From Stimuli to Minds: Enhancing Psychological Reasoning in LLMs via Bilateral Reinforcement Learning

Yichao Feng*, **Haoran Luo***, **Lang Feng, Shuai Zhao, Anh Tuan Luu**
Nanyang Technological University
College of Computing and Data Science
{yichao.feng, lang.feng, shuai.zhao, anhtuan.luu}@ntu.edu.sg, haoran.luo@ieee.org

Abstract

Large Language Models show promise in emotion understanding, social reasoning, and empathy, yet struggle with psychologically grounded tasks requiring inference of implicit mental states in complex, socially and contextually ambiguous settings. These limitations stem from lacking theory-aligned supervision and difficulty capturing nuanced mental processes in real-world narratives. To bridge this gap, we leverage expert-labeled scenarios and propose a trajectory-aware reinforcement learning framework imitating expert psychological reasoning. By integrating real-world stimuli with structured reasoning guidance, our approach enables compact models to internalize social-cognitive principles, perform nuanced inference, and support continual self-improvement. Experiments across benchmarks show expert-level interpretive capability across psychological tasks.

Code — https://github.com/Githubuseryf/Stimuli2Minds
Extended version — https://arxiv.org/abs/2508.02458

1 Introduction

Large Language Models (LLMs) show strong generalization across diverse language tasks (Wu 2025; Zhang et al. 2025; Feng et al. 2025b). Their emerging potential in psychological domains, such as emotion understanding (Kovacevic et al. 2024), social reasoning (Leng and Yuan 2023), and empathy recognition (Sorin et al. 2024) has drawn growing research interest. These tasks require not only linguistic comprehension but also nuanced inference of implicit mental states and emotional cues, often without supervision or defined ground truths. Unlike conventional language tasks, they involve rich psychological stimuli embedded in ambiguous, socially grounded scenarios shaped by diverse cultural norms, interpersonal dynamics, and lived experiences. Although LLMs show partial sensitivity to such cues, performance on psychologically grounded tasks remains limited from human-level competence (Ke et al. 2025).

To evaluate LLMs' cognitive reasoning, several benchmarks have emerged. **ToMbench** provides a structured framework for assessing Theory of Mind (ToM) across eight task types and 31 social-cognitive skills (Chen et al. 2024),

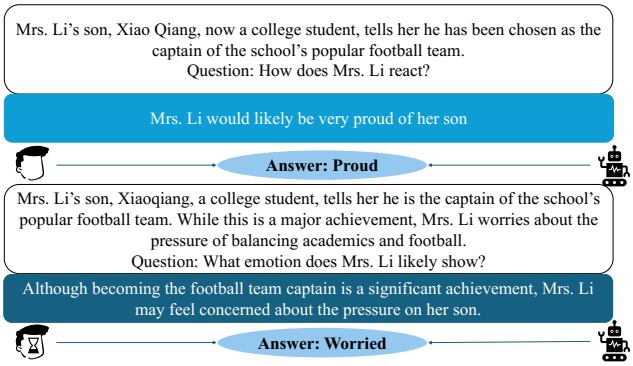

Figure 1: A sample ToMbench question presents two types of psychological stimuli pairs for demonstration.

showing that even advanced models like GPT-4o (Achiam et al. 2023) still lag behind humans. Other benchmarks such as Psychobench (Li et al. 2024) and CogBench (Coda-Forno et al. 2024) examine emotion inference and pragmatic reasoning. These efforts systematize evaluation of social cognition in LLMs. While LLMs show partial sensitivity to mental states, their reasoning remains fragmented, often guided by surface cues rather than coherent mental models, and they struggle to generalize in complex scenarios.

Despite recent advances, three key challenges remain. **(1) High quality data scarcity:** Benchmarks like ToMbench are small in scale (Wu et al. 2023), and many datasets use LLM-generated content (Hu et al. 2024), limiting their value for tuning psychological reasoning (Long et al. 2024). **(2) Reasoning mismatch across tasks:** Psychological tasks differ in cognitive demands. Theories as Epstein's (Epstein 1998) and Fuzzy-Trace Theory (Reyna and Brainerd 1998) separate fast, intuitive reasoning from slow, analytical thought. Uniform strategies may hurt performance on intuition-driven tasks (Ji et al. 2025). As an example shows in Figure 1. **(3) Poor generalization in small models:** Compact LLMs struggle to generalize. Many depend on LLM generated labels or costly prompting (Wang et al. 2024).

To address these limitations, we make three contributions. **First**, we construct a large-scale dataset, **StimuliQA**, grounded in professional psychological theory and real-world interviews. It includes over 3,000 annotated stim-

*These authors contributed equally.

Specific Emotions	A_pos	How much the text expresses positivity	368
	A_neg	How much the text expresses negativity	368
	B_hope	How much the text expresses hope	3279
	B_pride	How much the text expresses pride	3279
	B_grat	How much the text expresses gratitude	3279
	B_fear	How much the text expresses fear	3279
	B_disap	Degree of expressed disappointment	3279
	B_anger	How much the text expresses anger	3279
Targeted Emotions	C_acol	Anger directed towards social collective	232
	C_agov	Anger directed towards government	232
	C_anet	Anger directed towards social network	232
	C_apol	Anger directed towards political entities	232
	C_dnet	Social network-directed disappointment	594
	C_dsoc	Disappointment directed towards society	594
	C_dgov	Disappointment at government	594
	C_dpol	Political-directed disappointment	594
	C_dcol	Disappointment at the social collective	594
	C_pnet	Pride directed towards social network	906
	C_psoc	Pride directed towards society	906
	C_pgov	Pride directed towards government	906
	C_ppol	Pride directed towards political entities	906
	C_pcol	Pride directed towards social collective	906
	C_pnat	Pride directed towards the nation	906
	C_gnet	Gratitude directed towards social network	890
	C_gsoc	Gratitude directed towards society	890
	C_ggov	Gratitude directed towards government	890
	C_gpol	Gratitude directed towards political entities	890
	C_gcol	Gratitude directed towards social collective	890
	C_gnat	Gratitude directed towards the nation	890
Direct	D_pos	Positive change mentioned	3279
	D_neg	Negative change mentioned	3279
	D_red	Redemption mentioned in the coding unit	3279
Self-Identification	S_if	Did this event impact your life nationally?	1879
	S_p/n	How do you feel about this state event?	1879
	S_im	How defining is this event for the nation?	1879
	S_tp	Was this event a national turning point?	1878
	S_co	Connection to the nation through the event	1879
	P_p/n	How do you feel about this personal event?	1400
	P_im	Importance of this event to your identity	1400
	P_tp	Is this event a personal turning point?	1400
	P_co	Connection to the nation through the event	1400
Needs Fulfilment	Co_fu	Fulfilment of communion in the coding unit	1029
	Co_la	Lack of communion fulfilment	1029
	Ef_fu	Fulfilment of efficacy in the coding unit	1029
	Ef_la	Lack of efficacy fulfilment	1029
	Sl_fu	Fulfilment of self-esteem in the coding unit	1029
	Sl_la	Lack of self-esteem fulfilment	1029
	Sc_fu	Fulfilment of security in the coding unit	1029
	Sc_la	Lack of security fulfilment	1029
National Theme	E_att	Degree of expressed attachment	3279
	E_det	Degree of expressed detachment	3279
	H_hdw	Appreciation of hard work mentioned	3279
	H_pro	Appreciation of progress mentioned	3279
	H_nos	Longing for past, heritage in the unit	2426
	J_coh	Social cohesion mentioned in the unit	3279
	J_div	Diversity mentioned in the unit	3279
	J_dis	Discord mentioned in the unit	3279
	J_ineq	Inequality mentioned in the unit	3279

Figure 2: The figure summarizes key psychological parameters across our datasets with number of samples.

uli with 58 psychological variables, subsequently converted into question–answer (QA) pairs, as detailed in Figure 2. **Second**, we introduce **Psy-Interpreter**, a reinforcement learning (RL) framework inspired by dual-system psychological theories. It integrates a trajectory cache and bilateral reasoning (BR) to foster expert-like psychological analysis across diverse tasks. **Third**, we show that compact models trained with the Psy-Interpreter framework and rationale-augmented supervision can rival much larger systems across multiple benchmarks, underscoring the effectiveness of structured RL and reasoning-aware supervision in enhancing efficiency and generalization for LLMs.

We validate our approach through three experiments. **First**, our expert-labeled dataset consistently improves out-of-distribution (OOD) performance in mainstream post-training methods, surpassing datasets without expert annotations. **Second**, the Psy-Interpreter framework with dual-system training substantially outperforms standard training methods across diverse task types. **Third**, models trained under our framework generalize better to unseen psychological tasks and exhibit self-annotation capabilities that enable continual learning, further narrowing the gap between expert supervision and autonomous psychological reasoning.

Together, these contributions form a unified and practical framework for advancing social-cognitive reasoning in LLMs through real-world data, psychologically informed learning, and efficient model optimization. Based on our experiments, we argue that explicit knowledge injection and the imitation of expert psychological thought patterns collectively endow the model with expert-level interpretive capabilities, enabling reliable and context-sensitive reasoning.

2 Related Works

2.1 LLMs in Psychological Tasks

LLMs are increasingly evaluated on psychological tasks involving ToM and moral reasoning. Benchmarks show that top models such as GPT-4o underperform humans in belief reasoning, indicating reliance on superficial cues (Xiao et al. 2025; Strachan et al. 2024). Datasets like SimpleToM, CogBench, and SocialIQa extend evaluations to emotion and social inference (Gu et al. 2024; Coda-Forno et al. 2024; Sap et al. 2019b). In moral reasoning, LLMs may align with or even surpass humans in perceived ethical competence (Liu et al. 2025; Huang et al. 2023), though often via pattern matching. Empathy studies find LLMs capable of emotionally appropriate responses, occasionally preferred over human ones (Ayers et al. 2023; Sorin et al. 2024), albeit with limited contextual nuance (Yang, Ye, and Du 2024). Open models still lag behind proprietary ones (Li et al. 2024), underscoring the need for theory-informed benchmarks and training methods (Xie et al. 2024; Qiu et al. 2024).

2.2 CoT and Reinforcement Learning

Chain-of-Thought (CoT) prompting improves deliberate reasoning by encouraging stepwise thinking (Luo et al. 2025a; Shen et al. 2025), and is combined with RL to align outputs with human preferences (Ouyang et al. 2022). Process supervision and gradient-level feedback, such as in DeepSeek-R1, further enhance intermediate reasoning quality (Luo et al. 2025c). Frameworks like Reflexion and Tree of Thoughts introduce self-evaluation and planning to strengthen coherence (Luo et al. 2025b; Feng et al. 2025a).

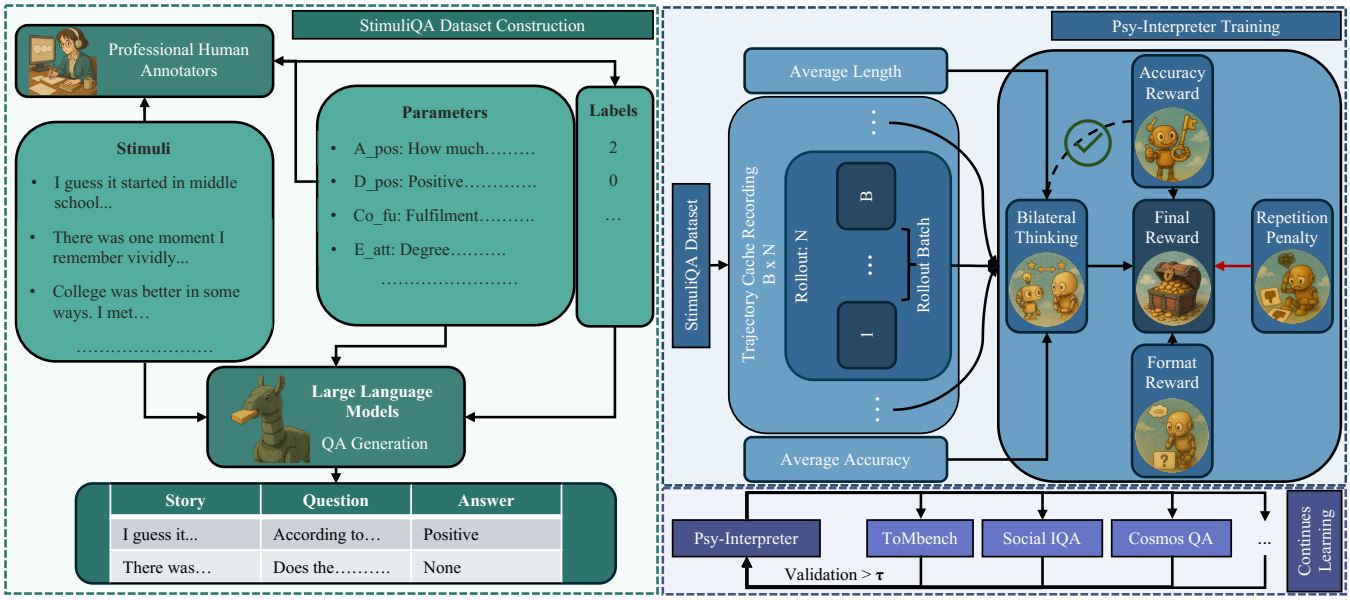

Figure 3: The framework comprises: *StimuliQA*, stimuli with expert psychological labels; *Psy-Interpreter*, a training framework tailored for psychological tasks; and *Continual Learning*, demonstrating continual learning capability through self-labeling.

3 Methodology

This section outlines our methodology for enhancing psychological reasoning as shown in the Figure 3. We first construct **StimuliQA**, a dataset pairing human-labeled narratives with LLM-generated QA. We then propose **Bilateral Reinforcement Learning**, integrating token accuracy, format compliance, reasoning depth, and repetition control into a unified reward. A *Trajectory Cache* stabilizes training via recent rollout tracking. Finally, **Continuous Learning** enables refinement through confident predictions.

3.1 Data Construction

We built **StimuliQA** to support psychological reasoning by combining expert annotations, aiming to inject knowledge and guide LLMs toward human-like reasoning. Comparison with other datasets can be found in the Appendix.

Stimuli Collection and Annotation We collected 3,280 real-life stimuli with emotional and social content. Human annotators labeled them based on variables above, and utilized LLMs to generate QA pairs from the labels.

Psychological Variable Design The dataset comprises 58 variables across three dimensions: **Emotional Reactions** (29 variables), which represent affective and social-emotional responses grounded in Lazarus's appraisal theory (Lazarus 1991) and further examined in narrative psychology settings as in (Han 2025); **Narrative Transformation** (12 variables), reflecting tone shifts and redemptive arcs, inspired by McAdams et al. (McAdams et al. 2001); and **Collective Psychology** (17 variables), encompassing indicators of self-worth and community connection based on Ryff & Keyes's model of psychological well-being (Ryff and Keyes 1995). This variable structure enables exploration of how individuals process and narrate life experiences.

3.2 Bilateral Reinforcement Learning

We propose a Bilateral RL frame with a *Trajectory Cache* that adopt *Trajectory-aware GRPO* based on Group Relative Policy Optimization (GRPO) (Shao et al. 2024) and *Bilateral Reward* to further promote structured reasoning.

Trajectory Cache Fixed rewards are often suboptimal across model scales or tasks. We use a *Trajectory Cache* to track recent performance and adjust rewards by trend, stabilizing estimation. With B batches and cache size C, the batch caches $B_c = B \times C$ summarize training dynamics.

T-GRPO Objective The goal of T-GRPO (Trajectory-aware Group Relative Policy Optimization) is defined as:

$$\mathcal{J}_{\text{T-GRPO}}(\theta) = \frac{1}{B_c G} \sum_{b=1}^{B_c} \sum_{i=1}^{G} \frac{1}{|o_{b,i}|} \sum_{t=1}^{|o_{b,i}|} \left[\min \left(r_{b,i,t}(\theta) \hat{A}_{b,i,t}, \right. \right.$$
$$\left. \left. \text{clip}(r_{b,i,t}(\theta), 1-\varepsilon, 1+\varepsilon) \hat{A}_{b,i,t} \right) \right] - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}), \tag{1}$$

where the importance ratio and normalized advantage are:

$$r_{b,i,t}(\theta) = \frac{\pi_\theta(o_{b,i,t} \mid q_b, o_{b,i,<t})}{\pi_{\theta_{\text{old}}}(o_{b,i,t} \mid q_b, o_{b,i,<t})}, \quad \hat{A}_{b,i,t} = \frac{r_{b,i}^{\text{Final}} - \bar{r}}{\sigma_r + \epsilon}, \tag{2}$$

with $\bar{r}$ and σ_r being the batch-level mean and standard deviation computed over all group-level final rewards $r_{b,i}^{\text{Final}}$ as defined in the *Bilateral Reward* section.

Bilateral Reward For our overall bilateral reward, each group (b, i) is assigned a final reward for $r_{b,i}^{\text{Final}}$ listed here:

$$r_{b,i}^{\text{Final}} = w^{\text{F1}} \cdot r_{b,i}^{\text{F1}} + w^{\text{fmt}} \cdot r_{b,i}^{\text{fmt}} + r_{b,i}^{\text{BR}} - r_{b,i}^{\text{rep}}, \tag{3}$$

where $w^{(\cdot)}$ denotes the weight and $r_{b,i}^{(\cdot)}$ denotes the reward term, which will be introduced in further detail below.

Answer Quality (r^{F1}) We compute $r_{b,i}^{\text{F1}}$ as the token-level F1 score between the predicted answer $\text{Ans}_{b,i}^{\text{pred}}$ the corresponding ground-truth answer $\text{Ans}_{b,i}^{\text{gold}}$. Formally:

$$r_{b,i}^{\text{F1}} = \frac{2P_{b,i}R_{b,i}}{P_{b,i} + R_{b,i}}, \tag{4}$$

$$P_{b,i} = \frac{|y_{b,i}^{\text{pred}} \cap y_{b,i}^{\text{gold}}|}{|y_{b,i}^{\text{pred}}|},\ R_{b,i} = \frac{|y_{b,i}^{\text{pred}} \cap y_{b,i}^{\text{gold}}|}{|y_{b,i}^{\text{gold}}|}, \tag{5}$$

where $y_{b,i}^{\text{pred}}$ & $y_{b,i}^{\text{gold}}$ are token sets after normalization, including lowercase conversion, punctuation cleanup.

Format Compliance (r^{fmt}) To ensure structural consistency, we define a binary reward for format correctness:

$$r_{b,i}^{\text{fmt}} = \begin{cases} 1, & \text{if } o_{b,i} \text{ matches predefined format constraints,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Format validation checks if the response is in the correct format (<think>... </think> <answer>... </answer>).

Bilateral Reasoning (r^{BR}) To encourage informative yet concise reasoning, we introduce a *bilateral reward* term that jointly considers relative reasoning length and answer quality. Let $L_{b,i} = |o_{b,i}|$ denote the output length for group (b, i), and define the batch-level averages as:

$$\bar{L} = \frac{1}{B_c G} \sum_{b=1}^{B_c} \sum_{i=1}^{G} L_{b,i}, \quad \bar{r}^{\text{F1}} = \frac{1}{B_c G} \sum_{b=1}^{B_c} \sum_{i=1}^{G} r_{b,i}^{\text{F1}}. \tag{7}$$

We then compute the bilateral reward as:

$$r_{b,i}^{\text{BR}} = \begin{cases} \delta, & \frac{L_{b,i}}{\bar{L}} < \tau_- \text{ and } r_{b,i}^{\text{F1}} > \bar{r}^{\text{F1}}, \\ \delta, & \frac{L_{b,i}}{\bar{L}} > \tau_+ \text{ and } r_{b,i}^{\text{F1}} > \bar{r}^{\text{F1}}, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where τ_- and τ_+ denote the lower and upper thresholds, respectively. This reward encourages the model to adopt an appropriate reasoning length, using concise responses for simpler cases and extended reasoning for complex ones.

Further, comparing $r_{b,i}^{\text{F1}}$ with $\bar{r}^{\text{F1}}$ ensures that length-based rewards are granted only to outputs that are both accurate and reliably reasoned, preventing the model from receiving incentives for merely producing longer or shorter responses without demonstrating genuine understanding.

Repetition Penalty (r^{rep}) To discourage verbosity and repetition, we compute a penalty over the <think> segment based on the ratio of repeated to total 4-grams:

$$r_{b,i}^{\text{rep}} = -\min\left(\tau_{\text{rep}}, \frac{\text{Repeat}_4(o_{b,i}^{\text{think}})}{\text{Total}_4(o_{b,i}^{\text{think}}) + \epsilon}\right), \tag{9}$$

where τ_{rep} denotes the maximum penalty score for penalizes semantic redundancy and promotes meaningful reasoning.

Other Rewards Additional general reward formulations-including the base reward, short reward, length reward and the length-based reward with repetition penalty (RP) were also utilized and systematically compared in the later section (see Section 4). Full details are provided in the Appendix.

3.3 Continuous Learning

To enable **self-evaluation** and **self-improvement** under low-resource settings, we propose a simple framework for *continuous learning*. The model not only generates predictions but also assesses their quality and improves over time, producing more **user-aligned** and **psychologically appropriate** responses. We define the learning criterion as:

$$\text{self\_train}(x) \iff \text{valid}(x) \wedge \text{confidence}(x) > \tau, \tag{10}$$

where x is a model output, $\text{valid}(x)$ verifies format and $\text{confidence}(x)$ is the model's estimated confidence score.

4 Experiments

This section is organized around five core research questions (RQs): **RQ1:** Can human-labeled psychological data improve model generalization under various training methods? **RQ2:** Do tailored reward functions enhance recognition of explicit and latent psychological states? **RQ3:** Does Psy-Interpreter perform well on OOD datasets? **RQ4:** Is continual learning generally effective? **RQ5:** Can Psy-Interpreter identify if a psychological question requires reasoning?

4.1 Experimental Setup

Datasets We used six different datasets: our StimuliQA and five OOD sets—ToMbench (Chen et al. 2024), SimpleToM (Gu et al. 2024), SocialIQa (Sap et al. 2019a), CosmosQA (Huang et al. 2019), and selected BIG-Bench Hard (Suzgun et al. 2023) subsets (disambiguation_qa, formal_fallacies, causal_judgement). For each, we extracted the correct option to construct QA pairs grounded in narrative context. As SimpleToM and SocialIQa answers were difficult to connect to their questions, we appended them to the question text. We sampled 200 training and 100 test examples per parameter, excluding four low-frequency ones (C_acol, C_agov, C_anet, C_apol), yielding 10,800 training and 5,400 test instances for training and testing proposes.

Metrics Consistent with prior QA research like(Deutsch, Bedrax-Weiss, and Roth 2021) and (Su et al. 2019), we use F1 score to evaluate. For OOD benchmarks, we adopt F1-based accuracy: each output is matched to the option with highest F1. If multiple options tie, we return E ("Not Applicable"); accuracy is computed on these selections.

Implementation Details We conduct three experiments. First, we train models on our dataset using two training paradigms (Table 1, Figure 4) and compare their performance against two comparison datasets, LlamaQA and MistralQA. Second, we train Qwen models on StimuliQA with varied reward functions and decoding strategies, evaluated on the held-out test set (Table 2). Third, we label logical reasoning samples on StimuliQA and use them to SFT-train Qwen models, thereby injecting knowledge without changing the generation format (Mecklenburg et al. 2024). As Reinforcement Learning from Human Feedback (RLHF) primarily re-ranks existing knowledge (Christiano et al. 2017; Yue et al. 2025; Li et al. 2025), we apply our RL framework to the SFT model to produce **Psy-Interpreter**, evaluated on five OOD datasets and further tested with a continual learning module. Full details are provided in the Appendix.

Training Sets	ToMbench		SimpleTom		SocialIQa		CosmosQA		BIG-Bench Hard	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Qwen2.5-0.5B-Instruct										
LlamaQA	11.95	26.89%	15.37	32.03%	15.55	30.60%	12.29	28.21%	11.06	40.47%
MistralQA	10.01	21.33%	20.43	26.88%	22.79	32.70%	10.30	24.29%	14.45	39.74%
StimuliQA	**16.61**	**34.13%**	**22.32**	**53.59%**	**25.37**	**40.74%**	**12.82**	**30.08%**	**16.23**	**48.91%**
Qwen2.5-1.5B-Instruct										
LlamaQA	12.14	27.52%	20.03	34.93%	21.80	40.28%	12.43	28.74%	12.39	44.10%
MistralQA	10.27	22.62%	25.02	33.74%	27.31	37.87%	11.31	26.33%	14.50	39.88%
StimuliQA	**18.26**	**37.94%**	**36.21**	**49.84%**	**27.32**	**41.91%**	**14.23**	**30.92%**	**23.51**	**56.04%**
Qwen2.5-3B-Instruct										
LlamaQA	15.91	32.69%	18.48	26.16%	34.62	49.59%	15.11	31.96%	13.63	44.98%
MistralQA	14.88	30.70%	35.02	33.33%	35.28	40.84%	12.69	27.34%	19.33	41.05%
StimuliQA	**22.06**	**42.38%**	**37.62**	**56.44%**	**43.40**	**53.74%**	**15.99**	**33.17%**	**19.74**	**54.88%**

Table 1: Comparison of SFT training on StimuliQA and two other datasets by both accuracy and F1 score.

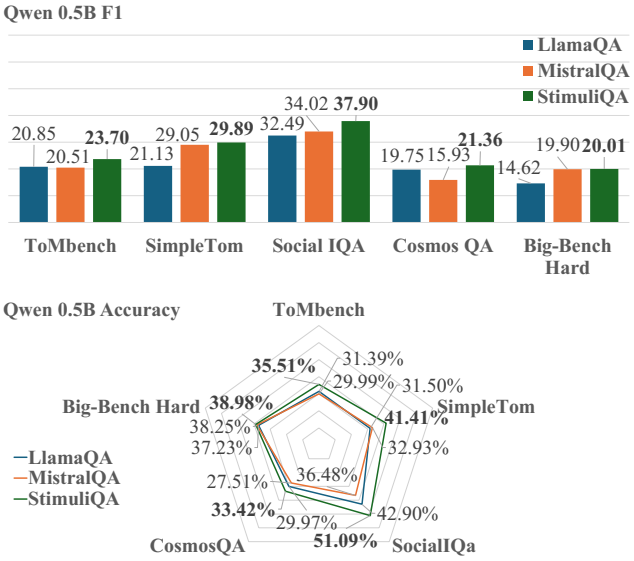

Figure 4: Comparison of GRPO training on StimuliQA and two other training datasets on Qwen0.5B. Full table with Qwen1.5B and 3B's results is located in the Appendix.

4.2 RQ1: Effect of Human-Labeled Data

One of our core hypotheses is that human-labeled data—especially from trained psychology students—can inject high-quality domain knowledge into compact models during SFT. We test this by comparing Qwen2.5-Instruct models (0.5B, 1.5B, 3B) trained on our **StimuliQA** dataset with those trained on synthetic data generated by Llama 3.3-70B and Mistral 8×7B. All models are evaluated on five OOD benchmarks: **ToMbench**, **SimpleToM**, **SocialIQa**, **CosmosQA**, and **BIG-Bench Hard**, with results summarized in Table 1 and Figure 4. This setup allows us to examine performance gaps across model sizes and tasks.

As shown in Table 1, models trained on **StimuliQA** outperform those trained on synthetic data across all model sizes and benchmarks, with especially strong gains in the 3B setting. For instance, on **SimpleToM**, Qwen2.5-3B achieves 37.62 F1 and 56.44% accuracy, far exceeding Llama 3.3 (18.48/26.16%) and Mistral 8×7B (35.02/33.33%). These substantial margins further reinforce our claim that: human-labeled data conveys more nuanced and structured psychological knowledge than synthetic outputs. The consistent OOD improvements indicate this knowledge is indeed generalizable, highlighting the professional quality and domain relevance of our dataset. Annotations by trained psychology students provide valuable domain expertise essential for effective model training. When used in GRPO training (Figure 4), human-labeled data still offers an advantage.

4.3 RQ2: Impact of Reward Design

Our results in Table 2 show that psychologically grounded reward functions substantially enhance model performance, and Figure 5 also shows a training accuracy boost. In particular, the BR reward, which integrates structure, emotion sensitivity, and length normalization, consistently outperforms baselines like Basic R1 or single-aspect rewards. For instance, BR raises overall F1 on Qwen2.5-1.5B from 34.46 to 39.98, with +5.66 and +5.71 F1 gains in *Needs Fulfilment* and *National Theme*, respectively. These gains suggest BR promotes not only token-level accuracy but also deeper emotional and moral reasoning. Beyond absolute improvements, BR yields more balanced results across all psychological dimensions, unlike length-based or repetition-penalized rewards, which often improve certain aspects at the cost of degrading others. This highlights BR's ability to align model behavior with complex, multi-faceted psychological goals.

Moreover, BR scales well across model sizes, offering +5.34 F1 on Qwen-3B and +5.74 on Qwen-0.5B, making it suitable for compute constrained settings. BR-trained models generate more coherent moral reasoning, richer self-reflection, and clearer emotional distinctions indicating a shift toward more theory-aligned internal representations.

Model	Specific Emotions	Targeted Emotions	Direction of Change	Self Identification	Needs Fulfilment	National Theme	Overall
			Qwen2.5-0.5B-Instruct				
Naïve	10.72	9.20	14.60	13.78	10.58	13.76	11.46
CoT	10.95(+0.23)	10.43(+1.23)	14.51(-0.09)	15.30(+1.52)	11.42(+0.84)	14.11(+0.35)	12.31(+0.85)
SFT	**14.08(+3.36)**	**19.70(+10.5)**	**29.27(+14.67)**	**21.29(+7.51)**	**22.07(+11.49)**	**29.79(+16.03)**	**21.70(+10.24)**
Basic R1	25.25	27.11	27.91	32.78	32.59	30.62	29.23
Length Reward	25.49(+0.24)	26.83(-0.28)	28.40(+0.49)	32.80(+0.02)	31.86(-0.73)	27.81(-2.81)	28.63(-0.60)
Short Reward	22.61(-2.64)	28.90(+1.79)	27.82(-0.09)	34.52(+1.74)	33.80(+1.21)	28.50(-2.12)	29.49(+0.26)
Length w RP	26.70(+1.45)	28.49(+1.38)	29.85(+1.94)	34.59(+1.81)	33.67(+1.08)	30.18(-0.44)	30.37(+1.14)
BR	**29.02(+3.77)**	**31.20(+4.09)**	**38.70(+10.79)**	**37.84(+5.06)**	**39.97(+7.38)**	**38.74(+8.12)**	**34.97(+5.74)**
			Qwen2.5-1.5B-Instruct				
Naïve	11.88	10.68	18.06	15.64	13.13	16.67	13.46
CoT	14.87(+2.99)	15.07(+4.39)	22.59(+4.53)	19.09(+3.45)	17.68(+4.55)	20.18(+3.51)	17.37(+3.91)
SFT	**19.13(+7.25)**	**21.48(+10.8)**	**33.00(+14.94)**	**31.32(+15.68)**	**29.43(+16.30)**	**31.58(+14.91)**	**26.27(+12.81)**
Basic R1	34.29	30.57	37.03	35.28	38.22	36.81	34.46
Length Reward	33.26(-1.03)	30.91(+0.34)	35.94(-1.09)	35.89(+0.61)	38.00(-0.22)	36.75(-0.06)	34.41(-0.05)
Short Reward	34.81(+0.52)	30.56(-0.01)	36.92(-0.11)	37.00(+1.72)	38.78(+0.56)	38.23(+1.42)	35.14(+0.68)
Length w RP	34.90(+0.61)	32.38(+1.81)	37.05(+0.02)	38.01(+2.73)	40.54(+2.32)	39.25(+2.44)	36.32(+1.86)
BR	**38.45(+4.16)**	**36.00(+5.43)**	**41.95(+4.92)**	**42.09(+6.81)**	**43.88(+5.66)**	**42.52(+5.71)**	**39.98(+5.52)**
			Qwen2.5-3B-Instruct				
Naïve	13.44	13.27	19.97	16.20	15.19	18.68	15.34
CoT	13.86(+0.42)	15.11(+1.84)	21.60(+1.63)	17.65(+1.45)	16.27(+1.08)	20.03(+1.35)	16.70(+1.36)
SFT	**22.45(+9.01)**	**22.04(+8.77)**	**34.47(+14.50)**	**31.68(+15.48)**	**27.24(+10.97)**	**32.73(+14.05)**	**26.95(+11.61)**
Basic R1	34.17	31.23	36.81	36.04	37.98	38.71	35.04
Length Reward	33.02(-1.15)	31.56(+0.33)	37.90(+1.09)	36.77(+0.73)	38.78(+0.80)	37.86(-0.85)	35.14(+0.10)
Short Reward	35.46(+1.29)	31.28(+0.05)	38.02(+1.21)	37.84(+1.80)	40.38(+2.40)	37.72(-0.99)	35.81(+0.77)
Length w RP	35.50(+1.33)	33.56(+2.33)	39.88(+3.07)	39.02(+2.98)	41.79(+3.81)	39.80(+1.09)	37.39(+2.35)
BR	**39.17(+5.00)**	**36.25(+5.02)**	**41.87(+5.06)**	**42.03(+5.99)**	**44.44(+6.46)**	**43.38(+4.67)**	**40.38(+5.34)**

Table 2: Comparison of different training and generation methods on our StimuliQA dataset across six psychological dimensions from our StimuliQA. Our BR reward consistently achieves the best overall performance across all six domains.

4.4 RQ3: Generalization of Psy-Interpreter

As shown in Table 3, **Psy-Interpreter** demonstrates strong zero-shot generalization across five diverse OOD benchmarks after injecting knowledge via SFT and GRPO training with our **StimuliQA** dataset. Despite being trained without any direct supervision on these test sets, it consistently outperforms other baselines such as Simple RL Zoo, NuExtract, and DeepSeek-R1 Distilled across all model sizes. Notably, the 3B variant achieves 28.17 F1 on **ToMbench**, 56.83 on **SimpleToM**, and 65.54 on **SocialIQa**, marking substantial gains over both in-house and open-source compact models. These results suggest that Psy-Interpreter successfully internalizes psychologically grounded knowledge during training, enabling it to reason beyond its original domain.

4.5 RQ4: Effectiveness of Continual Learning

Table 3 shows that our continual learning framework (**Psy-Interpreter-SFT**) achieves consistent improvements across all five OOD psychological benchmarks. For the 0.5B model, accuracy rises from 40.66% to **58.82%** on ToMbench, 51.30% to **74.04%** on SocialIQa, and 38.69% to **43.45%** on CosmosQA. F1 scores increase more sharply, 25.27 to **62.28** on ToMbench and 23.05 to **62.41** on BIG-Bench Hard, indicating better correctness and closer alignment with human-preferred reasoning. The 1.5B and 3B variants maintain these gains, reaching up to 82.82 F1 on SocialIQa and consistently surpassing their non-SFT counterparts. Despite their compact size, our models rival or exceed larger commercial LLMs: **Psy-Interpreter-SFT (3B)** outperforms GPT-4 nano (57.03) and Claude 3 Haiku (15.94).

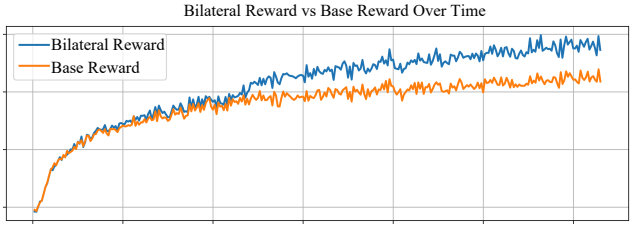

Figure 5: Base and Bilateral reward Training comparison.

Despite their smaller scale, our models often match or notably surpass larger commercial LLMs. **Psy-Interpreter-SFT (3B)** achieves an impressive F1 of 82.82 on SocialIQa, exceeding GPT-4 nano (57.03) and Claude 3 Haiku (15.94), and outperforming DeepSeek Reasoner on most tasks. Moreover, this performance gain is consistent across ToM reasoning, and narrative-understanding benchmarks.

Models	ToMbench		SimpleTom		SocialIQa		CosmosQA		BIG-Bench Hard	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
GPT-4 nano	22.11	46.22%	67.03	54.34%	57.03	68.03%	**15.82**	35.00%	9.30	41.75%
Claude 3 Haiku	18.81	49.65%	19.32	37.31%	15.94	52.64%	11.28	33.27%	7.07	40.15%
DeepSeek Reasoner	26.63	**56.02%**	**78.98**	**69.17%**	**63.93**	**70.36%**	13.13	35.23%	**40.52**	**69.93%**
Llama 3.3 70B Instruct	**27.42**	52.24%	76.43	61.46%	60.91	69.07%	15.27	**37.19%**	17.61	69.34%
Mistral 8×7B Instruct	12.03	34.46%	12.06	42.52%	12.14	42.23%	7.49	24.66%	5.39	26.42%
Qwen2.5-0.5B-Instruct										
Simple RL Zoo	13.38	14.42%	12.89	6.63%	16.56	12.23%	10.33	5.22%	9.12	6.86%
NuExtract	12.69	7.21%	12.20	1.39%	14.44	2.54%	8.02	2.50%	6.83	2.04%
New Merges Serialization	15.21	23.27%	17.30	13.86%	15.14	22.69%	9.80	22.69%	8.04	6.13%
Psy-Interpreter	25.27	40.66%	37.63	65.10%	43.01	51.30%	23.58	38.69%	23.05	48.76%
Psy-Interpreter-SFT	**62.28**	**58.82%**	**85.25**	**65.45%**	**76.58**	**74.04%**	**34.43**	**43.45%**	**62.41**	**56.06%**
Qwen2.5-1.5B-Instruct										
Simple RL Zoo	3.36	4.51%	9.02	6.51%	4.33	7.31%	2.91	2.50%	3.98	8.03%
Nemotron Reasoning	2.02	4.65%	6.38	6.10%	7.30	10.21%	1.69	5.26%	6.98	9.78%
DeepSeek-R1 Distilled	11.39	22.64%	26.80	26.68%	9.25	22.49%	7.33	12.14%	2.73	13.28%
Psy-Interpreter	26.54	49.30%	39.02	59.31%	56.40	63.94%	23.61	41.34%	22.33	50.07%
Psy-Interpreter-SFT	**65.94**	**62.25%**	**84.95**	**66.26%**	**80.55**	**78.86%**	**36.37**	**45.68%**	**64.04**	**58.69%**
Qwen2.5-3B-Instruct										
Transformer RL	16.25	18.86%	15.55	9.56%	25.20	39.33%	11.14	6.53%	9.33	5.26%
Tulu Trained	21.54	22.67%	33.83	16.97%	46.24	39.12%	14.44	24.28%	34.92	29.34%
Raspberry	15.46	15.96%	33.92	24.35%	28.93	26.06%	10.36	3.65%	13.23	6.57%
Psy-Interpreter	28.17	51.33%	56.83	58.73%	65.54	71.09%	25.60	44.33%	18.87	56.06%
Psy-Interpreter-SFT	**66.33**	**58.36%**	**84.98**	**67.92%**	**82.82**	**81.24%**	**36.88**	**47.56%**	**65.86**	**60.88%**

Table 3: Performance of our Psy-Interpreter and continual learning with SFT on all the OOD datasets, compared against strong commercial models and open-source baselines, with consistent gains across all five datasets.

4.6 RQ5: Length Distribution Comparison

Figure 6 illustrates the reasoning length distributions under the bilateral and base reward settings on the *Unexpected Outcome Test* from ToMbench, which comprises 300 samples with an intuitive question-difficulty hierarchy. We only include samples with non-zero F1 scores, as these responses provide reasoning chains that are more representative and interpretable than those from entirely incorrect outputs. Index 1 corresponds to simple stories with simple questions, Index 2 to the same stories with harder questions, and Index 3 to more complex stories with direct questions. The figure also reports the total number of valid samples under each reward configuration. As shown, the BR yields a more interpretable distribution: Index 1 responses remain concise, while Index 2 and Index 3 exhibit longer and more differentiated reasoning chains. In contrast, the base reward produces a noisier distribution with reduced separation between simple and complex QA pairs, occasionally generating overly long reasoning for direct emotion predictions. This indicates that our BR better aligns reasoning length with task difficulty, enhancing overall interpretability and efficiency. Moreover, the consistent accuracy gains in Tables 2 and 3 underscore the importance of carefully controlling reasoning length in complex psychological reasoning. In addition, our reward method was able to generate more correct samples at each difficulty level, yielding 12 additional samples for simple emotion prediction and 19 and 8 more for the harder ones, respectively, validating its effectiveness.

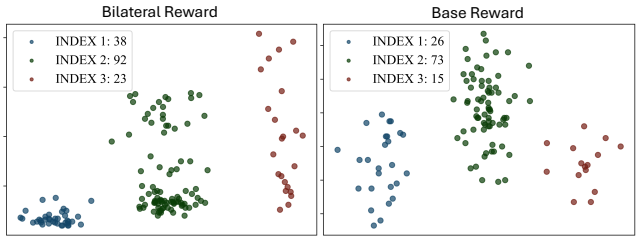

Figure 6: The figure shows the distribution of reasoning lengths under the bilateral reward and the base reward.

Figure 1 shows a representative reasoning sample, illustrating how the model iteratively refines its thought patterns, additional examples are provided in the appendix.

5 Conclusion

We first introduce the annotated dataset StimuliQA and Psy-Interpreter, an RL framework grounded in psychological theory that enhances emotional and cognitive reasoning in LLMs. Leveraging high-quality labeled data with structured RL substantially improves social-cognitive reasoning in compact models, yielding consistent gains across five diverse OOD benchmarks. Our Bilateral Reward aligns reasoning length with task complexity and improves accuracy, while Psy-Interpreter-SFT with continual learning module effectively narrows the gap with larger commercial LLMs.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ayers, J. W.; Poliak, A.; Dredze, M.; Leas, E. C.; Zhu, Z.; Kelley, J. B.; Faix, D. J.; Goodman, A. M.; Longhurst, C. A.; Hogarth, M.; and Smith, D. M. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 183(6): 589–596.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; et al. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15959–15983.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Coda-Forno, J.; Binz, M.; Wang, J. X.; and Schulz, E. 2024. CogBench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, 9076–9108.

Deutsch, D.; Bedrax-Weiss, T.; and Roth, D. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9: 774–789.

Epstein, S. 1998. Cognitive-experiential self-theory. In *Advanced personality*, 211–238. Springer.

Feng, L.; Xue, Z.; Liu, T.; and An, B. 2025a. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*.

Feng, Y.; Zhao, S.; Li, Y.; Xiao, L.; Wu, X.; and Luu, A. T. 2025b. Aspect-Based Summarization with Self-Aspect Retrieval Enhanced Generation. *arXiv preprint arXiv:2504.13054*.

Gu, Y.; Tafjord, O.; Kim, H.; Moore, J.; Bras, R. L.; Clark, P.; and Choi, Y. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Han, S.-T. 2025. Narrative-Centered Emotional Reflection: Scaffolding Autonomous Emotional Literacy with AI. *arXiv preprint arXiv:2504.20342*.

Hu, J.; Dong, T.; Gang, L.; Ma, H.; Zou, P.; Sun, X.; Guo, D.; Yang, X.; and Wang, M. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.

Huang, J.-t.; Wang, W.; Li, E. J.; Lam, M. H.; Ren, S.; Yuan, Y.; Jiao, W.; Tu, Z.; and Lyu, M. R. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2391–2401. Hong Kong, China: Association for Computational Linguistics.

Ji, Y.; Zhao, S.; Tian, X.; Wang, H.; Chen, S.; Peng, Y.; Zhao, H.; and Li, X. 2025. How Difficulty-Aware Staged Reinforcement Learning Enhances LLMs' Reasoning Capabilities: A Preliminary Experimental Study. *arXiv preprint arXiv:2504.00829*.

Ke, L.; Tong, S.; Cheng, P.; and Peng, K. 2025. Exploring the frontiers of llms in psychological applications: A comprehensive review. *Artificial Intelligence Review*, 58(10): 305.

Kovacevic, N.; Holz, C.; Gross, M.; and Wampfler, R. 2024. On Multimodal Emotion Recognition for Human-Chatbot Interaction in the Wild. In *Proceedings of the 26th International Conference on Multimodal Interaction*, 12–21.

Lazarus, R. S. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8): 819.

Leng, Y.; and Yuan, Y. 2023. Do LLM Agents Exhibit Social Behavior? *arXiv preprint arXiv:2312.15198*.

Li, Y.; Huang, Y.; Wang, H.; Zhang, X.; Zou, J.; and Sun, L. 2024. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*.

Li, Y.-C.; Xu, T.; Yu, Y.; Zhang, X.; Chen, X.-H.; Ling, Z.; Chao, N.; Yuan, L.; and Zhou, Z.-H. 2025. Generalist Reward Models: Found Inside Large Language Models. *arXiv preprint arXiv:2506.23235*.

Liu, W.; Luo, H.; Lin, X.; Liu, H.; Shen, T.; Wang, J.; Mao, R.; and Cambria, E. 2025. Prompt-R1: Collaborative Automatic Prompting Framework via End-to-end Reinforcement Learning. *arXiv preprint arXiv:2511.01016*.

Long, L.; Wang, R.; Xiao, R.; Zhao, J.; Ding, X.; Chen, G.; and Wang, H. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11065–11082. Bangkok, Thailand: Association for Computational Linguistics.

Luo, H.; E, H.; Chen, G.; Lin, Q.; Guo, Y.; Xu, F.; Kuang, Z.; Song, M.; Wu, X.; Zhu, Y.; and Tuan, L. A. 2025a. Graph-R1: Towards Agentic GraphRAG Framework via End-to-end Reinforcement Learning. arXiv:2507.21892.

Luo, H.; E, H.; Chen, G.; Zheng, Y.; Wu, X.; Guo, Y.; Lin, Q.; Feng, Y.; Kuang, Z.; Song, M.; Zhu, Y.; and Tuan, L. A. 2025b. HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. arXiv:2503.21322.

Luo, H.; E, H.; Guo, Y.; Lin, Q.; Wu, X.; Mu, X.; Liu, W.; Song, M.; Zhu, Y.; and Tuan, L. A. 2025c. KBQA-o1: Agentic Knowledge Base Question Answering with Monte Carlo Tree Search. arXiv:2501.18922.

McAdams, D. P.; Reynolds, J.; Lewis, M.; Patten, A. H.; and Bowman, P. J. 2001. When bad things turn good and good things turn bad: Sequences of redemption and contamination in life narrative and their relation to psychosocial adaptation in midlife adults and in students. *Personality and social psychology bulletin*, 27(4): 474–485.

Mecklenburg, N.; Lin, Y.; Li, X.; Holstein, D.; Nunes, L.; Malvar, S.; Silva, B.; Chandra, R.; Aski, V.; Yannam, P. K. R.; et al. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *arXiv preprint arXiv:2404.00213*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: NeurIPS 2022 (Conference Track)*.

Qiu, H.; He, H.; Zhang, S.; Li, A.; and Lan, Z. 2024. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 615–636. Miami, Florida, USA: Association for Computational Linguistics.

Reyna, V. F.; and Brainerd, C. J. 1998. Fuzzy-trace theory and false memory: New frontiers. *Journal of experimental child psychology*, 71(2): 194–209.

Ryff, C. D.; and Keyes, C. L. M. 1995. The structure of psychological well-being revisited. *Journal of personality and social psychology*, 69(4): 719.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019a. Social IQa: Commonsense Reasoning about Social Interactions. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4463–4473. Hong Kong, China: Association for Computational Linguistics.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, T.; Mao, R.; Wang, J.; Zhang, X.; and Cambria, E. 2025. Flow-guided direct preference optimization for knowledge graph reasoning with trees. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1165–1175.

Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; and Klang, E. 2024. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research*, 26: e52597.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

Su, D.; Xu, Y.; Winata, G. I.; Xu, P.; Kim, H.; Liu, Z.; and Fung, P. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd workshop on machine reading for question answering*, 203–211.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; and Wei, J. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13003–13051. Toronto, Canada: Association for Computational Linguistics.

Wang, X.; Li, C.; Chang, Y.; Wang, J.; and Wu, Y. 2024. NegativePrompt: Leveraging Psychology for Large Language Models Enhancement via Negative Emotional Stimuli. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6504–6512. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wu, X. 2025. Sailing by the Stars: A Survey on Reward Models and Learning Strategies for Learning from Rewards. *arXiv preprint arXiv:2505.02686*.

Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10691–10706. Singapore: Association for Computational Linguistics.

Xiao, Y.; Wang, J.; Xu, Q.; Song, C.; Xu, C.; Cheng, Y.; Li, W.; and Liu, P. 2025. Towards Dynamic Theory of Mind: Evaluating LLM Adaptation to Temporal Evolution of Human States. *arXiv preprint arXiv:2505.17663*.

Xie, H.; Chen, Y.; Xing, X.; Lin, J.; and Xu, X. 2024. PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling. *arXiv preprint arXiv:2412.13660*.

Yang, Q.; Ye, M.; and Du, B. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Yue, Y.; Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Song, S.; and Huang, G. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Zhang, J.; Feng, L.; Guo, X.; Wu, Y.; Dong, Y.; and Xu, D. 2025. TimeMaster: Training Time-Series Multimodal LLMs to Reason via Reinforcement Learning. *arXiv preprint arXiv:2506.13705*.