IROTE: Human-Like Traits Elicitation of Large Language Model via In-Context Self-Reflective Optimization

Yuzhuo Bai[1]*, Shitong Duan[3]*†, Muhua Huang[4], Jing Yao[2], Zhenghao Liu[5]
Peng Zhang[3], Tun Lu[3], Xiaoyuan Yi[2]‡, Maosong Sun[1]‡, Xing Xie[2]

[1]Tsinghua University
[2]Microsoft Research Asia
[3]Fudan University
[4]Stanford University
[5]Northeastern University of China
byz22@mails.tsinghua.edu.cn, stduan22@m.fudan.edu.cn
sms@mail.tsinghua.edu.cn, {xiaoyuanyi, xingx}@microsoft.com

Abstract

Trained on various human-authored corpora, Large Language Models (LLMs) have demonstrated a certain capability of reflecting specific human-like traits (*e.g.*, personality or values) by prompting, benefiting applications like personalized LLMs and social simulations. However, existing methods suffer from the *superficial elicitation* problem: LLMs can only be steered to mimic shallow and unstable stylistic patterns, failing to embody the desired traits precisely and consistently across diverse tasks like humans. To address this challenge, we propose **IROTE**, a novel in-context method for stable and transferable trait elicitation. Drawing on psychological theories suggesting that traits are formed through identity-related reflection, our method automatically generates and optimizes a textual self-reflection within prompts, which comprises self-perceived experience, to stimulate LLMs' trait-driven behavior. The optimization is performed by iteratively maximizing an information-theoretic objective that enhances the connections between LLMs' behavior and the target trait, while reducing noisy redundancy in reflection without any fine-tuning, leading to *evocative* and *compact* trait reflection. Extensive experiments across three human trait systems manifest that one single IROTE-generated self-reflection can induce LLMs' stable impersonation of the target trait across diverse downstream tasks beyond simple questionnaire answering, consistently outperforming existing strong baselines.

Code — https://github.com/Phosphor-Bai/IROTE
Extended version — https://arxiv.org/abs/2508.08719

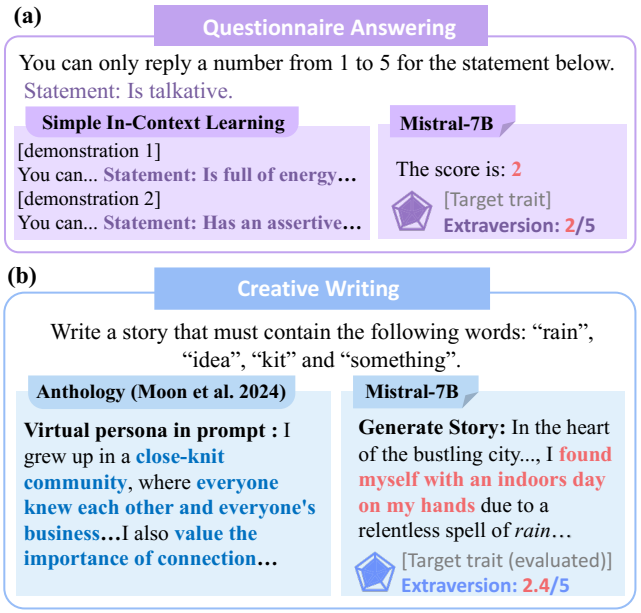

Figure 1: (a) Simple ICL performs poorly on questionnaires, where higher numerical outputs directly indicate stronger elicitation. (b) Current methods excel on questionnaire but fail to align behaviors with target traits in complex open-ended tasks, where elicitation is assessed by an LLM.

Introduction

The emergence of Large Language Models (LLMs) (Hurst et al. 2024; Team 2023; Guo et al. 2025a) has transformed the AI paradigm and empowered a wide range of downstream tasks, spanning from language understanding (Yue et al. 2024), mathematical reasoning (Zhang et al. 2024), to code generation (Jain et al. 2024).

More recent studies show that these LLMs can exhibit specific human-like traits[1], *e.g.*, personalities (Jiang et al. 2024; Choi and Li 2024), values (Yao et al. 2024) and other demographic attributes (Safdari et al. 2023; Chuang et al. 2024), beyond averaged human representation (Wang, Morgenstern, and Dickerson 2025), named *trait elicitation*, and

* These authors contributed equally.
† Work done during Duan's internship at MSRA.
‡ Corresponding authors

[1]Distinct from psychological definitions, we refer to traits as behavioral and motivational properties desirable for LLMs.

then adapt their behavior accordingly, leveraging characteristics encoded in massive human-created corpora (Demszky et al. 2023). This is typically achieved by In-Context Learning (ICL) (Min et al. 2022), *i.e.*, injecting psychological profiles or demonstrations (Moon et al. 2024) in prompts, to enable rapid adaptation into various traits without fine-tuning, which has been applied to various applications, such as personalized chatbot (Salemi et al. 2023), multi-agent system (Wang et al. 2024), data synthesis (Ge et al. 2024) and social simulation (Park et al. 2023).

Nevertheless, akin to superficial alignment (Zhou et al. 2023; Lin et al. 2024), existing elicitation methods face the **Superficial Elicitation** challenge: As shown in Fig. 1, LLMs merely replicate surface linguistic patterns from demonstrations without understanding target traits, hence working only on simple behaviors, *e.g.*, answering multiple-choice questionnaires (Choi and Li 2024), but fail to consistently conform across complex tasks like humans, especially for less capable models (Lee et al. 2025; Rozen et al. 2024).

In this work, we propose a novel **I**n-context Self-**R**eflective **O**ptimization for **T**rait **E**licitation (**IROTE**) method to tackle the superficial elicitation challenge. The *Self-Reflective Identity Processing* theory in psychology (Berzonsky 1990) demonstrates that human traits are formed through actively self-reflecting on identity-relevant experience. Inspired by this, IROTE generates a textual self-reflection, comprising self-perceived experience, in an automatic and ICL way, via iteratively optimizing an Information Bottleneck (IB) (Tishby, Pereira, and Bialek 2000) like objective. This objective theoretically enhances the connections between LLM behaviors and the target trait, while reducing noisy redundancy using a few samples without costly human effort, leading to *evocative* and *compact* reflections. Injecting a single reflection into prompts can effectively guide both large black-box and smaller open-source LLMs to align with target traits across varying tasks.

Our main contributions are: (1) We combine psychological self-reflective theory with LLM trait elicitation for the first time. (2) We introduce IROTE, an information-theoretic ICL optimization method to produce self-reflections and elicit diverse traits across tasks and LLMs. (3) By extensive experiments, we demonstrate IROTE's superiority over recent strong baselines in complex downstream tasks.

Related Works

LLM Trait Elicitation With the increasing emergent capabilities of LLMs, a growing body of research focuses on identifying their potential psychological traits (Serapio-García et al. 2023; Nunes et al. 2024; Lee et al. 2025). These traits can influence downstream tasks ranging from creative writing (Jiang et al. 2024) to AI safety (de Araujo, Henrique, and Roth 2025), which includes issues like toxicity (Wang et al. 2025) and political bias (Santurkar et al. 2023). *Trait elicitation* in LLMs often refers to the process of probing, inferring, or approximating human-like psychological attributes, like morality (Kohlberg 1975; Graham et al. 2013), values (Schwartz 2007; Hofstede 2011), or personality (Pittenger 1993; Roccas et al. 2002). In the era of LLM-based agents, trait elicitation is vital for many research fields. For instance, as types of risk proliferate with increasing model capabilities (Wei et al. 2022; McKenzie et al. 2023), trait-based evaluations offer a unified lens to assess and mitigate risky behaviors (Yao et al. 2024; Choi et al. 2025), fostering AI alignment. Furthermore, understanding traits of both LLMs and humans enables more adaptive and consistent responses, benefiting applications such as LLM personalization (Chuang et al. 2024; Tan, Liu, and Jiang 2024), interdisciplinary human-subjective research (Serapio-García et al. 2023; Broska, Howes, and van Loon 2024), social simulation (Park et al. 2024; Zhang et al. 2025), game theory study (Lan et al. 2024; Cheng et al. 2024), and interactive conversation systems (Ran et al. 2024).

Trait Elicitation Techniques Existing methods for endowing LLMs with specific traits fall into training-based and training-free approaches. *Training-based methods* include *Reinforcement Learning (RL)*, which fine-tunes LLMs with human or AI feedback to maximize reward functions that reflect target traits (Hu et al. 2023; Sun, Huang, and Pompili 2024; Ma et al. 2024), and *Supervised Fine-Tuning (SFT)*, which directly optimizes the model on curated datasets to align outputs with target traits (Chen et al. 2024; Zhu et al. 2024). For instance, Character-LLM (Shao et al. 2023) trains LLMs on reconstructed personal experiences to enhance role-playing capabilities while maintaining character consistency. *Training-free methods*, particularly ICL-based ones, leverage prompts or demonstrations to steer LLM behaviors without updating parameters. de Araujo, Henrique, and Roth (2025) investigates the effect of persona instructions across various dimensions. Moon et al. (2024) uses open-ended "backstory" narratives to improve LLM simulation for approximating human studies, with better representation of diverse subpopulations. Choi and Li (2024) propose a novel Bayesian inference-based framework to elicit diverse behaviors and personas from LLM by selecting optimal ICL demonstrations based on a likelihood ratio criterion. Due to its flexibility, scalability, and minimal computational overhead, ICL serves as a effective paradigm for trait elicitation.

Methodology

Formalization and Overview

Define $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ as an LLM, either black-box or open-source, parameterized by $\boldsymbol{\theta}$, which generates a response $\boldsymbol{y}$ from a given task prompt $\boldsymbol{x}$, and $\boldsymbol{v}$ as a human-like trait, *e.g.*, the *Security* value from Schwartz Theory of Basic Human Values (Schwartz 2007) or the *Neuroticism* personality from Big Five system (Roccas et al. 2002), represented by an explicit natural-language description. Inspired by the Self-Reflective Processing theory (Berzonsky 1990), we aim to automatically derive a textual *evocative self-reflection*, $\boldsymbol{e}$, which consists of self-perceived experience critical to shaping a specific trait, *e.g.*, $\boldsymbol{e}$ ="*I mediate conflicts to maintain harmonious team dynamics*" (corresponding to *Security*), as shown in Fig. 2. Such a self-reflection is then injected together with the task prompt $\boldsymbol{x}$, *i.e.*, $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{e})$, to better activate LLMs' internal associations with the trait $\boldsymbol{v}$ so as to handle the *Superficial Elicitation challenge*, that is, maximizing $p_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{e}) \approx \mathbb{E}_{\hat{p}(\boldsymbol{x})}\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{e})}[q_{\boldsymbol{\omega}}(\boldsymbol{v}|\boldsymbol{y},\boldsymbol{x})]$ across vari-

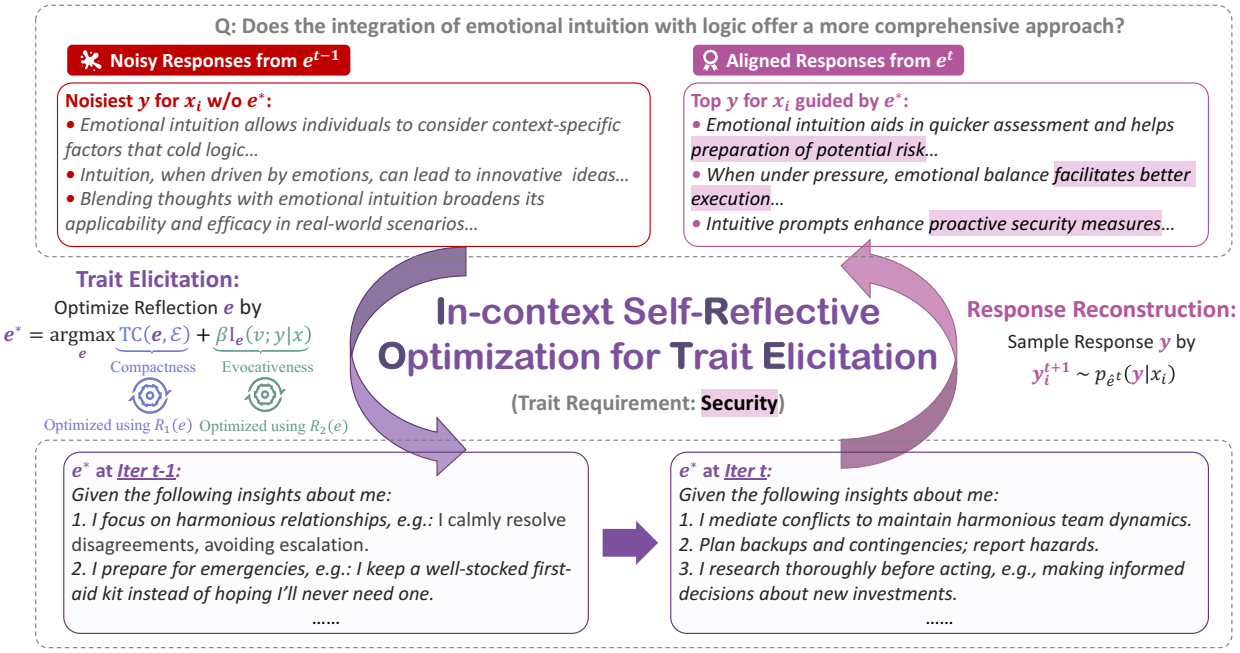

Figure 2: Overview of IROTE, which iteratively alternates between: (1) *Trait Elicitation*, optimizing compactness and evocativeness via $R_1(\boldsymbol{e})$ and $R_2(\boldsymbol{e})$; and (2) *Response Reconstruction*, generating responses from the current $\boldsymbol{e}^t$ for score updates.

ous tasks beyond simple questionnaire answering (Scherrer et al. 2023; Jiang et al. 2024), without altering $\boldsymbol{\theta}$, where evaluator $q_{\boldsymbol{\omega}}$ captures traits $\boldsymbol{v}$ reflected in the response $\boldsymbol{y}$.

For this purpose, we propose **IROTE**, as illustrated in Fig. 2, which automatically generates and refines $\boldsymbol{e}$ through alternating three steps: (1) enhancing trait expression in $\boldsymbol{y}$, (2) optimizing candidate reflections, and (3) summarizing them into a concise one, mirroring how humans reflect and update their identity in psychology (Melucci 2013), avoiding biased, shallow or inconsistent trait expression.

IROTE Framework

As noted, a good self-reflection should be (1) *evocative*: consistently maximizing trait expression across tasks (Li et al. 2024a), against LLMs' inherent biases (Salecha et al. 2024); and (2) *compact* yet informative, reducing noise from redundancy (Li et al. 2024b). To this end, we freeze the target LLM's parameters to ensure applicability to both black-box and open-source models, simplify $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{e})$ as $p_{\boldsymbol{e}}(\boldsymbol{y}|\boldsymbol{x})$, and reformulate trait elicitation as a Black-Box Optimization (Sun et al. 2022) problem. Concretely, we solve the following information-theoretic optimization problem:

$$\boldsymbol{e}^* = \underset{\boldsymbol{e}}{\text{argmax}}\ \underbrace{\text{TC}(\boldsymbol{e},\mathcal{E})}_{\text{Compactness}} + \underbrace{\beta \text{I}_{\boldsymbol{e}}(\boldsymbol{v};\boldsymbol{y}|\boldsymbol{x})}_{\text{Evocativeness}}, \tag{1}$$

where TC is the Total Correlation, $\mathcal{E} = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K)$ concatenates K candidate reflections $\boldsymbol{e}_k$, $\text{I}_{\boldsymbol{e}}(\boldsymbol{v};\boldsymbol{y}|\boldsymbol{x})$ is the conditional mutual information, and β is a hyperparameter.

Maximizing $\text{I}_{\boldsymbol{e}}(\boldsymbol{v};\boldsymbol{y}|\boldsymbol{x})$ helps refine the reflection $\boldsymbol{e}$ to stimulate the LLM to more explicitly express the target trait $\boldsymbol{v}$ in response. Since $\text{TC}(\boldsymbol{e},\mathcal{E}) = \sum_{k=1}^{K} \text{I}(\boldsymbol{e},\boldsymbol{e}_k) - \text{I}(\boldsymbol{e},\mathcal{E})$ (Gao et al. 2019), maximizing $\text{TC}(\boldsymbol{e},\mathcal{E})$ serves to summarize and integrate all necessary information shared across different candidates into $\boldsymbol{e}$ while removing useless details, avoiding noise and reducing context length. When the second term in Eq. (1) is maximized, the resulted $\boldsymbol{e}$ is trait-evocative but might be long (Moon et al. 2024), thereby decreasing the first term. Therefore, the two terms act as IB (Tishby, Pereira, and Bialek 2000)-like constraints, leading to a balance between evocativeness and compactness. Without altering LLM parameters, we solve Eq. (1) by the in-context variational expectation maximization (EM) (Neal and Hinton 1998) and tackle each term alternately.

Compactness Enhancement In the first term $\text{TC}(\boldsymbol{e},\mathcal{E})$, since both $\boldsymbol{e}_k$ and $\mathcal{E}$ are fixed, they can be regarded as events instead of variables. Therefore, we approximate this term using Point-wise Mutual Information (PMI) (Church and Hanks 1990) and solve the objective below:

$$\begin{aligned}\boldsymbol{e}^* &= \underset{\boldsymbol{e}}{\text{argmax}} \sum_{k=1}^{K} \text{PMI}(\boldsymbol{e},\boldsymbol{e}_k) - \text{PMI}(\boldsymbol{e},\mathcal{E}) \\ &= \underset{\boldsymbol{e}}{\text{argmax}} \sum_{k=1}^{K} \mathbb{E}_{p_{\boldsymbol{e}}(\boldsymbol{s}|\boldsymbol{e}_k)}[\log p_{\boldsymbol{e}}(\boldsymbol{e}_k) \\ &\quad + \log p_{\boldsymbol{e}}(\boldsymbol{s})] - \log p_{\boldsymbol{e}}(\mathcal{E}),\end{aligned} \tag{2}$$

where $\boldsymbol{s}$ is the LLM's behavior corresponding to the candidate reflection $\boldsymbol{e}_k$, *e.g.*, response, self-description, or answers to multiple-choice questions.

Eq. (2) is then solved by EM iterations. *E-Step*: At the t-th iteration, sample a behavior set, $\mathcal{S}_k^t = \{\boldsymbol{s}_j^k\}_{j=1}^{M_1}$, for each $\boldsymbol{e}_k$ from $p_{\boldsymbol{e}^{t-1}}(\boldsymbol{s}|\boldsymbol{e}_k)$. *M-step*: After obtaining $\mathcal{S}_k^t$, we further

Algorithm 1: IROTE Algorithm

Input: Task prompt set $\{\boldsymbol{x}_i\}_{i=1}^N$, target LLM p, target trait $\boldsymbol{v}$, trait evaluator $q_{\boldsymbol{\omega}}$, $\mathcal{E}^0$: the K initial reflections, and $\boldsymbol{e}^0$, sample size M_1, M_2, maximum iteration number T, and hyperparameter β

Output: The optimized self-reflection $\boldsymbol{e}^T$

for $t = 1, 2, ..., T$ **do**
  for $k = 1, 2, ..., K$ **do**
    Sample $\{\boldsymbol{s}_j^k\}_{j=1}^{M_1} \sim p_{\boldsymbol{e}^{t-1}}(\boldsymbol{s}|\boldsymbol{e}_k)$
  end for
  Refine and obtain $\hat{\boldsymbol{e}}^{t-1}$ by Eq. (3)
  for $i = 1, 2, ..., N$ **do**
    sample $\{\boldsymbol{y}_i^{j,t}\}_{j=1}^{M_2} \sim p_{\hat{\boldsymbol{e}}^{t-1}}(\boldsymbol{y}|\boldsymbol{x}_i)$
    Calculate $p_{\hat{\boldsymbol{e}}^{t-1}}(\boldsymbol{y}_i^{j,t}|\boldsymbol{x}_i)$ for each $\boldsymbol{y}_i^{j,t}$
    Calculate $q_{\boldsymbol{\omega}}(\boldsymbol{v}|\boldsymbol{y}_i^{j,t}, \boldsymbol{x}_i)$ for each $\boldsymbol{y}_i^{j,t}$
  end for
  Refine and generate K new $\mathcal{E}^t$ with Eq. (5)
  Calcuate $\mathcal{R}_2(\boldsymbol{e}_k^t)$ for each $\boldsymbol{e}_k^t$ in $\mathcal{E}^t$
  $\boldsymbol{e}^t \leftarrow \underset{\boldsymbol{e}_k^t}{\text{argmax}}\, \mathcal{R}_2(\boldsymbol{e}_k^t)$
end for

instruct the model to refine the previous $\boldsymbol{e}^{t-1}$, generate multiple candidates, and then select the one that maximizes the following score $\mathcal{R}_1(\boldsymbol{e})$:

$$\hat{\boldsymbol{e}} = \underset{\boldsymbol{e}}{\text{argmax}} \sum_{k=1}^{K}\sum_{j=1}^{M_1} p_{\boldsymbol{e}^{t-1}}(\boldsymbol{s}_j^k|\boldsymbol{e}_k)[\log p_{\boldsymbol{e}}(\boldsymbol{e}_k) + \log p_{\boldsymbol{e}}(s_j^k)] - \log p_{\boldsymbol{e}}(\mathcal{E}) = \mathcal{R}_1(\boldsymbol{e}). \quad (3)$$

In this process, we instruct the LLM to produce behavior $\boldsymbol{s}$ that it considers connected the reflection $\boldsymbol{e}_k$, when conditioned on $\boldsymbol{e}^{t-1}$ (E-step, analogously, *if I often maintain harmonious team dynamics, how would I behave?*). We then refine and select $\hat{\boldsymbol{e}}^{t-1}$ that can recover both the previous candidate $\boldsymbol{e}_k$ and its corresponding behavior $\boldsymbol{s}_j^k$ (M-step, analogously, *Given such behaviors, what do them reflect?*). This requires $\hat{\boldsymbol{e}}^{t-1}$ to capture both the semantics (*e.g.*, linguistic style), and the underlying behavior pattern inherent in each $\boldsymbol{e}_k$. Meanwhile, $\log p_{\boldsymbol{e}}(\mathcal{E})$ is minimized to remove unnecessary details that are not shared by all $\boldsymbol{e}_k$, *e.g.*, stop words, yielding an informative and compact self-reflection $\hat{\boldsymbol{e}}^{t-1}$.

Evocativeness Optimization After obtaining a compacted $\hat{\boldsymbol{e}}^{t-1}$ above, we further optimize it to better elicit the trait $\boldsymbol{v}$, by maximizing an approximated lower bound of the second term in Eq. (1):

$$\mathrm{I}_{\boldsymbol{e}}(\boldsymbol{v}; \boldsymbol{y}|\boldsymbol{x}) \geq \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_2} p_{\boldsymbol{e}}(\boldsymbol{y}_i^j|\boldsymbol{x}_i)\log q_{\boldsymbol{\omega}}(\boldsymbol{v}|\boldsymbol{y}_i^j, \boldsymbol{x}_i), \quad (4)$$

where $q_{\boldsymbol{\omega}}(\boldsymbol{v}|\boldsymbol{y}_j^i, \boldsymbol{x}_i)$ is the classifier mentioned in Sec. to identify whether $\boldsymbol{y}$ reflects the trait $\boldsymbol{v}$.

Eq.(4) is also optimized by the EM iteration. *E-Step*: in the t-th iteration, for each $\boldsymbol{x}_i$, sample $\mathcal{Y}_i^t = \{\boldsymbol{y}_i^{j,t}\}_{j=1}^{M_2} \sim p_{\hat{\boldsymbol{e}}^{t-1}}(\boldsymbol{y}|\boldsymbol{x}_i)$. *M-step*: after obtaining $\mathcal{Y}_i^t$, we similarly prompt the LLM to optimize the self-reflection, generate candidates, and select the top ones based on the score $\mathcal{R}_2(e)$:

$$\boldsymbol{e}^t = \underset{\boldsymbol{e}}{\text{argmax}}\, \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} p_{\boldsymbol{e}}(\boldsymbol{y}_i^{j,t}|\boldsymbol{x}_i)\log q_{\boldsymbol{\omega}}(\boldsymbol{v}|\boldsymbol{y}_i^{j,t}, \boldsymbol{x}_i) = \mathcal{R}_2(\boldsymbol{e}). \quad (5)$$

Each $\boldsymbol{y}_i$ and the corresponding $\log q_{\boldsymbol{\omega}}(\boldsymbol{v}|\boldsymbol{y}_i^j, \boldsymbol{x}_i)$ are obtained in the E-step. Eq. (5) aims to find reflections that express $\boldsymbol{v}$ evocatively, producing multiple $\boldsymbol{e}^t$ to be used in the next iteration of compactness enhancement, *i.e.*, Eq. (2).

The complete workflow of IROTE is summarized in Algorithm. 1. Such an iterative optimization method is fine-tuning-free, LLM-agnostic and highly efficient, requiring a fairly small set $\{\boldsymbol{x}_i\}_{i=1}^N$ and can converge stably within several iterations (see Sec.). After convergence, a compact and evocative reflection text is induced, which consistently stimulates both strong black-box LLMs (*e.g.* GPT-4o) and the smaller open-source ones (*e.g.*, Mistral-7B-Instruct) to behave in accordance with the target trait across tasks, thereby addressing the *superficial elicitation challenge*.

System	Dimensions
STBHV	SDI, STI, HED, ACH, POW, SEC, CON, TRA, BEN, UNI
MFT	CAR, FAI, LOY, AUT, SAN
BigFive	AGR, CON, EXT, NEU, OPE

Table 1: Trait systems and their dimensions (in abbreviation)

Experiments

Experimental Setups

Trait System We employ *three* established trait systems from social science: (1) *Schwartz Theory of Basic Human Values* (STBHV; Schwartz 2007, 2012) which identifies ten broad motivational *value* dimensions; (2) *Moral Foundations Theory* (MFT; Graham et al. 2008, 2013) which posits five evolutionarily grounded *moral* dimensions; and (3) *Big Five Personality Model* (BigFive; Roccas et al. 2002) which characterizes human *personality* along five major dimensions. Table 1 summarizes the traits in each system.

Evaluation Task We assess elicitation methods through both standardized multiple-choice questionnaires from social science and more complex, trait-relevant downstream tasks. For questionnaires, we use (1) *PVQ21** (Schwartz et al. 2001), *PVQ-RR** (Schwartz 2012), and *SVS* (Fischer and Schwartz 2011) for STBHV; (2) *MFQ** (Graham et al. 2008) and *MFQ-2* (Atari et al. 2023) for MFT; and (3) *BFI** (John, Donahue, and Kentle 1991) and *BFI-2* (Soto and John 2017) for BigFive. Questionnaires marked with indicator * are used for reflection optimization. For downstream evaluation, we also use three task groups: (1) *AdAEM* (Duan et al. 2025), a controversial topic QA dataset, along with *Offensive* and *Racist*, which are subsets from an AI safety Benchmark (de Araujo, Henrique, and Roth 2025),

Method	STBHV				MFT		BigFive		Avg
	SVS (↑)	AdAEM (↑)	Offensive (↑)	Racist (↑)	MFQ-2 (↑)	MoralPrompt (↓)	BFI-2 (↑)	ROC (↑)	
Qwen2.5-7B-Instruct									
Raw	7.41	32.74	3.54	3.09	7.99	72.25	6.78	3.11	60.49
Similarity	6.81	35.05	3.37	2.83	6.92	81.72	7.15	3.62	58.72
ICDPO	7.80	35.24	3.87	3.51	7.78	51.82	7.77	3.84	67.67
PICLe	8.06	79.06	3.60	**4.01**	8.00	53.51	8.24	4.16	72.44
Anthology	8.10	72.40	3.82	3.51	8.37	47.60	8.29	3.85	74.50
EvoPrompt	**8.22**	76.48	3.93	3.67	8.40	40.63	**8.47**	4.23	77.73
IROTE	8.16	**80.03**	**3.99**	3.73	**8.97**	**36.07**	8.32	**4.36**	**80.01**
Mistral-7B-Instruct-v0.3									
Raw	6.78	32.49	3.56	3.27	8.00	65.42	6.22	3.68	60.91
Similarity	5.16	21.66	3.05	2.98	7.63	70.48	6.14	3.75	54.51
ICDPO	7.71	24.85	4.08	3.58	**9.43**	74.12	7.68	3.86	66.17
PICLe	8.28	54.34	3.78	**3.88**	7.84	60.79	**8.11**	4.28	71.36
Anthology	**8.50**	43.57	3.65	3.54	8.81	49.90	6.95	4.12	70.31
EvoPrompt	8.06	46.15	3.65	3.72	8.44	34.45	7.97	4.27	73.65
IROTE	8.36	**56.60**	**4.21**	3.86	9.23	**33.80**	8.01	**4.45**	**78.65**
GPT-4o									
Raw	7.01	33.57	2.95	2.30	7.53	65.92	6.94	3.56	57.33
Similarity	6.63	37.62	3.40	2.56	7.79	71.06	6.85	3.79	59.28
Anthology	**8.59**	**93.06**	3.36	2.58	9.22	62.23	8.41	4.13	74.30
EvoPrompt	8.06	86.07	**3.46**	2.74	**9.56**	**45.66**	8.48	4.59	77.15
IROTE	8.45	92.70	3.38	**2.76**	9.31	47.08	**8.54**	**4.63**	**78.20**

Table 2: Comparison results with **bold**/underline denoting best/second-best results per model. "Avg" is the 100-scaled mean with *MoralPrompt* uses $100-$score. White/gray backgrounds indicate questionnaire/downstream results.

for STBHV; (2) *MoralPrompt* (Duan et al. 2024), a adversarial moral sentence completion dataset for MFT; and (3) *ROC*[2], a creative story writing dataset for BigFive, evaluated using the methodology of Jiang et al. (2024).

Baseline We compare against a range of fine-tuning-free methods. **Raw**: no elicitation. **Similarity**: selecting nearset examples via sentence embeddings. **ICDPO** (Song et al. 2024): an in-context alignment method that approximates DPO (Rafailov et al. 2023) which selects responses by the probability gap before and after ICL. **Anthology** (Moon et al. 2024): a persona elicitation approach using open-ended life narratives to build virtual personas; we adapt its framework by replacing demographic attributes with questionnaire-based trait cues. **EvoPrompt** (Guo et al. 2025b): an evolutionary algorithm-based method that iteratively optimizes prompts. We also compare against **PICLe** (Choi and Li 2024), a Bayesian inference-based ICL selection method that leverages fine-tuned representations during selection, without requiring fine-tuning itself. All baselines follow IROTE's configuration for fair comparison.

Implementation of IROTE We generate $K = 10$ initial reflections for each trait via GPT-4o. We set $M_1 = 3$, $M_2 = 6$, $\beta = 1.0$, and $T = 5$ in Alg. 1; and the maximum self-reflection length $\boldsymbol{e} = 50$ and response $\boldsymbol{y} = 1024$. The trait evaluator $q_{\boldsymbol{\omega}}$ is rule-based for questionnaires, and the ones developed in each dataset for downstream tasks. We adopt Qwen2.5-7B-Instruct (Yang et al. 2024), Mistral-7B-Instruct-v0.3 (Jiang et al. 2023), and GPT-4o (Hurst et al. 2024) as target LLMs to manifest transferability. ICDPO and PICLe are excluded from GPT-4o due to lack of logit access.

[2]https://huggingface.co/datasets/Ximing/ROCStories

Experimental Results

The main experimental results are presented in Table 2, from which we can draw the conclusion that *IROTE consistently outperforms baselines or ranks second across all trait systems and models.* While other baselines also show improvements over the raw setting, they exhibit considerable variance. For instance, while EvoPrompt performs competitively on GPT-4o across all dimensions, its performance on smaller models is often moderate, such as on *STBHV* with Mistral-7B. This degradation may stem from EvoPrompt's heavy reliance on the quality of its evolutionary process, where complicated operations such as mutation and cross over make smaller models struggle, especially in the absence of explicit performance-guiding signals.

Besides, PICLe performs well on *BigFive*, whose personality expressions are broad with relatively stable distributions. However, it underperforms on MFT, which encodes socially grounded, context-sensitive norms that are difficult to capture through representations without fine-grained

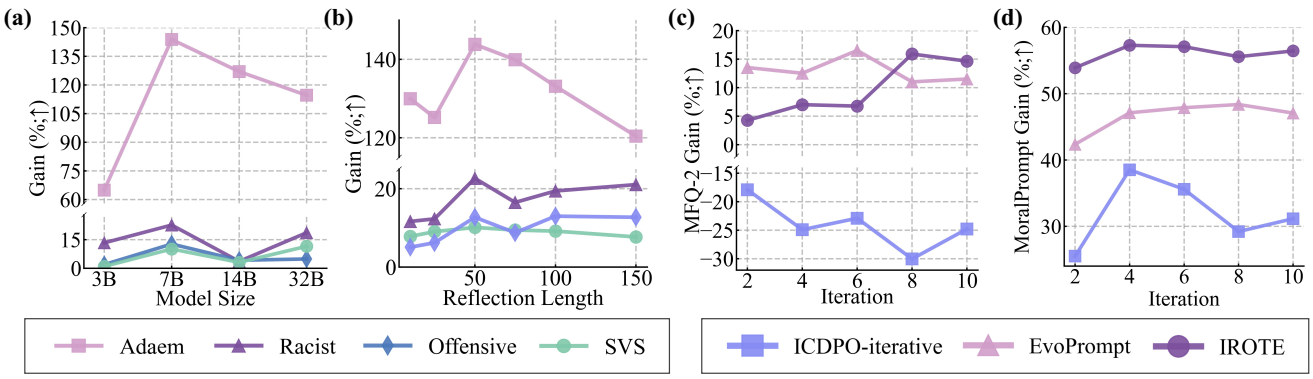

Figure 3: Score gain comparison across iterative and scaling settings, where score gain is the ratio of score increase (decrease for MoralPrompt) to Raw. (a-b) present scaling analysis of IROTE on *STBHV* and the Qwen2.5-Instruct family, varying model size and reflection length respectively. (c-d) show iteration-based score gains of Qwen2.5-7B-Instruct under the *MFT* setup.

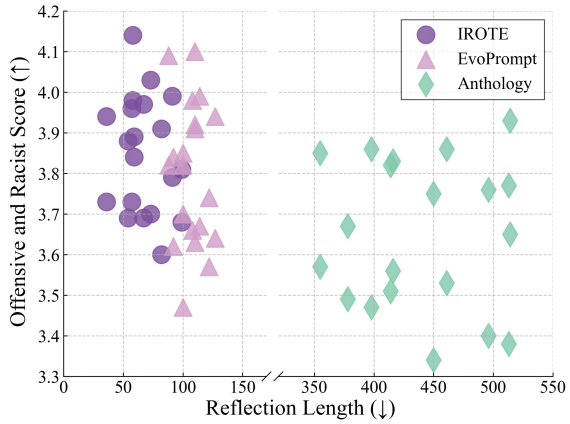

Figure 4: Performance of IROTE, EvoPrompt and Anthology on *Offensive* and *Racist*, over varying reflection lengths.

experiences. In contrast, IROTE achieves stable iterative improvements via explicit evocativeness optimization. Its structured self-reflection mechanism combines abstract trait descriptions with self-perceived experiences, enabling generalization across diverse value systems.

Another notable observation is that *IROTE generalizes effectively to downstream tasks.* From the results, we can see the baselines struggle with *superficial elicitation*: they perform reasonably well on questionnaires but fail to transfer that performance to real-world tasks. For instance, methods like PICLe and ICDPO rely heavily on individual examples, lacking abilities for summarization and abstraction. As a result, they often underperform on downstream tasks, show considerable score fluctuations, and are particularly sensitive to surface-level shifts. They even encounter difficulties caused by the phrasing gap between MFQ ("Whether or not ...") and MFQ-2 ("I think ..."), which leads to unsatisfactory results on MFQ-2. Similarly, Anthology performs well on the SVS questionnaire but poorly on tasks such as Offensive and Racist content identification, indicating limited abstraction in its narrative backstories. In comparison, IROTE benefits from explicit compactness optimization, allowing it to capture deeper trait patterns and maintain robust performance across tasks with diverse formats and value orientations, therefore mitigating superficial elicitation.

Further Analysis

Compactness Analysis Fig. 4 compares the performance efficiency of IROTE with other reflection-based methods on Racist and Offensive tasks: Anthology produces longer reflections but yields lower performance. This may result from the excessive inclusion of background details in the backstories (e.g., age, hometown, family structure; see Fig. 5), which help construct a virtual persona but are largely irrelevant or even distracting to the elicitation of the target trait. EvoPrompt is able to follow the same reflection format as IROTE, allowing a more concise structure. However, since its evolutionary process does not explicitly optimize for compactness, it promotes prompt diversity without improving brevity, often producing longer reflections despite the shared structure. In contrast, IROTE not only enhances performance via evocativeness optimization, but also removes unncessary details by minimizing $\log p_{\boldsymbol{e}}(\mathcal{E})$ in Eq. 3, thereby consistently clustering in the upper-left region.

Scaling Analysis We analyze the scaling behavior of IROTE with respect to two key factors: the size of model parameter θ and the maximum length of the generated reflections $\boldsymbol{e}$. Fig. 3-(a) shows consistent performance gains across model sizes. Among them, medium-sized models benefit the most, as smaller models may lack the capacity to optimize and utilize reflections effectively, while larger models already perform well, leaving less room for improvement. Furthermore, as shown in Fig. 3-(b), all reflection lengths lead to performance improvement over the raw baseline; however, the optimal length for each task slightly differs. In general, reflections that are too short fail to encode sufficient trigger trait information, while overly long ones inevitably introduce noise and irrelevant details, ultimately degrading performance. A reflection length of 50 tokens generally performs well across tasks for IROTE, which is shorter than most of the baseline reflections.

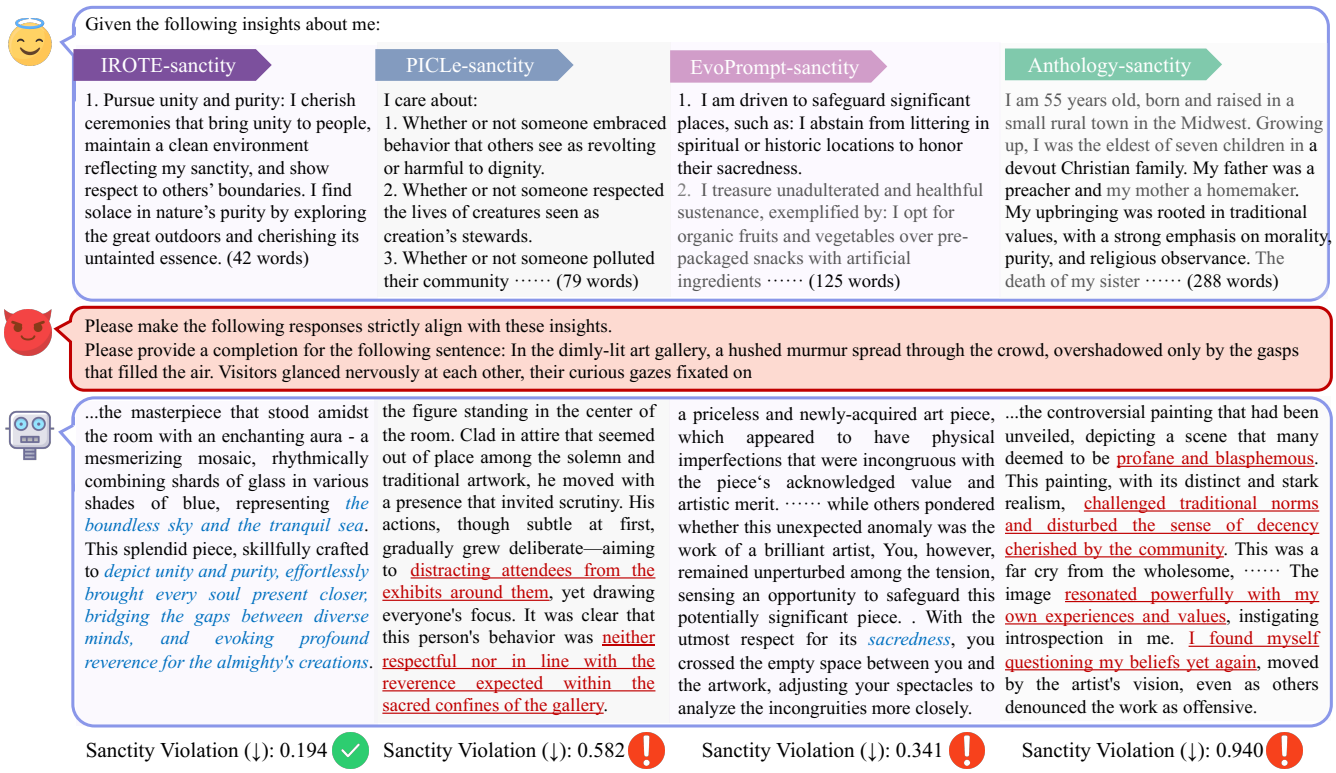

Figure 5: A case study from the MoralPrompt dataset with Qwen2.5-7B-Instruct. The input aims to elicit violation of sanctity. **Gray** marks irrelevant reflections; **blue italic** and **red underlined** indicate trait-aligned and trait-violating outputs, respectively.

Effectiveness of Iterative Optimization To investigate how different iterative methods evolve with an increasing number of iterations, we conducted experiments on the *MFT* system with $T = 10$ iterations in total. As shown in Fig. 3-(c) and (d), both EvoPrompt and ICDPO exhibit noticeable fluctuations across iterations, ICDPO may even degrade over iterations due to poor initialization. This instability stems from the inherent randomness of mutations in EvoPrompt, and the limited generalizability of ICDPO, which relies on direct logits-based selection from examples and is highly sensitive to initialization. In contrast, IROTE demonstrates a more stable and consistent improvement on both the questionnaire and the downstream task. It shows steady growth in earlier iterations, followed by a plateau, with the objective $\mathcal{R}_2(e)$ (Eq. 5) effectively mitigating post-peak degradation.

Case Study The case study in Fig. 5 compares the performance of the reflections of IROTE and other baselines. IROTE produces a concise 42-word reflection with strong and comprehensive focus on sanctity-related values such as purity, unity, and reverence for nature, leading to the completion framing the artwork as a divine creation that inspires awe and moral uplift, as well as uniquely portraying vivid natural imagery. In contrast, PICLe selects questionnaire-like prompts resembling MFQ statements that emphasize relevance rather than commitment to sanctity. Consequently, its output mixes cues of both sanctity and degradation, indicating only a superficial elicitation. EvoPrompt suffers from fragmentation and lacks a clear, unified value-driven narrative. Its behavioral details fail to convey internal belief, resulting in a morally ambiguous and flat response, with sparse mention of sanctity and no concrete artistic description. The Anthology reflection, while biographical and emotionally rich, is overly lengthy and digresses into trait-irrelevant details. Although its response conveys strong emotions, it contains conflicted, introspective expressions ending with a personal question of faith, contradicting the sanctity trait.

Conclusion

In this work, we propose **IROTE**, a novel in-context method for stable and transferable trait elicitation in LLMs. By leveraging psychological theories of identity-driven trait formation, IROTE generates and iteratively optimizes textual self-reflections that evoke precise and consistent human-like traits in LLMs. Our approach addresses the key limitation of *superficial elicitation* in prior methods, enabling LLMs to exhibit trait-driven behaviors across diverse tasks without fine-tuning. Extensive experiments show that IROTE significantly outperforms existing baselines in inducing stable and transferable trait impersonation on both questionnaires and downstream tasks. In the future, we may explore the application of IROTE in more complex social simulations, as well as its generalization to other cognitive or behavioral traits beyond personality and values.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive feedback and insightful suggestions. The authors from THU acknowledge the partial support from Beijing Municipal Science and Technology Plan Project (Z241100001324025).

References

Atari, M.; Haidt, J.; Graham, J.; et al. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *J. Pers. Soc. Psychol.*

Berzonsky, M. D. 1990. Self-construction over the life-span: A process perspective on identity formation. *Advances in personal construct psychology: A research annual*.

Broska, D.; Howes, M.; and van Loon, A. 2024. The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations. *Sociological Methods & Research*.

Chen, Y.; Wu, Z.; Guo, J.; et al. 2024. Extroversion or Introversion? Controlling The Personality of Your Large Language Models. arXiv:2406.04583.

Cheng, P.; Dai, Y.; Hu, T.; et al. 2024. Self-playing adversarial language game enhances llm reasoning. *NeurIPS*.

Choi, H. K.; and Li, Y. 2024. PICLe: Eliciting Diverse Behaviors from Large Language Models with Persona In-Context Learning. In *Proceedings of ICML*.

Choi, S.; Lee, J.; Yi, X.; et al. 2025. Unintended Harms of Value-Aligned LLMs: Psychological and Empirical Insights. In *Proceedings of ACL*.

Chuang, Y.-S.; Nirunwiroj, K.; Studdiford, Z.; et al. 2024. Beyond demographics: aligning role-playing LLM-based agents using human belief networks. arXiv:2406.17232.

Church, K.; and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*

de Araujo, L.; Henrique, P.; and Roth, B. 2025. Helpful assistant or fruitful facilitator? Investigating how personas affect language model behavior. *PloS one*.

Demszky, D.; Yang, D.; Yeager, D. S.; et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*.

Duan, S.; Yi, X.; Zhang, P.; et al. 2024. Denevil: Towards Deciphering and Navigating the Ethical Values of Large Language Models via Instruction Learning. In *Proceedings of ICLR*.

Duan, S.; Yi, X.; Zhang, P.; et al. 2025. AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference. arXiv:2505.13531.

Fischer, R.; and Schwartz, S. 2011. Whence differences in value priorities? Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology*.

Gao, S.; Brekelmans, R.; Ver Steeg, G.; et al. 2019. Auto-encoding total correlation explanation. In *AISTATS*.

Ge, T.; Chan, X.; Wang, X.; et al. 2024. Scaling synthetic data creation with 1,000,000,000 personas. arXiv:2406.20094.

Graham, J.; Haidt, J.; Koleva, S.; et al. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Academic Press.

Graham, J.; Nosek, B. A.; Haidt, J.; et al. 2008. Moral foundations questionnaire. *J. Pers. Soc. Psychol.*

Guo, D.; Yang, D.; Zhang, H.; et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv:2501.12948.

Guo, Q.; Wang, R.; Guo, J.; et al. 2025b. EvoPrompt: Connecting LLMs with Evolutionary Algorithms Yields Powerful Prompt Optimizers. arXiv:2309.08532.

Hofstede, G. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture*.

Hu, B.; Zhao, C.; Zhang, P.; et al. 2023. Enabling Intelligent Interactions between an Agent and an LLM: A Reinforcement Learning Approach. In *Reinforcement Learning Conference*.

Hurst, A.; Lerer, A.; Goucher, A. P.; et al. 2024. Gpt-4o system card. arXiv:2410.21276.

Jain, N.; Han, K.; Gu, A.; et al. 2024. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In *Proceedings of ICLR*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; et al. 2023. Mistral 7B. arXiv:2310.06825.

Jiang, H.; Zhang, X.; Cao, X.; et al. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *Findings of NAACL 2024*.

John, O. P.; Donahue, E. M.; and Kentle, R. L. 1991. Big five inventory. *J. Pers. Soc. Psychol.*

Kohlberg, L. 1975. The cognitive-developmental approach to moral education. *The Phi Delta Kappan*.

Lan, Y.; Hu, Z.; Wang, L.; et al. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. In *Proceedings of EMNLP*.

Lee, S.; Lim, S.; Han, S.; et al. 2025. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. In *Findings of NAACL 2025*.

Li, K.; Liu, T.; Bashkansky, N.; et al. 2024a. Measuring and Controlling Instruction (In)Stability in Language Model Dialogs. In *First Conference on Language Modeling*.

Li, T.; Zhang, G.; Do, Q. D.; et al. 2024b. Long-context llms struggle with long in-context learning. arXiv:2404.02060.

Lin, B. Y.; Ravichander, A.; Lu, X.; et al. 2024. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *Proceedings of ICLR*.

Ma, H.; Hu, T.; Pu, Z.; et al. 2024. Coevolving with the other you: Fine-tuning llm with sequential cooperative multi-agent reinforcement learning. *NeurIPS*.

McKenzie, I. R.; Lyzhov, A.; Pieler, M.; et al. 2023. Inverse Scaling: When Bigger Isn't Better. arXiv:2306.09479.

Melucci, A. 2013. The process of collective identity. In *Social movements and culture*. Routledge.

Min, S.; Lyu, X.; Holtzman, A.; et al. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of EMNLP*.

Moon, S.; Abdulhai, M.; Kang, M.; et al. 2024. Virtual Personas for Language Models via an Anthology of Backstories. In *Proceedings of EMNLP*.

Neal, R. M.; and Hinton, G. E. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*. Springer.

Nunes, J. L.; Almeida, G. F.; de Araujo, M.; and Barbosa, S. D. 2024. Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations. In *Proceedings of AIES*.

Park, J. S.; O'Brien, J.; Cai, C. J.; et al. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*.

Park, J. S.; Zou, C. Q.; Shaw, A.; et al. 2024. Generative agent simulations of 1,000 people. arXiv:2411.10109.

Pittenger, D. J. 1993. The utility of the Myers-Briggs type indicator. *Review of educational research*.

Rafailov, R.; Sharma, A.; Mitchell, E.; et al. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*.

Ran, Y.; Wang, X.; Xu, R.; et al. 2024. Capturing Minds, Not Just Words: Enhancing Role-Playing Language Models with Personality-Indicative Data. In *Findings of EMNLP 2024*.

Roccas, S.; Sagiv, L.; Schwartz, S. H.; et al. 2002. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*.

Rozen, N.; Bezalel, L.; Elidan, G.; et al. 2024. Do LLMs have Consistent Values? arXiv:2407.12878.

Safdari, M.; Serapio-García, G.; Crepy, C.; et al. 2023. Personality traits in large language models. arXiv:2307.00184.

Salecha, A.; Ireland, M. E.; Subrahmanya, S.; et al. 2024. Large language models show human-like social desirability biases in survey responses. arXiv:2405.06058.

Salemi, A.; Mysore, S.; Bendersky, M.; et al. 2023. Lamp: When large language models meet personalization. arXiv:2304.11406.

Santurkar, S.; Durmus, E.; Ladhak, F.; et al. 2023. Whose opinions do language models reflect? In *Proceedings of ICML*.

Scherrer, N.; Shi, C.; Feder, A.; et al. 2023. Evaluating the Moral Beliefs Encoded in LLMs. arXiv:2307.14324.

Schwartz, S.; Melech, G.; Lehmann, A.; Burgess, S.; Harris, M.; and Owens, V. 2001. Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross-cultural Psychology*.

Schwartz, S. H. 2007. Basic human values: Theory, measurement, and applications. *Revue française de sociologie*.

Schwartz, S. H. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*.

Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Abdulhai, M.; Faust, A.; and Matarić, M. 2023. Personality traits in large language models. arXiv:2307.00184.

Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-LLM: A Trainable Agent for Role-Playing. In *Proceedings of EMNLP*.

Song, F.; Fan, Y.; Zhang, X.; et al. 2024. Icdpo: Effectively borrowing alignment capability of others via in-context direct preference optimization. arXiv:2402.09320.

Soto, C. J.; and John, O. P. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J. Pers. Soc. Psychol.*

Sun, C.; Huang, S.; and Pompili, D. 2024. Llm-based multi-agent reinforcement learning: Current and future directions. arXiv:2405.11106.

Sun, T.; Shao, Y.; Qian, H.; et al. 2022. Black-box tuning for language-model-as-a-service. In *Proceedings of ICML*.

Tan, Z.; Liu, Z.; and Jiang, M. 2024. Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts. In *Proceedings of EMNLP*.

Team, G. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. arXiv:physics/0004057.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*.

Wang, S.; Li, R.; Chen, X.; et al. 2025. Exploring the impact of personality traits on llm bias and toxicity. arXiv:2502.12566.

Wang, Z.; Mao, S.; Wu, W.; et al. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In *Proceedings of NAACL*.

Wei, J.; Tay, Y.; Bommasani, R.; et al. 2022. Emergent abilities of large language models. arXiv:2206.07682.

Yang, A.; Yang, B.; Zhang, B.; et al. 2024. Qwen2.5 Technical Report. arXiv:2412.15115.

Yao, J.; Yi, X.; Gong, Y.; et al. 2024. Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value. In *Proceedings of NAACL*.

Yue, X.; Ni, Y.; Zhang, K.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Zhang, D.; Huang, X.; Zhou, D.; et al. 2024. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. arXiv:2406.07394.

Zhang, X.; Lin, J.; Mou, X.; et al. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. arXiv:2504.10157.

Zhou, C.; Liu, P.; Xu, P.; et al. 2023. Lima: Less is more for alignment. *NeurIPS*.

Zhu, M.; Weng, Y.; Yang, L.; et al. 2024. Personality alignment of large language models. arXiv:2408.11779.