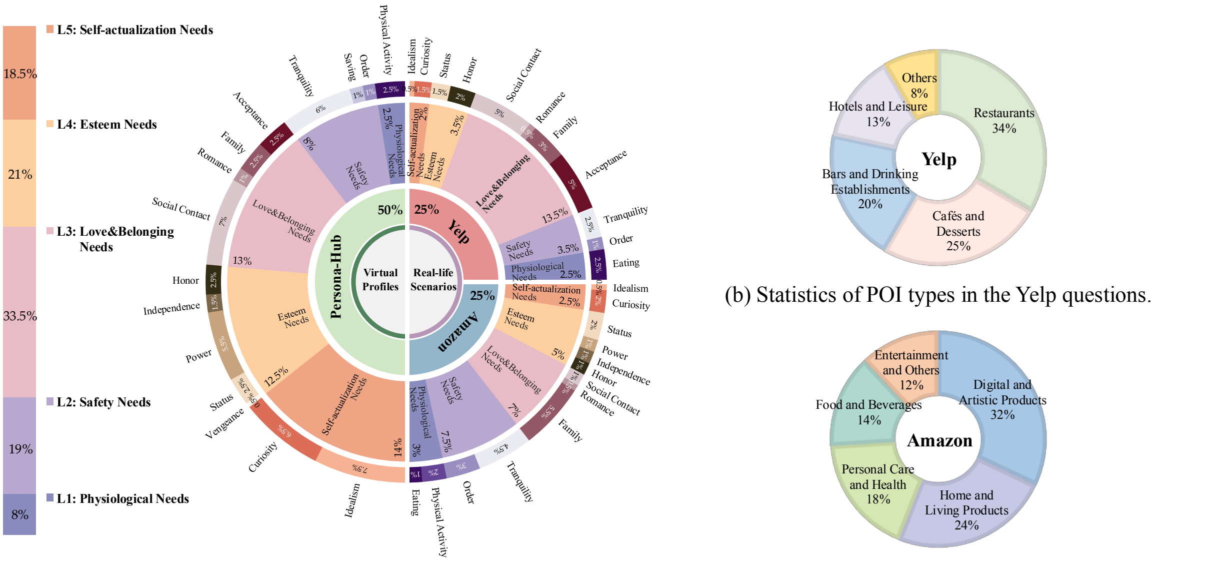
The answer can be obtained directly, without the complex reasoning.

An Example of Motivational Reasoning Question:

Dr. Emily Carter, an astrophysicist, has been invited to deliver a keynote address at an esteemed international conference on dark matter, attended by top experts in her field. The conference is known for its rigorous standards and high expectations for speakers. In the days before her speech, she spends long evenings preparing her presentation, often sacrificing rest and personal time. Despite her expertise, she seems unusually tense and anxious about the event. What is the most likely motivation behind her meticulous preparation for the speech?
Options:
A. She wants to impress her colleagues and gain recognition in the field.
B. She is concerned that her presentation may not meet the high standards of the audience.
C. She hopes her presentation will lead to future research funding.",
D. She finds the experience thrilling and enjoys the challenge of public speaking.
E. She wishes to build professional connections with other scientists.
F. She feels driven to share her work to advance scientific knowledge. ✗
Correct Answer: B

Figure 1: The difference between the existing traditional benchmark, such as SocialIQa from Sap et al. (2019) and our proposed MotiveBench.

Mimicking human behavior in specific scenarios has been a crucial task for autonomous agents, forming the foundation for various applications such as problem-solving, testing, and simulation (Schatzmann et al., 2007). Previous studies employ either rule-based (Keizer et al., 2010; Wilkins, 2014) or machine learning approaches (Asri et al., 2016; Kreyssig et al., 2018) to replicate human interactions in isolated and controlled environments. With the advent of large language models (LLMs) like GPT-4 (Achiam et al., 2023) and LLaMA (Dubey et al., 2024), an increasing number of studies have been adopting LLM-based agents (Aher et al., 2023; Argyle et al., 2023; Boiko et al., 2023; Kang and Kim, 2023; Mehta et al., 2023; Hong et al., 2023; Wang et al., 2024c) due to the remarkable capabilities of LLMs in general problem-solving, reasoning, and autonomous action-taking. However, a critical question remains underexplored: **Can current LLMs truly understand and exhibit human-like motivations and behaviors?** The complexity of human behavior dynamics poses new challenges for LLMs, which are



(a) Data sources and statistics of different need levels in MotiveBench.

(b) Statistics of POI types in the Yelp questions.

(c) Statistics of product categories in the Amazon questions.

Figure 2: The statistical overview of MotiveBench. It contains 200 diverse profiles and real-world scenarios, along with 600 motivational and behavioral reasoning questions, covering multiple finely-grained levels of needs.

distinct from the challenges in understanding and manipulating the physical world. Delving deeply into this research question can lay a reliable foundation for LLM-based social simulations.

Recent literature has attempted to explore the gap between LLMs and real humans from a narrow behavioral perspective. Xie et al. (2024) demonstrated the feasibility of simulating human trust behavior with LLM agents in the Trust Game (Berg et al., 1995), while Zhou et al. (2023) analyzed differences between LLMs and humans in social interactions (Lee et al., 2024). However, these studies focus on specific subdomains and fail to provide broader insights into human-like behavior. In the study of motivations and behaviors in LLMs, Sap et al. (2019) introduced SocialIQa, a benchmark for commonsense reasoning in social contexts. While it includes some basic reasoning tasks on intentions behind behaviors, it lacks detailed, comprehensive, and challenging assessments, as shown in Figure 1. Existing similar benchmarks (Rashkin et al., 2018a,b; Talmor et al., 2018) also exhibit several clear limitations: **(1) Simplistic contexts**, lacking detailed scenarios and character profiles, leading to information asymmetry compared to real-world situations; **(2) Overly explicit information**, with tasks solvable through basic pattern matching without requiring human-like reasoning; and **(3) Limited theoretical grounding**, failing to systematically capture the multi-level nature of human motivation.

To address the above challenges, we propose MotiveBench, a comprehensive evaluation benchmark, consisting of 200 diverse profiles and scenarios, along with 600 motivational and behavioral reasoning questions. Figure 2 shows an overview of MotiveBench. We strive to cover and balance the proportions of different levels of needs in the benchmark, ultimately formulating the questions. Specifically, it has the following advantages: **(1) Diverse scenarios, profiles, motivations, and behaviors.** We utilize diverse profiles from the Persona-Hub (Ge et al., 2024) dataset, along with real-world motivation and behavior data from platforms like Amazon and Yelp, as the basis for question generation. **(2) Human-in-the-loop multi-agent framework to enhance efficiency and quality.** We propose a multi-agent collaboration framework that efficiently generates high-quality questions across a range of difficulties, requiring minimal human effort to ensure validity. **(3) Grounded in authoritative psychological theories to ensure comprehensive evaluation.** Our test questions cover the five levels of Maslow's Hierarchy of Needs (Maslow, 1943), as well as the 16 basic desires of human nature from the Reiss Motivation Profile (Reiss, 2004).

To the best of our knowledge, MotiveBench is the first benchmark for evaluating LLMs' human-like motivation-behavior reasoning. **Our results show that even the most advanced LLMs fall short in achieving human-like motivation and behavior reasoning.** Beyond the quantified results, comprehensive experiments reveal notable insights, such as significant differences in motivational reasoning between LLMs and humans, and their limitations in data annotation for human social behaviors. We hope our research provides practical guidelines for applying LLMs in social simulations and contributes to their future humanization. The dataset, benchmark, and code will be released upon paper acceptance to support further research.

2 MOTIVEBENCH Preliminaries

2.1 Three Types of Reasoning Tasks

Generally, in a specific scenario, an individual with a certain profile will perform a behavior based on a particular motivation. We define this as a complete behavioral quadruple: **Scenario**, **Profile**, **Motivation**, and **Behavior**. The scenario provides the context and external triggers (Yang et al., 2009). Profile shapes an individual's understanding of the behavior and the way they act (Bandura, 2001; Eagly and Wood, 2012). Motivation is the internal driving force behind an individual's actions, based on their needs, goals, or emotional state (Deci and Ryan, 2012; Shayganfar et al., 2016). Behavior is the result of the interaction between scenario, profile, and motivation (Graham, 1991). Therefore, we define three types of reasoning tasks:

1) **Motivational Reasoning Question.** Given a specific scenario, profile, and detailed behavior information, the task is to infer the motivation behind the individual's behavior.

2) **Behavioral Reasoning Question.** Given a specific scenario, profile, and detailed motivation information, the task is to infer the most likely behavior the individual would perform.

3) **Motive&Behavior Reasoning Question.** This more challenging task closely aligns with the ultimate goal of autonomous agents, which is to infer the most reasonable motivation and behavior when only the scenario and profile are provided.

2.2 Fine-Grained Needs Hierarchy

Maslow's hierarchy of needs theory (Maslow, 1943) suggests that human actions are driven by various needs, which are divided into five levels: Physiological Needs, Safety Needs, Love and Belonging Needs, Esteem Needs, and Self-actualization Needs. When lower-level needs are met, individuals will seek to fulfill higher-level needs (Jr, 1991).

Furthermore, Reiss (2004) proposes 16 more granular categories to provide a broader and more informative range of motivations. These include Curiosity, Idealism, Honor, Independence, Power, Status, Vengeance, Acceptance, Family, Romance, Social Contact, Order, Saving, Tranquility, Eating, and Physical Activity. Although the Reiss theory offers more detailed insights into motivation, the broader range of abstract concepts can be difficult to manage. Inspired by Rashkin et al. (2018a), we adopt a hybrid method in which the Reiss Motive Profile labels are categorized as sub-categories within Maslow's framework.

3 MOTIVEBENCH Construction

We construct MOTIVEBENCH from scratch to avoid relying on existing scales or test items from psychology or sociology, thereby mitigating potential data leakage or contamination issues. Figure 3 illustrates the AI-Human collaboration framework.

3.1 Data Collection and Pre-processing

To obtain diverse scenarios and profiles, we collect data from Persona-Hub proposed by Ge et al. (2024), as well as real-life platforms like Amazon (Hou et al., 2024) and Yelp[1].

Persona-Hub contains diverse profiles, such as *"A fellow astrophysicist who specializes in the study of dark matter and provides valuable insights and critiques to the author's research."* Based on them, we can synthesize a diverse range of scenarios, motivations, and behaviors using the fine-grained hierarchy of needs outlined in Section 2.2.

In addition, review texts on platforms like Amazon and Yelp offer rich real user intent data, such as reasons for visiting a specific point of interest (POI) or purchasing a product. This provides a valuable reference for collecting data grounded in real-world scenarios. Therefore, we first collect review data from 33 Amazon product domains and 22 Yelp business categories to ensure diversity. We then employ LLMs, such as LLaMA3.1-70B and Qwen2.5-72B, to filter high-quality reviews that align with Maslow's hierarchy of needs, focusing on motivations rooted in real-life contexts. For example, in reviews of smartphones, we prioritize motivations such as purchasing for better communication with family, reflecting "social needs", or selecting a phone with a long battery life for convenience during travel, reflecting "safety needs", rather than only focusing on product attributes (such as performance or appearance) as the motivation for purchase. An example is shown in Figure 4.

[1] https://www.yelp.com/dataset

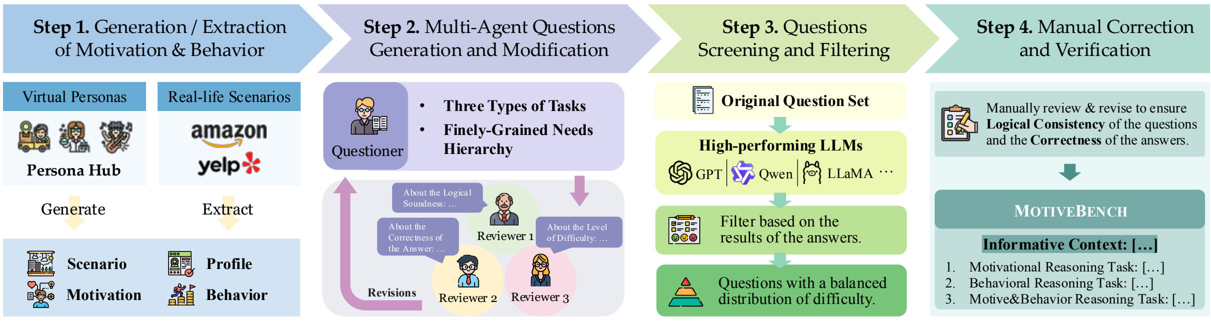

Step 1. Generation / Extraction of Motivation & Behavior
Step 2. Multi-Agent Questions Generation and Modification
Step 3. Questions Screening and Filtering
Step 4. Manual Correction and Verification
Virtual Personas
Real-life Scenarios
Persona Hub
amazon
yelp
Generate
Extract
Scenario
Profile
Motivation
Behavior
Questioner
• Three Types of Tasks
• Finely-Grained Needs Hierarchy
About the Logical Soundness: ...
About the Correctness of the Answer: ...
About the Level of Difficulty: ...
Reviewer 1
Reviewer 2
Reviewer 3
Revisions
Original Question Set
High-performing LLMs
GPT
Qwen
LLaMA ···
Filter based on the results of the answers.
Questions with a balanced distribution of difficulty.
Manually review & revise to ensure Logical Consistency of the questions and the Correctness of the answers.
MotiveBench
Informative Context: [...]
1. Motivational Reasoning Task: [...]
2. Behavioral Reasoning Task: [...]
3. Motive&Behavior Reasoning Task: [...]

Figure 3: A step-by-step questions generation and correction pipeline using AI-Human collaboration framework.

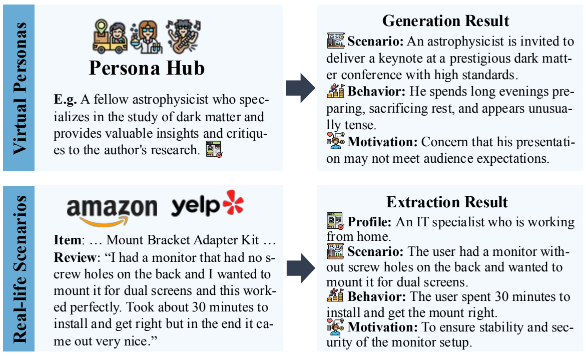

Virtual Personas
Persona Hub
E.g. A fellow astrophysicist who specializes in the study of dark matter and provides valuable insights and critiques to the author's research.
Generation Result
Scenario: An astrophysicist is invited to deliver a keynote at a prestigious dark matter conference with high standards.
Behavior: He spends long evenings preparing, sacrificing rest, and appears unusually tense.
Motivation: Concern that his presentation may not meet audience expectations.
Real-life Scenarios
amazon
yelp
Item: ... Mount Bracket Adapter Kit ...
Review: "I had a monitor that had no screw holes on the back and I wanted to mount it for dual screens and this worked perfectly. Took about 30 minutes to install and get right but in the end it came out very nice."
Extraction Result
Profile: An IT specialist who is working from home.
Scenario: The user had a monitor without screw holes on the back and wanted to mount it for dual screens.
Behavior: The user spent 30 minutes to install and get the mount right.
Motivation: To ensure stability and security of the monitor setup.

Figure 4: Generation and extraction of motivations and behaviors from virtual personas and real-life scenarios.

3.2 Questions Generation and Modification

Existing benchmarks mostly rely on manual construction, which requires significant effort (Sabour et al., 2024; Chen et al., 2024). To reduce financial and labor costs, we employ a multi-agent framework to formulate questions and generate options.

Specifically, we begin by using an LLM-based questioner to formulate complete question content based on the quadruples obtained in the previous stage. Next, three LLM-based reviewers provide feedback from: 1) the logical soundness of the question, 2) the correctness of the answer, and 3) the level of difficulty. The feedback is then compiled and sent back to the questioner for revisions. This process is iterated until no further suggestions are made, or the iteration threshold is reached. To minimize the hidden biases introduced by a single LLM, we use a diverse set of models, such as LLaMA3.1-70B and Qwen2.5-72B, for question creation and modification. The detailed prompts of our agents can be found in the Appendix F.

3.3 Questions Screening and Filtering

A well-rounded benchmark should include questions of varying difficulty to evaluate LLMs. Since controlling difficulty during question generation is challenging, we use high-performing LLMs to answer and label each question. Based on this, we categorize the questions into difficult, medium, and easy levels, ensuring a balanced distribution of all three difficulty levels in the final benchmark. Ultimately, We curate a subset of high-quality, diverse scenarios, retaining 100 scenarios for Person-Hub, 50 for Amazon, and 50 for Yelp. These scenarios span different difficulty levels and provide balanced coverage. Notably, this scale of context-rich evaluation proves sufficient for assessing LLMs' performance, yielding statistically robust and methodologically sound conclusions (Sabour et al., 2024).

3.4 Manual Correction and Verification

After filtering the questions, we authors manually review and revise each question to ensure high quality. We carefully refine the logical consistency of questions and the consensus of the correct options, it taks about 6 minutes per question. Due to the hallucination issues (Huang et al., 2023) in LLMs that often introduce logical or factual errors, this step is essential. Our analysis reveals that some generated questions lack clear human consensus, highlighting LLMs' limitations in capturing human reasoning's complexity and nuance. Notably, we minimize the impact of demographic backgrounds on answers and demonstrate this by recruiting diverse annotators in follow-up experiments. The final statistics of MotiveBench are shown in Table 1, where the context length is significantly longer than that of previous benchmarks (e.g., SocialIQa in Sap et al. (2019): 20.16 tokens per context, and CommonsenseQA in Talmor et al. (2018): 13.41 tokens per context).

Finally, we recruit 28 participants and select 15 (diverse backgrounds) to assess MCQ correctness and reasonableness while preventing answer leakage. Each question is independently annotated by five annotators, yielding 98.17% consistency (≥ 3) and 93.00% agreement with predefined answers. Detailed analyses are in the Appendix C. We cor-

	#S	#Q	Average Token Count		
			Context	Cor.Opt	Fal.Opt
MotiveBench	**200**	**600**	**96.45**	**13.83**	**13.31**
Persona	100	300	97.13	15.53	14.60
Amazon	50	150	87.19	12.57	12.51
Yelp	50	150	104.35	11.65	11.51

Table 1: Data statistics of our proposed MotiveBench. (#S: number of scenarios, #Q: number of questions, Cor.Opt: correct options, and Fal.Opt: false options.)

rect inconsistent questions and those that contradict predefined answers based on the annotations, ensuring that the final answer of each question aligns accurately with human consensus.

4 Experiments

4.1 Experimental Setup

For each scenario, there are three types of questions presented in the form of MCQ. The scenario is only considered correct when all three questions within the same scenario are answered correctly. The specific format is shown in Appendix B.

We evaluate LLMs in two settings: one using vanilla prompting with task instructions (Base), and the other employing chain-of-thought reasoning (CoT) (Wei et al., 2022). The prompts we used are detailed in Appendix F. Given that LLMs have been shown to exhibit a bias towards the order of choices (Zheng et al., 2023), we introduce random variations in the choice order by generating 6 permutations. This ensures the correct option appears in all possible positions, while the incorrect options are randomly shuffled each time. We report the average of the 6 results as the final performance.

We evaluate 29 popular LLMs, as listed in Appendix D. For all open-source models, we use the vLLM[2] and set the temperature parameter to 0 to ensure result stability. For closed-source models, we access them through the Azure OpenAI API[3].

4.2 Main Results

Table 2 summarizes the performance of various LLMs across three domains, with detailed task-specific results in Appendix E. Below, we analyze the results and highlight several key findings.

First of all, since our MotiveBench inherently reflect human consensus, the accuracy of each model serves as an indicator of its capacity for human-like motivation-behavior reasoning. It is evident that even the most advanced model, GPT-4o, exhibits a performance gap compared to human-level reasoning, with an evaluation result of 81.08%. Furthermore, analysis of questions where GPT-4o failed reveals persistent differences in reasoning patterns with human. These discrepancies will be examined in detail in Section 4.3 Insight 2.

Secondly, among all the LLMs we evaluated, GPT-4o demonstrates the strongest capability in MotiveBench. Notably, within the open-source model series, the Qwen2.5 series demonstrates strong performance, with smaller models (14B, 32B) achieving capabilities comparable to the 72B model and even GPT-4o. Similarly, LLaMA 3.1-70B also shows good results. However, other series, especially Baichuan2, exhibit weaker reasoning capabilities for motivation and behavior tasks. From the perspective of model size, small-scale models (<10B) achieve an average accuracy of 53.10%, medium-scale models (10B-34B) 65.17%, and large-scale models (>34B) 73.63%. These findings suggest a clear trend of improved motivational and behavioral reasoning ability as model size increases. The pattern is better visualized in Figure 5.

Thirdly, CoT does not enhance the performance of LLMs in MotiveBench. In fact, results from most models indicate that CoT may lead to a decrease in performance. This effect is particularly pronounced in models with smaller parameter sizes (<=34B), where accuracy drops by 6.35%, compared to a 1.88% decrease in larger models (>=70B). This decline may occur because CoT simplifies tasks, but when the model's reasoning diverges from human cognitive patterns on motivations and behaviors, it hampers alignment with human cognition, reducing performance. In contrast, upgrading models and increasing their size significantly improve human-like motivational reasoning, as illustrated in Figure 5.

Another interesting finding is that, LLMs perform poorly in understanding *"love & belonging"* needs, which are related to emotional aspects. In Table 3 we break the overall score into five motivation aspects, as introduced in Section 2.2. While some studies suggest that LLMs can provide emotional value comparable to or even surpassing that of humans—for instance, the Replika chatbot reduced suicidal ideation for 3 percent of users (Maples et al., 2024)—they still exhibit limitations in emotional understanding and reasoning (Sabour et al., 2024). This may be attributed to: (1) LLMs excel in providing emotional value

[2] https://docs.vllm.ai

[3] https://azure.microsoft.com

Motive&Behavior Reasoning Ability	**Vitural Profiles**		**Real-life Scenarios**				**Overall**	
	Persona-Hub		**Amazon**		**Yelp**			
LLMs	**Base**	**CoT**	**Base**	**CoT**	**Base**	**CoT**	**Base**	**CoT**
Baichuan2-7B-Chat	33.33	**30.33**	37.67	**35.67**	28.67	24.33	33.25	**30.17**
Baichuan2-13B-Chat	**42.33**	25.67	**49.33**	32.00	**47.33**	**30.33**	**45.33**	28.42
ChatGLM3-6B	39.83	31.33	44.00	28.00	40.67	32.33	41.08	30.75
GLM4-9B-Chat	**56.33**	**56.50**	**75.33**	**75.00**	**66.33**	**67.33**	**63.58**	**63.83**
Yi1.5-6B-Chat	36.67	45.33	47.67	55.33	45.00	53.00	41.50	49.75
Yi1.5-9B-Chat	55.67	51.83	68.33	63.33	62.67	58.33	60.58	56.33
Yi1.5-34b-Chat	**60.33**	**58.00**	**72.33**	**72.67**	**75.67**	**65.00**	**67.17**	**63.42**
Phi3-mini-4k-Instruct	57.33	44.83	67.67	48.33	57.67	47.00	60.00	46.25
Phi3-small-8k-Instruct	58.50	**63.17**	76.33	**81.00**	68.33	**76.33**	65.42	**70.92**
Phi3-medium-4k-Instruct	**66.67**	61.50	**80.00**	69.33	**76.33**	68.00	**72.42**	65.08
Phi3.5-mini-Instruct	57.17	55.67	70.33	64.67	62.67	59.00	61.83	58.75
Phi3.5-MoE-Instruct	**66.67**	53.00	72.33	69.33	73.00	74.33	69.67	62.42
Llama2-7B-Chat	17.50	21.83	21.33	24.00	14.67	27.00	17.75	23.67
Llama2-13B-Chat	45.00	33.67	53.00	37.67	45.33	42.33	47.08	36.83
Llama2-70B-Chat	48.83	57.50	60.67	61.00	53.00	61.33	52.83	59.33
Llama3.1-8B-Instruct	59.50	52.50	74.00	68.67	63.67	62.00	64.17	58.92
Llama3.1-70B-Instruct	**74.00**	**68.50**	**84.00**	**83.33**	**79.00**	**75.67**	**77.75**	**74.00**
Qwen-7B-Chat	42.33	39.17	54.00	51.00	47.33	43.00	46.50	43.08
Qwen-14B-Chat	59.17	51.67	69.33	63.33	68.33	59.00	64.00	56.42
Qwen-72B-Chat	65.17	63.50	78.00	79.33	69.67	65.00	69.50	67.83
Qwen2-7B-Instruct	63.00	64.50	73.00	74.33	72.00	71.67	67.75	68.75
Qwen2-72B-Instruct	71.17	**70.17**	83.67	79.33	79.33	65.00	76.33	71.17
Qwen2.5-7B-Instruct	64.83	62.67	74.33	70.67	63.67	62.33	66.92	64.58
Qwen2.5-14B-Instruct	71.67	67.33	80.00	77.67	79.00	73.33	75.58	71.42
Qwen2.5-32B-Instruct	74.00	68.83	**90.33**†	**87.00**†	**82.00**	76.67	**80.08**	**75.33**
Qwen2.5-72B-Instruct	**74.17**	68.50	87.33	83.33	77.67	**78.33**	78.33	74.67
GPT-3.5-Turbo 1106	57.83	58.83	79.67	78.67	63.00	64.33	64.58	65.17
GPT-4o mini 2024-07-18	72.33	69.50	84.67	**85.00**	77.00	75.33	76.58	74.83
GPT-4o 2024-05-13	**75.50**†	**72.00**†	**89.67**	84.67	**83.67**†	**79.33**†	**81.08**†	**77.00**†

Table 2: Evaluation Results for MOTIVEBENCH across 7 popular model families in 3 domains, including Base and CoT prompting. The best results in each series are highlighted in **Bold**, with the best overall results marked by †.

by mimicking surface-level language patterns, creating a sense of understanding without deep causal reasoning. (2) The expression of *"love & belonging"* needs in the text data is often implicit or ambiguous, as it primarily involves internal processes. Since the model lacks direct exposure to comprehensive and real-world social psychology case data, it struggles to handle such issues effectively.

4.3 In-Depth Analysis

Insight 1: Comparison with Existing Benchmarks. We aim to introduce a new evaluation dimension—Motive. To study the difference between MOTIVEBENCH and existing benchmarks, we leverage LiveBench (White et al., 2024), which typically assess general capabilities like coding, mathematics, reasoning, data analysis, language comprehension, and instruction following.

Figure 6 shows the Pearson correlation coefficients (Cohen et al., 2009) of rankings across different ability dimensions for several popular LLMs. It is evident that Motive is distinct from existing evaluation dimensions, with an average correlation coefficient of 0.8175. This suggests that by introducing the Motive dimension, we can explore patterns or relationships in human capabilities that traditional evaluation metrics do not observe.

Insight 2: Differences Between GPT-4o and Human in Motivation and Behavior. In Table 2, we observe that GPT-4o is the most advanced model in our benchmark. Therefore, we are curious to investigate the situations in which this model fails to demonstrate human-like intelligence. For questions answered incorrectly by GPT-4o, we examine its reasoning and thought processes. We have summarized the following findings:

1) Over-Rationalization, Lacking Emotional Insight. GPT-4o often relies on logical reasoning, neglecting broader practical experience or emotional context, leading to reasoning that may be disconnected from real-world complexities.

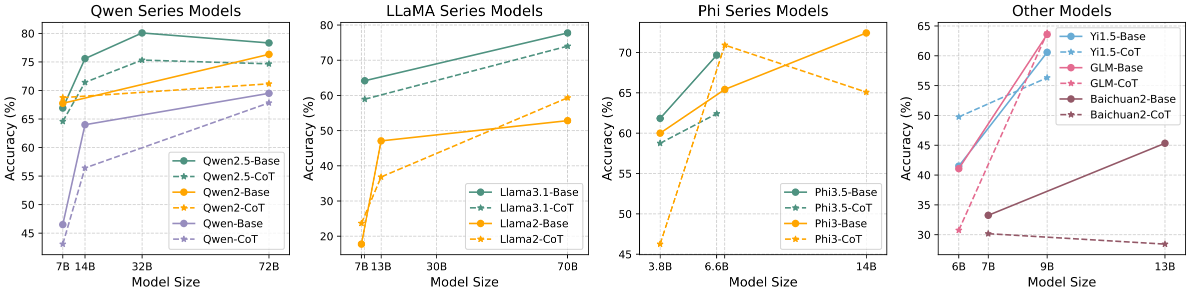


Figure 5: Scaling of various model families across their different versions and sizes in motivational reasoning ability.

Needs Hierarchy	GPT series		LLaMA series		Qwen series		Phi series		Yi series		AVG.
	4o	4o mini	3.1 70B	3.1 8B	2.5 72B	2.5 32B	3.5 MoE	3 Medium	1.5 34B	1.5 9B	
Level 1: Physiological	85.83	80.00	84.31	75.97	76.67	88.33	80.69	82.50	69.86	64.44	78.86
Level 2: Safety	83.20	79.69	82.49	64.31	81.50	86.45	80.58	74.34	68.32	68.72	76.96
Level 3: Love & Belonging	77.27	72.86	69.93	62.68	74.05	76.16	64.81	68.71	62.49	54.28	68.32
Level 4: Esteem	80.01	77.27	74.83	61.07	74.92	81.58	62.31	65.32	65.96	57.55	70.08
Level 5: Self-actualization	82.83	80.65	81.40	69.20	84.52	82.20	74.35	76.04	70.57	63.57	76.53

Table 3: Hierarchy of needs-oriented evaluation results for different model families and their strongest models.

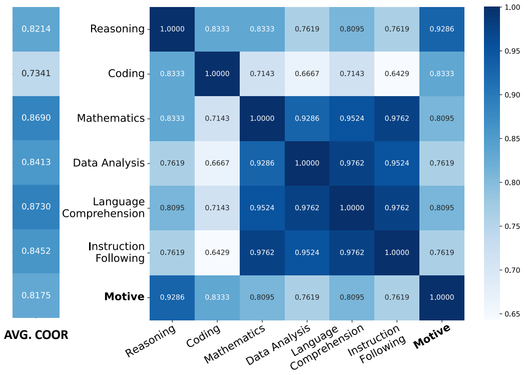


Figure 6: The correlation coefficients between the evaluation results of six general capabilities and the Motive.

2)Weak Logical Precision, Prone to General Assumptions. GPT-4o's reasoning can be overly simplistic, often based on general assumptions or external knowledge, without fully addressing the specific context or details of a situation.

3)Overly Idealistic, Ignoring Complex Realities. It tends to assume people follow social norms or moral codes, ignoring more complex or real-world challenges that could affect behavior.

4)Lack of Awareness of Behavioral Impact. It may prioritize actions that seem easy or plausible but fail to consider their actual effectiveness or long-term impact in real-life scenarios.

Figure 7 presents an example of GPT-4o demonstrating weak logical coherence and overly idealistic. Detailed discussion and interpretation, and more cases can be found in the Appendix G.

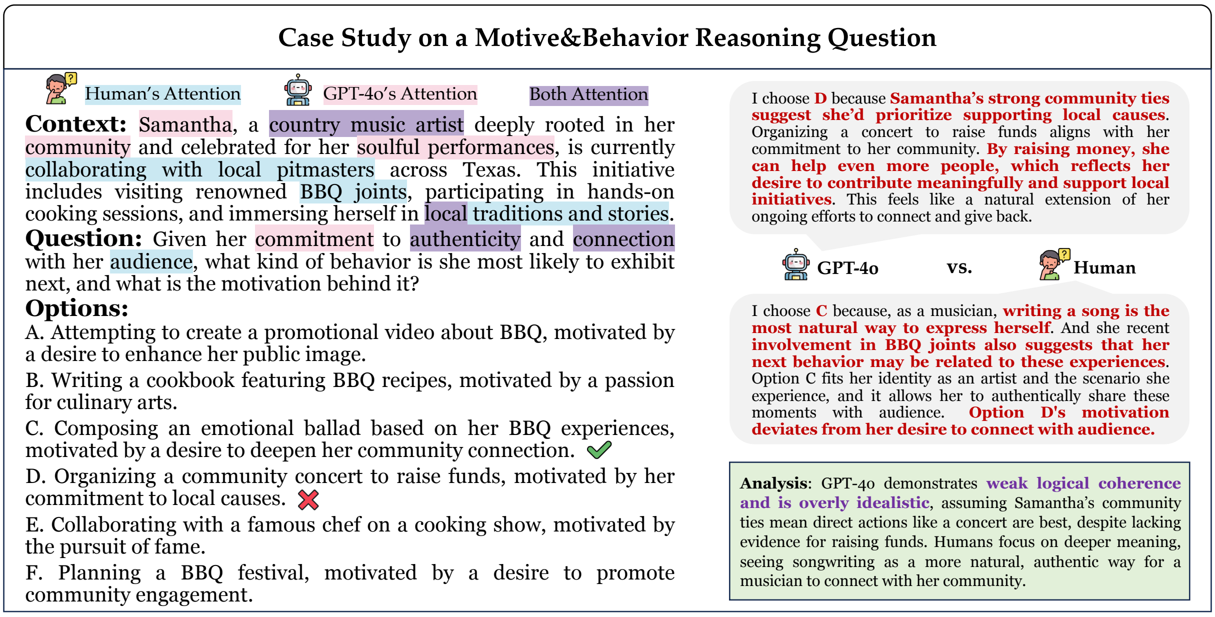
Insight 3: Limitations of LLMs in Complete Data Annotation. Recently, many researchers have been curious about whether "large language models replace human annotators" or "replace human participants in social science experiments". By using our MOTIVEBENCH as the lens, we find that LLMs for annotation presents several challenges:

1) Logical or Factual Errors. LLMs may generate inaccurate or misleading questions due to limited understanding of psychological theories, resulting in the content of the questions or options lacking logical consistency. Therefore, we introduce a manual correction step to ensure the reliability of the reasoning questions.

2) Limited Understanding of Human Dynamics. The tasks we consider often involve nuanced psychological and sociological dimensions, which LLMs may struggle to accurately capture due to the complexity of human thought processes.

3)Annotator-specific Bias. Relying on a single LLM introduces biases, distorting the objective interpretation of psychological data. To minimize such biases, our framework uses multiple models at different stages of the pipeline and incorporates human labeling to revise questions at the final step.

Insight 4: Trends of Motivational Intelligence in LLMs of Varying Sizes Our experiments with LLMs of varying sizes reveal distinct trends in motivational intelligence. Smaller models (6B-13B) struggle with complex reasoning, long-chain inferences, and emotional understanding in the informative contexts we provide, often producing superficial responses. Medium-sized models (13B-34B) show improved performance, are able to capture motivational cues, provided better contextual understanding and somewhat human-like thinking



Case Study on a Motive&Behavior Reasoning Question

Human's Attention GPT-4o's Attention Both Attention

Context: Samantha, a country music artist deeply rooted in her community and celebrated for her soulful performances, is currently collaborating with local pitmasters across Texas. This initiative includes visiting renowned BBQ joints, participating in hands-on cooking sessions, and immersing herself in local traditions and stories.

Question: Given her commitment to authenticity and connection with her audience, what kind of behavior is she most likely to exhibit next, and what is the motivation behind it?

Options:

A. Attempting to create a promotional video about BBQ, motivated by a desire to enhance her public image.

B. Writing a cookbook featuring BBQ recipes, motivated by a passion for culinary arts.

C. Composing an emotional ballad based on her BBQ experiences, motivated by a desire to deepen her community connection. ✓

D. Organizing a community concert to raise funds, motivated by her commitment to local causes. ✗

E. Collaborating with a famous chef on a cooking show, motivated by the pursuit of fame.

F. Planning a BBQ festival, motivated by a desire to promote community engagement.

I choose **D** because **Samantha's strong community ties suggest she'd prioritize supporting local causes**. Organizing a concert to raise funds aligns with her commitment to her community. **By raising money, she can help even more people, which reflects her desire to contribute meaningfully and support local initiatives**. This feels like a natural extension of her ongoing efforts to connect and give back.

GPT-4o vs. **Human**

I choose **C** because, as a musician, **writing a song is the most natural way to express herself**. And she recent **involvement in BBQ joints also suggests that her next behavior may be related to these experiences**. Option C fits her identity as an artist and the scenario she experience, and it allows her to authentically share these moments with audience. **Option D's motivation deviates from her desire to connect with audience.**

Analysis: GPT-4o demonstrates **weak logical coherence and is overly idealistic**, assuming Samantha's community ties mean direct actions like a concert are best, despite lacking evidence for raising funds. Humans focus on deeper meaning, seeing songwriting as a more natural, authentic way for a musician to connect with her community.

Figure 7: A case study on motivation and behavior reasoning: we analyze GPT-4o's attention within the question and its explicitly generated reasoning process to uncover key differences between its approach and human cognition.

patterns. In addition, large models (70B+), particularly GPT-4o and Qwen2.5-72B, excel in complex reasoning and nuanced motivational modeling. They integrate scenario and character's profile information effectively, understanding motivations, emotions, and logical relationships, thus showing behavior patterns closer to human intelligence.

However, current LLMs still differ from humans in some aspects of motivational reasoning (see Insight 2), posing challenges in fully open-ended, dynamic scenarios like real-time strategy games and unpredictable social interactions—key frontiers for agent simulation and decision-making systems.

5 Related Work

Currently, the intellectual capabilities of LLMs have reached an unexpected level, but they also display a certain unreliability—sometimes becoming confused, while at other times demonstrating abilities that far exceed human in specific evaluations. But how far are they from human-like cognition? An increasing number of studies are now approaching this issue from Theory of Mind (ToM) perspective. Sap et al. (2022) find that GPT-3 lacks social intelligence by using SOCIALIQA (Sap et al., 2019) and ToMi (Le et al., 2019). Ullman (2023) demonstrate that small variations that preserve the principles of ToM can drastically alter the outcomes. Similar results can also be found from Shapira et al. (2023). Strachan et al. (2024) observe that GPT-4 excelled at identifying indirect requests, false beliefs, and misdirection, but struggled with detecting faux pas, and it still lags behind humans in overall ToM performance (Chen et al., 2024). In addition, some studies have conducted experiments from a behavioral perspective. Xiao et al. (2023) reveal that current LLMs struggle to align their behaviors with assigned characters. Furthermore, Zhou et al. (2023) and Wang et al. (2024a) propose interactive and sandbox benchmarks, showing GPT-4 excels in conversational scenarios but struggles to exhibit social commonsense reasoning and deal with social tasks (Lee et al., 2024).

Different from existing research, we aim to explore a new dimension—Motive, which examines the alignment between LLMs and human behavior in dynamic and unstructured environments.

6 Conclusions

We introduce MOTIVEBENCH, the first systematic benchmark to evaluate the human-like motivational and behavioral reasoning ability of LLMs with detailed, realistic situations. Our results reveal that even the most advanced LLMs still fall short in achieving human-like motivational reasoning. Furthermore, most LLMs struggle with understanding *"love & belonging"* needs. In-depth analysis reveals specific differences in motivational reasoning between current LLMs and humans. By introducing MOTIVEBENCH, we aim to provide insights from a new dimension, enabling future models to exhibit more human-like cognition processes.

Limitations

Fully Automated Questions Generation

In the current pipeline for generating questions in MotiveBench, we still rely on human effort to manually check and revise the quality of questions. This approach poses challenges for automatically refreshing the benchmark to avoid data contamination for future LLM releases and for scaling to a larger set of test questions.

To address these challenges, it is necessary to train a revision model using our existing manually corrected data, with the goal of fully automating the entire question generation pipeline. This would enable the benchmark to dynamically update itself, thereby maximizing its value in the rapidly evolving era of LLMs. To achieve this, a potential approach involves leveraging advanced techniques, focusing particularly on fine-tuning existing pretrained models with our manually corrected dataset. This process will help the model learn the intricate patterns and nuances required for high-quality question generation. Additionally, we plan to incorporate continuous learning mechanisms, allowing the model to adapt to new data and evolving trends. By doing so, we aim to enhance not only the accuracy and relevance of the generated questions but also ensure that the benchmark remains aligned with the latest developments in the field of large language models. The result will be a more dynamic, scalable, and efficient system capable of keeping pace with advancements in AI technology.

From Situational QA to Realistic Simulations

MotiveBench fundamentally employs the "situational question-answering" paradigm, where LLMs are prompted to answer questions about the next immediate step in various scenarios. However, this paradigm still deviates from real-world human social activities, where individuals take a sequence of actions, form longer life stories, and behave spontaneously without being asked, "what will you do?".

To overcome this limitation, we can consider using the paradigm of LLMs performing role-playing in a simulation sandbox (Wang et al., 2024b). Given an initial scenario, LLMs act as specific characters with preset personas to engage in daily activities or achieve goal-oriented tasks. By analyzing the behavior trajectories elicited in the sandbox, we can assess LLMs' motivational reasoning and proactive action-taking capabilities in a more comprehensive manner.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *arXiv preprint arXiv:1607.00070*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology*, 52(1):1–26.

Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Barbara Brehm. 2014. *Psychology of health and fitness*. FA Davis.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology*, 1(20):416–436.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alice H Eagly and Wendy Wood. 2012. Social role theory. *Handbook of theories of social psychology*, 2:458–476.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Sandra Graham. 1991. A review of attribution theory in achievement contexts. *Educational Psychology Review*, 3:5–39.

Martin Hagger and Nikos Chatzisarantis. 2005. *The social psychology of exercise and sport*. McGraw-Hill Education (UK).

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Henry J Bindel Jr. 1991. Becoming an effective classroom manager: A resource for teachers, by bob f. steere.

Yeonghun Kang and Jihan Kim. 2023. Chatmof: An autonomous ai system for predicting and generating metal-organic frameworks. *arXiv preprint arXiv:2308.01423*.

Alan E Kazdin, American Psychological Association, et al. 2000. *Encyclopedia of psychology*, volume 8. American Psychological Association Washington, DC.

Simon Keizer, Milica Gasic, Filip Jurcicek, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Parameter estimation for agenda-based user simulation. In *Proceedings of the SIGDIAL 2010 Conference*, pages 116–123.

Florian Kreyssig, Iñigo Casanueva, Pawel Budzianowski, and Milica Gasic. 2018. Neural user simulation for corpus-based policy optimisation for spoken dialogue systems. *arXiv preprint arXiv:1805.06966*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. 2024. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*.

Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj mental health research*, 3(1):4.

AH Maslow. 1943. A theory of human motivation. *Psychological Review google schola*, 2:21–28.

Nikhil Mehta, Milagro Teruel, Patricio Figueroa Sanz, Xin Deng, Ahmed Hassan Awadallah, and Julia Kiseleva. 2023. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. *arXiv preprint arXiv:2304.10750*.

AI Meta. 2024. Introducing llama 3.1: Our most capable models to date. *Meta AI Blog*.

Rémi Radel, Dusan Pjevac, Karen Davranche, Fabienne d'Arripe Longueville, Serge S Colson, Thomas Lapole, and Mathieu Gruet. 2016. Does intrinsic motivation enhance motor cortex excitability? *Psychophysiology*, 53(11):1732–1738.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018b. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Steven Reiss. 2004. Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of general psychology*, 8(3):179–193.

Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

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

Mahni Shayganfar, Charles Rich, and Candace L Sidner. 2016. An overview of affective motivational collaboration theory. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. 2024a. Towards objectively benchmarking social intelligence for language agents at action level. *arXiv preprint arXiv:2404.05337*.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2024b. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. *arXiv preprint arXiv:2412.05631*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

David E Wilkins. 2014. *Practical planning: extending the classical AI planning paradigm*. Elsevier.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How far are we from believable ai agents? a framework for evaluating the believability of human behavior simulation. *arXiv preprint arXiv:2312.17115*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yu Yang, Stephen J Read, and Lynn C Miller. 2009. The concept of situations. *Social and Personality Psychology Compass*, 3(6):1018–1037.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Details of the Hierarchy of Needs

Maslow's hierarchy of needs (Maslow, 1943) is a motivational theory proposed by American psychologist Abraham Maslow in 1943, which explains the prioritization and fulfillment of human needs. He categorized human needs into five levels, ranging from basic survival needs to higher psychological needs, ultimately culminating in self-actualization. Below is a detailed description of each level of need, along with corresponding example scenarios from our MOTIVEBENCH as shown in Figure 8.

- **Physiological Needs**: Physiological needs are the most fundamental survival needs for humans, including air, water, food, warmth, and sleep. These needs form the base of the hierarchy, and only once they are met can individuals engage in other activities.

- **Safety Needs**: Once physiological needs are met, humans seek security, which includes physical safety, financial stability, health, and environmental consistency. This need reflects an individual's desire for order, protection, and predictability in the future.

- **Love and Belonging Needs**: Also known as social needs, people seek to build interpersonal relationships, gain friendships, experience love, and feel a sense of belonging to a social group. This need involves the desire to integrate into society, be accepted, and interact with others.

- **Esteem Needs**: Esteem needs are divided into intrinsic esteem (self-confidence, self-respect, and independence) and extrinsic esteem (recognition, appreciation, and status from others). Satisfying these needs boosts an individual's sense of self-worth and social standing.

- **Self-Actualization Needs**: The highest level of need, self-actualization refers to the individual's pursuit of fulfilling their potential and achieving their ideal self. Self-actualization is typically manifested through creativity, personal growth, and realizing one's own value.

Furthermore, the Reiss Motivation Profile (RMP) (Reiss, 2004), proposed by psychologist Steven Reiss, is also a theoretical model that aims to understand an individual's motivation by analyzing their preferences across 16 fundamental needs. Each person responds differently to these 16 needs, and this variation determines their behavior, decision-making, and lifestyle. The core assumption of the RMP model is that individual behaviors are driven by these needs, and the intensity of these needs shapes one's behavioral patterns.

Fundamental human needs such as **Curiosity**, **Idealism**, and **Honor** motivate individuals to explore knowledge, uphold ethical principles, and seek validation from society, representing a pursuit of cognitive growth and moral fulfillment. Desires for control and autonomy are embodied in **Independence** and **Power**, where those driven by independence prioritize freedom and self-governance, while those seeking power focus on exerting influence and leadership. After satisfying cognitive and control-oriented needs, social and emotional desires come to the forefront: **Acceptance**, **Family**, and **Romance** emphasize the importance of close relationships and emotional intimacy, whereas **Social Contact** and **Status** reflect the longing for social integration and external recognition. Practical concerns for stability are seen in the needs for **Order** and **Saving**, which focus on creating structure and security in life. Simultaneously, the pursuit of personal well-being emerges through **Tranquility** and **Physical Activity**, highlighting the desire for peace of mind and physical health. Finally, more primal or visceral needs, such as **Vengeance** and **Eating**, speak to coping with perceived injustice and savoring the pleasures of sensory experiences, offering a reminder of the complexities inherent in human motivation.

B Question Format Example

Figure 9 shows an example of three types of tasks in the same scenario, from Persona-Hub.

Level 5: Self-Actualization Needs	**Example Scenario:** Samantha, a graduate student majoring in analytical psychology, is deeply engaged in her thesis research. Recently, she attended an international conference and participated in a challenging debate with Dr. Martin, a respected philosopher, on topics related to free will and determinism. Afterward, she continued to reflect on the discussion, thinking about its potential connections to her academic work. Additionally, she had received constructive feedback from her advisor on ways to develop her research further.
Level 4: Esteem Needs	**Example Scenario:** Suresh, an avid local football fan from Gujarat, operates a well-known blog focused on local sports events. He has supported his favorite team for many years and frequently uses his platform to highlight their achievements. Recently, he was present at a critical match between two rival teams, during which his favorite team lost due to a disputed penalty call. In the aftermath of the match, despite facing backlash and threats from upset fans of the rival team, Suresh dedicated several hours to writing a detailed post that critiqued the referee's decision and its implications for the game.
Level 3: Love & Belonging Needs	**Example Scenario:** As Harry dashed through the downpour, he stumbled upon a quaint little pub that seemed like an oasis amidst the stormy evening, its warm glow inviting him to take shelter. Recently moved to a new city, Harry felt increasingly isolated. Inside the pub, he was drawn to the lively atmosphere and began trying different craft beers. Nervous about attending a wedding alone, he found a friendly bartender and chatted about shared interests like classic rock music. After an hour of engaging conversation, Harry invited the bartender to join him at the wedding.
Level 2: Safety Needs	**Example Scenario:** Owen, an IT specialist, is meticulous and focuses on creating a secure and hazard-free home office environment, taking great care in managing every aspect of his workspace to avoid potential risks. Recently, he invested in a range of safety measures, including fireproof storage for sensitive documents and high-quality surge protection for his electronics. He often discusses office safety practices with his peers and follows several technology newsletters.
Level 1: Physiological Needs	**Example Scenario:** Yara, a college student balancing two part-time jobs and maintaining a demanding academic schedule, browses online reviews one evening in search of an affordable brunch spot. She comes across a quaint diner known for its sizable breakfast platters and quick service. Yara decides to go for brunch the next morning. She arrives at the diner, orders a large breakfast platter, and devours her meal efficiently. After the meal, she shows notable satisfaction.

Figure 8: Example scenarios for Maslow’s hierarchy of needs in MotiveBench.

An example from Persona-Hub

1. Motivational Reasoning Question

Context: Dr. Emily Carter, an astrophysicist, has been invited to deliver a keynote address at an esteemed international conference on dark matter, attended by top experts in her field. The conference is known for its rigorous standards and high expectations for speakers. In the days before her speech, she spends long evenings preparing her presentation, often sacrificing rest and personal time. Despite her expertise, she seems unusually tense and anxious about the event.

Question: What is the most likely motivation behind her meticulous preparation for the speech?",

Options:

A. She wants to impress her colleagues and friends.

B. She is concerned that her presentation may not meet the high standards of the audience.

C. She hopes her presentation will lead to future research funding.

D. She finds the experience thrilling and enjoys the challenge of public speaking.

E. She wishes to build professional connections with other scientists.

F. She feels driven to share her work to advance scientific knowledge.

Correct Answer: B

2. Behavioral Reasoning Question

Context: Dr. Emily Carter is a respected astrophysicist with a strong commitment to her work and her family. She has received an offer for a prestigious position as a visiting professor at a top university overseas, which would enhance her career and provide financial benefits. Despite her professional success, she is very close to her family, especially her younger brother, who requires her support due to a disability. In addition to her work, she volunteers in her community to help families in similar situations.

Question: Faced with this opportunity, driven by her desire to balance her career and family responsibilities, what behavior is Dr. Carter most likely to exhibit in response to this offer?

Options:

A. Accept the position immediately, viewing it as a unique chance to advance her career without hesitation.

B. Reject the offer outright, choosing to dedicate herself fully o her family and the needs of her brother.

"C. Propose a modified working arrangement that allows her to fulfill both her professional duties and family obligations.

D. Take a temporary leave from her current job to focus on supporting her family during this transition.

E. Agree to the position but plan to take frequent extended breaks to care for her brother.

F. Delay her decision for several months, considering how this change might affect her personal life.

Correct Answer: C

3. Motive&Behavior Reasoning Question

Context: Dr. Emily Carter, a renowned astrophysicist, has just returned from a year-long sabbatical where she conducted groundbreaking research on dark matter. Upon returning, she faces a significant backlog of urgent emails and requests from colleagues and students, including a tight deadline for a grant application and a critical experiment that requires her oversight. Furthermore, her younger brother, diagnosed with a chronic illness during her absence, needs ongoing support.

Question: Given this scenario, what kind of behavior is Dr. Carter most likely to exhibit next, and what might motivate that behavior?

Options:

A. Prioritize her research to recover lost ground, motivated by the desire to reclaim her status in the academic community.

B. Request an extended leave to fully support her brother, motivated by her commitment to family obligations.

C. Strategically delegate tasks to her team to manage her workload, motivated by the need to harmonize her professional duties with personal responsibilities.

D. Concentrate exclusively on her career, neglecting her brother's needs, motivated by an unyielding ambition for professional achievement.

E. Organize meetings to share her research insights and delegate responsibilities, motivated by a desire to reintegrate into her professional environment effectively.

F. Work excessive hours to complete all tasks independently, motivated by a strong sense of personal obligation.

Correct Answer: C

Figure 9: An example question from Persona-Hub.

C Human Annotation Results

C.1 Recruitment of Data Annotators

From Dec.18, 2024 to Jan.13, 2025, we recruited participants through a crowdsourcing platform to annotate questions. A total of 28 individuals registered for the study. Based on their performance in a small set of trial annotations, 15 participants were selected (rejection rate: 46.43%).

To comprehensively assess the impact of demographics and perspectives on answer selection, we recruited annotators from **diverse backgrounds**, including fields such as minority language translation, public administration, metallurgical engineering, law, chemistry, IoT engineering, art, social demography, and computer science. Their **education levels** range from undergraduate to PhD, with an **average age** of 23.7.

C.2 Requirements for Annotation Task

Each annotator is responsible for evaluating both the **correctness** and **reasonableness** of the multiple-choice questions. The task is divided into two main components:

- For each MCQ, annotators assign a total of 6 points across the available options based on their assessment of the correctness of each option. Higher scores indicate greater correctness. The points must be allocated as whole integers, with the total sum of points for each question always equaling 6.

- Annotators also evaluate the overall reasonableness of each question. A question is deemed unreasonable if any of the following conditions are met: 1) The scenario presented is logically inconsistent or flawed; 2) Insufficient background information is provided, making it impossible to derive any meaningful answer; 3) None of the answer options are reasonable or relevant; 4) There are three or more answer options with equally high correctness scores, causing ambiguity and reducing the clarity of the question.

To ensure objectivity and thoroughness, each question is independently reviewed by five different annotators. Each annotator dedicates 8 hours to annotate a set of 200 questions. To maintain the integrity of the task and prevent any potential answer leakage or bias, annotators are restricted to viewing and annotating only one question per scenario.

This approach ensures that they have no exposure to related questions within the same context, safeguarding the independence of their judgments.

C.3 Analysis of Annotation Results

A comprehensive analysis of annotation quality, question consistency, and response accuracy reveals the following key findings:

- **Strong inter-annotator consistency**: Pearson correlation coefficients between annotator pairs range from 0.692 to 0.868, indicating a high level of alignment in their evaluations. Moreover, the absence of unreasonable cases highlights the clarity and robustness of the questions, as well as the accuracy of the annotations.
- **High agreement rates**: 98.17% of the questions achieved agreement among at least three annotators, while 82.50% reached consensus among four annotators. This demonstrates exceptional consistency in judgment and further confirms the reasonableness of the questions.
- **Accurate consensus answers**: By aggregating the weights assigned by annotators to each option and selecting the one with the highest cumulative score, we find that consensus answers match our pre-defined correct answers at a rate of 93.00%, thereby validating the accuracy of our predefined answers.

These annotation results suggest that the questions were generally perceived as well-constructed, reasonable, and logically sound by the annotators. The high consistency in the annotations further indicates that the evaluative criteria were clear and effectively applied across different annotators. Additionally, the analysis reveals that demographic factors, such as individual backgrounds or perspectives, appear to have little to no influence on the final answer choices, suggesting that our MotiveBench can make an objective and reasonable assessment of motivational intelligence.

Table 4: Overview of Annotation Results

Number of Incorrect Questions	0
Number of Correct Questions	600
$\geq$ **3 Agreement (%)**	98.17
$\geq$ **4 Agreement (%)**	82.50
Agreement w/ Answers (%)	93.00

D Evaluated LLMs

We evaluate 29 popular LLMs across a range of parameter sizes, including several models from the GPT series (Hurst et al., 2024) (GPT-4o 2024-05-13, GPT-4o mini 2024-07-18, and GPT-3.5-Turbo 1106), the LLaMA series (Meta, 2024; Touvron et al., 2023) (LLaMA 3.1 and LLaMA 2), the Qwen series (Hui et al., 2024; Yang et al., 2024; Bai et al., 2023) (Qwen 2.5, Qwen 2, and Qwen), the Phi series (Abdin et al., 2024) (Phi 3.5 and Phi 3), the GLM series (GLM et al., 2024) (ChatGLM 3 and GLM 4), as well as other models like Baichuan 2 (Yang et al., 2023) and Yi 1.5 (Young et al., 2024). These models span a wide spectrum of architectures and parameter configurations, offering a comprehensive evaluation of current LLM performance across various tasks and benchmarks. All of our experiments are conducted on a machine with four A100 80GB GPUs.

E Detailed Results of three tasks

Tables 5 and 6 present the performance of different models on motivation reasoning, behavior reasoning, and motivation-behavior reasoning tasks. We observe that larger models tend to perform better than smaller models.

F Experiment Prompts

Table 7 provides a comprehensive overview of the prompts used for model evaluation in MotiveBench, including both the base prompt and the CoT (Chain-of-Thought) prompt. Additionally, Tables 8, 9, and 10 detail the specific prompts designed for the questioner, reviewer, and modifier roles within our multi-agent question generation and refinement framework.

To ensure high-quality and well-calibrated question modifications, we empirically set the maximum number of modification rounds to 5. The process proceeds iteratively, with the three reviewers assessing different aspects of the generated questions. If no further issues are detected within this threshold, the refinement process is terminated early to maintain efficiency while preserving quality. This structured approach balances thoroughness and computational efficiency, ensuring that the final questions meet our predefined standards.

$\mathbb{MRQ}$: Motivational Reasoning Question
$\mathbb{BRQ}$: Behavioral Reasoning Question
$\mathbb{MBQ}$: Motive&Behavior Reasoning Question

Base Method	**Vitural Profiles**			**Real-life Scenarios**					
	Persona-Hub			**Amazon**			**Yelp**		
LLMs	$\mathbb{MRQ}$	$\mathbb{BRQ}$	$\mathbb{MBQ}$	$\mathbb{MRQ}$	$\mathbb{BRQ}$	$\mathbb{MBQ}$	$\mathbb{MRQ}$	$\mathbb{BRQ}$	$\mathbb{MBQ}$
Baichuan2-7B-Chat	**73.33**	63.00	55.83	63.67	63.00	73.33	69.00	55.33	65.33
Baichuan2-13B-Chat	**73.33**	**70.17**	**66.17**	**75.33**	**79.67**	**78.67**	**79.67**	**73.00**	**77.67**
ChatGLM3-6B	82.33	68.83	62.00	72.00	73.33	74.00	76.00	70.00	74.00
GLM4-9B-Chat	**90.83**	**80.17**	**76.33**	**92.67**	**87.67**	**88.33**	**93.33**	**83.00**	**84.67**
Yi1.5-6B-Chat	74.33	62.83	54.00	79.67	78.00	70.33	83.00	73.00	68.33
Yi1.5-9B-Chat	90.33	**80.33**	75.67	**92.00**	83.33	87.33	88.33	83.00	84.67
Yi1.5-34b-Chat	**93.33**	80.17	**79.33**	90.67	**87.33**	**92.33**	**95.33**	**85.67**	**91.00**
Phi3-mini-4k-Instruct	87.67	80.50	76.17	84.67	85.67	89.00	86.33	76.67	87.00
Phi3-small-8k-Instruct	87.83	80.83	75.17	91.00	86.67	90.67	86.00	82.00	84.67
Phi3-medium-4k-Instruct	92.67	**86.67**	**82.33**	91.33	**92.00**	**93.67**	**97.00**	**86.00**	90.67
Phi3.5-mini-Instruct	87.67	81.17	77.17	87.67	85.67	87.33	89.00	79.00	88.00
Phi3.5-MoE-Instruct	**93.33**	86.50	81.17	**92.67**	84.67	90.00	93.33	85.00	**91.00**
Llama2-7B-Chat	58.33	44.00	47.33	55.00	44.00	62.33	56.67	41.33	48.00
Llama2-13B-Chat	82.33	73.33	68.83	79.33	78.67	80.33	81.00	74.00	74.67
Llama2-70B-Chat	84.67	71.83	66.67	86.67	82.00	79.00	81.67	73.33	75.00
Llama3.1-8B-Instruct	92.83	82.50	74.83	**91.33**	89.00	88.33	89.33	80.00	87.33
Llama3.1-70B-Instruct	**94.67**	**90.17**	**87.17**	90.67	**95.00**	**95.67**	**94.67**	**89.67**	**93.33**
Qwen-7B-Chat	79.00	72.00	63.83	76.33	76.33	75.67	77.67	71.00	76.33
Qwen-14B-Chat	89.17	84.17	76.17	89.33	86.67	86.33	90.00	81.00	86.67
Qwen-72B-Chat	92.17	84.67	81.83	92.33	90.00	92.67	91.67	84.00	88.33
Qwen2-7B-Instruct	89.17	85.83	81.17	89.67	89.33	89.33	93.33	83.67	91.67
Qwen2-72B-Instruct	93.67	88.00	85.83	93.33	93.00	94.33	93.33	89.67	92.67
Qwen2.5-7B-Instruct	89.67	84.00	83.67	91.00	87.33	91.00	84.00	82.00	91.67
Qwen2.5-14B-Instruct	94.33	88.67	84.67	92.00	90.67	93.33	**95.00**	88.00	**94.00**
Qwen2.5-32B-Instruct	95.17	**89.33**	87.00	**97.33**	**93.67**	**96.67**	**95.00**	**93.00**	92.67
Qwen2.5-72B-Instruct	**95.33**	88.83	**87.17**	94.67	**93.67**	96.33	94.00	88.67	**94.00**
GPT-3.5	89.67	85.17	76.00	93.00	91.33	91.00	93.33	83.00	84.00
GPT-4o mini	93.67	87.50	87.33	94.33	92.33	96.33	95.33	88.00	91.33
GPT-4o	**95.00**	**89.50**	**87.83**	**95.00**	**95.00**	**99.33**	**96.33**	**93.00**	**93.67**

Table 5: The experimental results of the vanilla prompt-based method on the three types of tasks.

MRQ: Motivational Reasoning Question
BRQ: Behavioral Reasoning Question
MBQ: Motive&Behavior Reasoning Question

CoT Method	**Vitural Profiles**			**Real-life Scenarios**					
	Persona-Hub			**Amazon**			**Yelp**		
LLMs	MRQ	BRQ	MBQ	MRQ	BRQ	MBQ	MRQ	BRQ	MBQ
Baichuan2-7B-Chat	**76.00**	**61.17**	**57.17**	**66.67**	**62.67**	**71.67**	**66.00**	53.00	**59.00**
Baichuan2-13B-Chat	48.17	56.33	47.33	55.33	60.00	61.67	56.33	**54.33**	55.67
ChatGLM3-6B	71.83	61.83	54.83	63.33	65.67	66.00	64.00	59.00	66.67
GLM4-9B-Chat	**89.67**	**81.17**	**75.17**	**92.00**	**89.00**	**89.33**	**92.33**	**81.67**	**89.00**
Yi1.5-6B-Chat	84.50	75.50	69.00	79.67	81.00	84.33	81.67	75.00	82.00
Yi1.5-9B-Chat	87.50	**80.67**	72.00	87.00	81.67	**87.00**	**86.00**	**81.67**	83.67
Yi1.5-34b-Chat	**90.00**	80.33	**78.33**	**89.00**	**89.00**	85.67	**86.00**	80.00	**85.33**
Phi3-mini-4k-Instruct	80.33	73.00	66.00	80.00	73.33	76.67	80.67	68.67	79.00
Phi3-small-8k-Instruct	**89.83**	**84.00**	**83.17**	**94.67**	**89.33**	**94.00**	94.67	**85.67**	**93.00**
Phi3-medium-4k-Instruct	88.50	81.33	78.50	87.33	83.00	85.67	91.33	80.33	87.67
Phi3.5-mini-Instruct	81.50	77.83	73.67	82.33	79.00	82.00	85.67	76.67	81.67
Phi3.5-MoE-Instruct	77.33	74.00	67.83	84.33	81.67	81.33	**95.33**	83.00	91.00
Llama2-7B-Chat	66.00	58.17	45.33	59.33	59.00	63.00	66.00	57.00	62.67
Llama2-13B-Chat	79.50	68.17	58.67	70.00	73.33	72.33	78.67	66.67	72.67
Llama2-70B-Chat	88.33	82.17	76.00	85.33	85.00	84.00	89.33	80.33	85.33
Llama3.1-8B-Instruct	81.17	74.33	69.50	83.67	81.33	83.33	84.67	79.33	83.67
Llama3.1-70B-Instruct	**90.67**	**85.67**	**80.50**	**92.00**	**94.67**	**95.00**	**94.00**	**85.00**	**90.67**
Qwen-7B-Chat	78.83	72.67	64.67	77.67	77.33	79.33	80.67	66.00	75.67
Qwen-14B-Chat	86.50	80.50	73.00	89.33	83.00	83.00	89.33	79.00	81.33
Qwen-72B-Chat	91.00	84.33	81.50	87.00	87.33	88.67	88.67	83.67	87.33
Qwen2-7B-Instruct	88.67	84.83	83.50	91.00	89.33	88.33	93.00	83.00	93.00
Qwen2-72B-Instruct	91.00	84.83	83.17	87.00	87.33	88.67	88.67	83.67	87.33
Qwen2.5-7B-Instruct	88.67	84.83	81.67	88.33	91.33	87.33	86.67	80.33	90.00
Qwen2.5-14B-Instruct	**93.33**	85.50	82.67	90.67	90.00	94.33	93.33	86.33	90.67
Qwen2.5-32B-Instruct	93.00	**86.83**	**84.50**	**95.00**	**94.00**	**95.67**	94.67	**87.67**	90.67
Qwen2.5-72B-Instruct	91.67	84.00	84.00	93.00	92.67	94.33	**95.33**	**87.67**	**93.67**
GPT-3.5	88.50	83.33	79.00	92.67	92.67	89.67	92.67	79.67	87.00
GPT-4o mini	92.33	**89.33**	84.83	**93.33**	93.00	94.67	94.33	85.67	**93.00**
GPT-4o	**93.00**	87.67	**87.33**	91.67	**93.33**	**96.33**	**96.00**	**88.67**	92.00

Table 6: The experimental results of the CoT prompt-based method on the three types of tasks.

Base Prompt for Evaluation

The following is a {Question_Type}. You should: {Type_Interpretation}. Carefully read the given question, fully immerse yourself in the role of the character described, and reason based on the information provided. Your answer should rely strictly on the given details.
Note:
1. Based on the content of the given question, please infer the most likely answer.
2. You must select one answer from the given options: "A, B, C, D, E, F" as the most likely choice. Even if the question does not provide sufficient information to determine the correct answer, you should randomly choose one option as your output.
3. The result can only return **one character without any other explanation**.

Question: {Question_Content}

Options: {Options}

CoT Prompt for Evaluation

The following is a {Question_Type}. You should: {Type_Interpretation}. Carefully read the given question, fully immerse yourself in the role of the character described, and reason based on the information provided. Your answer should rely strictly on the given details.
Note:
1. Based on the content of the given question, please think step by step and infer the most likely answer.
2. You must select one answer from the given options: "A, B, C, D, E, F" as the most likely choice. Even if the question does not provide sufficient information to determine the correct answer, you should randomly choose one option as your output.
3. Please first think through the question step by step, analyze the reasoning process for the possible answers, and finally output the most likely answer's letter. **The last line of your reply should only contain one character of your final choice.**

Question: {Question_Content}

Options: {Options}

Illustration

{Question_Type}: {Type_Interpretation}
1. **Motivational Reasoning Question**: Based on the given scenario and the character's profile, determine the most likely motivation behind the character's behavior.
2. **Behavioral Reasoning Question**: Based on the given scenario and the character's profile, determine the most likely behavior the character would take next, given the motivation.
3. **Motive&Behavior Reasoning Question**: Based on the given scenario and the character's profile, determine the most likely motivation the character would develop next and the corresponding behavior would take.

{Options}
To minimize the potential bias caused by the order of options, we will randomize the order of options six times and calculate the average result from these six experiments. Specifically, the order sequences used will be: [1, 2, 3, 4, 5, 6], [6, 5, 4, 3, 2, 1], [3, 1, 6, 5, 4, 2], [2, 3, 5, 6, 1, 4], [5, 4, 1, 2, 6, 3], and [4, 6, 2, 1, 3, 5], ensuring that each option appears in every possible position across the six sequences.

Table 7: Prompts for evaluation.

Prompts of Questioner

Consider the four elements of scenario, profile, motivation, and behavior. In a given scenario, a character with a specific profile will perform a certain behavior based on a certain motivation. You are a professional psychologist and sociologist, skilled at creating challenging reasoning questions based on given scenarios to test participants' motivation and behavior reasoning abilities.

Please create three questions based on the given scenario:

1. Motivational Reasoning Question: Given a complex scenario, a specific profile, and a given behavior, infer the most likely motivation behind the character's behavior. The question should not contain any direct description related to the predicted motivation.
2. Behavioral Reasoning Question: Given a complex scenario, a specific profile, and a given motivation, infer the most likely behavior the character will perform based on that motivation. The question should not contain any direct description related to the predicted behavior.
3. Motive&Behavior Reasoning Question: This is a more advanced test. The question should only include the complex scenario and the character's profile. Using only the complex scenario and specific profile, infer the most likely motivation the character will have and the corresponding behavior they will perform.

To summarize, all three questions are based on the same story scenario and character profile setup. The motivation reasoning question requires the addition of a behavior in the question stem and asks the participant to infer the motivation for that behavior. The behavior reasoning question requires the addition of a motivation in the question stem and asks the participant to infer the behavior that may result from that motivation. The motivation and behavior reasoning question does not need any additional information and requires the participant to infer both the motivation and behavior of the character based on the given scenario and profile.

Note:

1. You will be provided with a simple scenario description. Please rewrite this scenario by correcting any logical inconsistencies, and add relevant details to make the scenario, profile, motivation, and behavior more vivid and complex.
2. Choose the most appropriate motivation and behavior to create the questions. However, ensure that the motivation and behavior are only related to real human needs, not to any POIs or products in the text.
3. The three questions are independent of each other and should be answered separately, meaning that each question should only rely on its own stem and not contain any information from the others. Therefore, please ensure that each question has enough rich and complex scenario and profile information to support correct reasoning.
4. Each question should have only one correct answer, along with five distractors. The distractors must be related to certain parts of the information in the question. Please analyze why each option is correct or incorrect.
5. The question stem must include irrelevant or redundant information that creates distractions and challenges. This is necessary to ensure each question is challenging. The correct answer must not be explicitly stated in the question.

Table 8: Detailed prompts of questioner in the multi-agent framework.

Prompts of Reviewers

You are a strict and discerning psychologist and sociologist, capable of precisely identifying issues in the given behavior and motivation reasoning questions and offering improvement suggestions.

I will provide you with three behavior and motivation reasoning questions. Please evaluate them based on the following aspects:

1. **Reasonableness of the Question Information and Type**: Specifically, all three questions should contain a concrete scenario and character profile. The motivation reasoning question should include additional behavioral information about the character. The behavior reasoning question should include additional motivational information about the character. The motivation and behavior reasoning question should not contain any direct clues about the motivation or behavior.
2. **Logical Consistency and Reasonableness of the Four-Tuple**: Assess whether, in the given scenario, a character with a specific profile would logically perform the stated behavior based on the provided motivation.
3. **Sufficiency of Information to Derive the Correct Answer**: Examine whether the information provided in each question is enough to infer the correct answer. If not, suggest modifications to the scenario or character profile to make the information clearer or more comprehensive.
4. **Challenge and Difficulty of the Question**: Evaluate whether the question presents an appropriate level of difficulty and challenge for the respondent.
5. **Correct Answer Must Not Be Explicitly Stated**: Ensure that the correct answer does not appear explicitly in the question information and can only be deduced through reasoning steps.
6. **Clarity and Plausibility of Distractor Options**: Evaluate whether the incorrect options are misleading and whether they correspond to distracting information within the question. If they do not, suggest adding the relevant distracting information or modifying the options.
7. **Adequate Distractors and Redundant Information**: Ensure that each question includes enough irrelevant or redundant information to make the question challenging, but without disrupting the logic needed to deduce the correct answer.
8. **Objectivity and Neutrality of the Question**: Ensure that the question is presented in a neutral and objective manner, with no implicit suggestion of the correct answer.

Please provide specific modification suggestions for the question set and give your feedback to the question author in a reasonable tone. Summarize your evaluation into a single paragraph of suggestions.

(All the aspects listed above are of concern, and each reviewer will be asked to focus on different aspects.)

Table 9: Detailed prompts of reviewers in the multi-agent framework.

Prompts of Modifier

Consider the four elements of scenario, character profile, motivation, and behavior. In the given scenario, a character with a specific profile will perform a certain behavior based on a particular motivation. You are a professional psychologist and sociologist, skilled in refining motivation and behavior reasoning test questions and providing relevant suggestions for improvement.

The specific types of the three questions are as follows:

1. Motivational Reasoning Question: Based on a complex scenario, a specific character profile, and a given behavior, deduce the most likely motivation behind the character's action. The question should not include any description related to the predicted motivation.
2. Behavioral Reasoning Question: Based on a complex scenario, a specific character profile, and a given motivation, deduce the most likely behavior the character will perform based on that motivation. The question should not include any description related to the predicted behavior.
3. Motive&Behavior Reasoning Question: The question should only include a complex scenario and character profile. This is a more difficult question type, where the respondent must deduce the most likely behavior and corresponding motivation of the character based solely on the scenario and character profile.

In summary, all three questions are based on the same story scenario and character profile settings. For the motivation reasoning question, an additional behavior is given, and the task is to deduce the motivation behind that behavior. In the behavior reasoning question, an additional motivation is given, and the task is to deduce the behavior that would most likely result from that motivation. The motivation and behavior reasoning question, however, does not provide any additional information, requiring the respondent to deduce both the motivation and behavior from the scenario and character profile. It is crucial that the story scenario and character profile in the question are rich enough to support reasoning and lead to the correct answer.

Specific Requirements:

1. Carefully consider each suggestion based on the given questions and selectively make reasonable changes to the questions.
2. Do not delete the distracting information related to the incorrect answers, as this is necessary to ensure the questions remain challenging.
3. The three questions are independent of each other and are to be answered separately. Respondents should only reason based on the question provided, without seeing any other information. Therefore, ensure that each question has sufficiently rich and complex scenario and character profile information.
4. After making revisions, analyze each option to determine why it is correct or incorrect. If there are any issues, modify the question again to ensure the uniqueness of the correct answer.

Table 10: Detailed prompts of modifier in the multi-agent framework.

G Detailed Discussion & Case Study

G.1 General Analysis of GPT-4o's Errors

We analyze instances where GPT-4o's responses deviate from human consensus in motivational and behavioral reasoning tasks. Specifically, for questions where GPT-4o selects incorrect answers, we prompt the model to explain its reasoning behind the chosen option and analyze why the correct answer is appropriate. By systematically comparing incorrect responses with those chosen by human participants, we identify distinct error patterns in GPT-4o's reasoning. These errors fall into four primary categories:

- **Over-Rationalization, Lacking Emotional Insight.** The model prioritizes logical coherence over emotional or social nuances that influence human decision-making.
- **Weak Logical Precision, Prone to General Assumptions.** The model makes broad generalizations, leading to logical imprecision and a failure to anchor its reasoning in specific contextual evidence.
- **Overly Idealistic, Ignoring Complex Realities.** GPT-4o assumes that idealistic or aspirational actions are more probable, even when contextual evidence suggests a more pragmatic or personally relevant behavior.
- **Lack of Awareness of Behavioral Impact.** The model overlooks the real-world impact of behaviors and motivations, often misjudging the alignment between a character's intent and plausible actions.

G.2 Detailed Case Study

The Figure 7 provides an example of GPT-4o's reasoning flaws in a motive and behavior reasoning question. The scenario involves Samantha, a country music artist with deep community ties, who participates in a BBQ collaboration and values authenticity and audience connection. The task asks what behavior she is most likely to exhibit next, considering her motivations.

GPT-4o selects Option D (organizing a community concert to raise funds), justifying this choice by assuming that Samantha's community ties naturally imply a desire to support local causes through fundraising. However, this reasoning is **overly idealistic** because there is no contextual evidence suggesting she is motivated by philanthropy. Instead, the scenario emphasizes her **personal artistic expression and audience engagement**, making composing a song about her BBQ experiences (Option C) the more natural choice.

Furthermore, GPT-4o's reasoning demonstrates **weak logical precision** by assuming that community involvement directly translates into organizing fundraising events, rather than considering more direct forms of artistic expression that align with Samantha's profession. In contrast, human participants recognize that a musician would most naturally process experiences through songwriting rather than logistical event organization.

Beyond this case, our analysis includes multiple other examples illustrating these four error categories. These cases appear in the Tables 11, 12, 13, 14, 15, where we provide a systematic breakdown of how GPT-4o's errors manifest across different behavioral reasoning tasks. Refining model training to emphasize emotional intelligence, contextual precision, and realistic behavioral motivations could enhance its alignment with human-like reasoning patterns.

Case 1 - Motive&Behavior Reasoning Question

Context: Samantha, a grateful accident survivor, was involved in a severe car crash six months ago. She has since undergone multiple surgeries and intense physical therapy. As a freelance writer before the accident, Samantha now spends much of her free time reading and writing poetry, which she shares occasionally on her personal blog. She has been sharing her recovery journey on social media and feels strongly about using her experience to make a positive impact. Recently, she has gained a lot of traction and connected with many individuals through her posts, deepening her sense of responsibility to those who are still struggling.

Question: What kind of behavior is Samantha most likely to exhibit next, and what is the motivation behind it?
Options:
A. Motivation: Cultivating a personal brand; Behavior: Posting artistic photos of her daily life and updates about her writing process.
B. Motivation: Highlighting the importance of emotional resilience; Behavior: Hosting online webinars focused on mental health strategies.
C. Motivation: Encouraging community support; Behavior: Organizing small group meetups for accident survivors to share their experiences.
D. Motivation: Seeking validation from peers; Behavior: Posting emotionally charged poetry on social media to gain likes and shares.
E. Motivation: Gaining recognition for her journey; Behavior: Collaborating with influencers to promote her story.
F. Motivation: Finding solace through expression; Behavior: Writing a memoir to reflect on her healing process.

Correct Answer: C
GPT-4o's Answer: B

Analysis: In this scenario, the appropriate action should be to offer emotional support, not discuss mental health strategies. GPT-4o's choice (hosting a webinar on mental health strategies) focuses on professional methods, which doesn't align with Samantha's current situation. She seeks to inspire others through her personal experiences, not teach strategies. GPT-4o's reasoning is too theoretical, lacking the empathy and life experience humans use to understand motivations. Humans, considering Samantha's struggles, would focus on actions that resonate with her personal healing process, such as sharing her story to help others.

Table 11: Case 1 on a Motive&Behavior Reasoning Question.

Case 2 - Motive&Behavior Reasoning Question

Context: James, a seasoned stockbroker specializing in tech and software stocks, has recently noticed a growing interest among younger investors in sustainable and socially responsible investments. Despite his initial skepticism, he recognizes the potential effects of this trend on his career. Additionally, he faces increasing competition from newer, digitally-savvy brokers capitalizing on this shift. Furthermore, he has come across various articles detailing the increasing demand for sustainable investments.

Question: Given these circumstances, what kind of behavior is James most likely to exhibit next, and what could be the motivation behind it?

Options:
A. He will develop a marketing strategy aimed at promoting sustainable tech stocks, driven by the desire to connect with a younger audience interested in socially responsible investments.
B. He will partner with a fintech firm specializing in sustainable investments, motivated by the necessity to broaden his service offerings and enhance client retention.
C. He will begin writing articles for finance magazines, driven by the ambition to share his insights on the importance of sustainable investing among his peers.
D. He will initiate a webinar series focusing on sustainable investment trends, motivated by the goal of showcasing his expertise and engaging with potential clients.
E. He will host social events for potential investors, driven by the intention to foster relationships and promote discussions around sustainable investing.
F. He will create a newsletter highlighting sustainable investment options, motivated by the aim of educating clients about emerging trends in the market.

Correct Answer: A
GPT-4o's Answer: D

Analysis: When faced with new market trends, humans typically prioritize directly addressing market demands and customer interests. For example, the growing interest of young investors in sustainable investments leads to a marketing strategy tailored to this group, which aligns with real market needs. GPT-4o tends to suggest that James might showcase his expertise through a webinar, but this motivation focuses more on "self-promotion" rather than directly responding to market demands or attracting a specific group, failing to address the competitive pressures and market changes.

Table 12: Case 2 on a Motive&Behavior Reasoning Question.

Case 3 - Motive&Behavior Reasoning Question

Context: Yara, a new mother with a newborn baby girl who has a history of allergies, recently dined at a café that provided detailed ingredient lists and used allergen-safe cooking methods. She was satisfied with the café's attention to allergen management and its ability to cater to her needs. Yara is also known to actively participate in community groups focused on managing allergies in children.

Question: In this scenario, what is Yara most likely to do next, and what is her primary motivation?

Options:
A. Sharing her experience with others, motivated by her commitment to helping the allergy community.
B. Thanking the café staff, motivated by appreciation for their allergen-safe practices.
C. Researching other restaurants, motivated by a desire for variety in dining options.
D. Leaving a negative review elsewhere, motivated by frustration over previous dining challenges.
E. Avoiding dining out altogether, motivated by concerns about public allergens.
F. Offering advice to another parent in the café, motivated by her interest in parenting discussions.

Correct Answer: B
GPT-4o's Answer: A

Analysis: GPT-4o's analysis overlooks emotion-driven behavior by focusing on Yara's rational and altruistic motives, assuming she would share her experience to help others. This perspective ignores the possibility of a direct emotional response, such as expressing gratitude for the cafe service. Furthermore, GPT-4o overinterprets Yara's background in community management, predicting that her actions would be more focused on helping others or sharing experiences, rather than simply thanking the staff. In contrast, humans are more likely to recognize that, despite Yara's involvement in the community, her immediate interaction with the cafe staff would be influenced by her emotional response, such as gratitude, fitting the context of the situation.

Table 13: Case 3 on a Motive&Behavior Reasoning Question.

Case 4 - Motive&Behavior Reasoning Question

Context: Samantha owns a travel agency that specializes in personalized service and unique travel experiences. Recently, she's been thinking about ways to make her business more environmentally friendly and believes that adopting sustainable practices could also boost her agency's reputation. During a meeting with her accountant, Mark, they reviewed various financial strategies to implement her ideas. Although the agency already has basic recycling and uses digital communication to reduce waste, Samantha is determined to make a bigger impact in the competitive travel market.

Question: What kind of behavior is Samantha most likely to exhibit next, and what is the motivation behind it?

Options:
A. Announcing a new program to contribute most profits to local environmental projects, motivated by a desire to build the agency's reputation for community involvement.
B. Rushing to install solar panels on all properties without detailed cost planning, motivated by an urgent need to show visible commitment to sustainability.
C. Delaying new projects until further discussions with stakeholders, motivated by caution about potential financial risks.
D. Expanding the recycling program to engage customers in eco-friendly actions, motivated by a focus on community-based solutions.
E. Launching a promotional campaign about the agency's past sustainable practices, motivated by the desire to draw media attention.
F. Organizing workshops for employees on sustainable practices, motivated by a goal to enhance internal awareness.

Correct Answer: D
GPT-4o's Answer: F

Analysis: GPT-4o overlooked the emphasis on "enhancing market competitiveness" in the question and focused excessively on the superficial logic of "sustainability." However, the purpose of sustainability is to enhance market competitiveness, and merely raising internal employees' awareness does not contribute to improving market competitiveness.

Table 14: Case 4 on a Motive&Behavior Reasoning Question.

Case 5 - Motive&Behavior Reasoning Question

Context: Fiona, a young woman working as an editor for a prestigious publishing house, lives alone in a vibrant urban neighborhood known for its diverse cultures. One afternoon, after a challenging week at her job, she decides to visit Bella Vita, a charming pizzeria in a more upscale area. Bella Vita is famous for its delicious pizzas and warm Italian atmosphere, complete with nostalgic music and friendly staff. As she sits by the window, enjoying the sunlight, she finds herself laughing softly, and her exhaustion starts to fade. The cozy ambiance surrounds her, bringing her feelings of comfort and joy. Fiona highly values her personal time, often enjoying these quiet moments for reflection and renewal, while also cherishing fond memories of family gatherings at similar Italian restaurants.

Question: As she listens to the familiar tunes and observes families enjoying meals together, based on what motivation is she most likely to exhibit what behavior next?

Options:
A. Reach out to a friend to share her experience, motivated by her desire for emotional connection.
B. Jot down her thoughts about the atmosphere, driven by her need for self-expression.
C. Plan to revisit the restaurant with her family, inspired by her longing for shared memories.
D. Explore other nearby restaurants, motivated by her curiosity about the local dining scene.
E. Compliment the staff for their service, reflecting her appreciation for kindness and hospitality.
F. Take a photograph to post online, motivated by her interest in sharing aesthetic moments with others.

Correct Answer: C
GPT-4o's Answer: B

Analysis: GPT-4o tends to over-rely on explicit textual details while overlooking implicit behavioral tendencies and deeper emotional motivations. For instance, it often focuses on directly stated traits in the prompt (e.g., "She is an editor, so she may prefer writing") and limits its reasoning to surface-level information, ignoring how emotions like nostalgia might influence behavior. In contrast, humans naturally consider the emotional undertones within a situation, such as how a familial atmosphere may evoke empathy and drive planning. Additionally, GPT-4o primarily relies on explicit contextual details to infer motivations, whereas humans are more sensitive to subtle emotional cues embedded in the broader scenario, allowing for a more nuanced understanding of behavior.

Table 15: Case 5 on a Motive&Behavior Reasoning Question.