PsyMem: Fine-grained Psychological Alignment and Explicit Memory Control for Advanced Role-Playing LLMs

Xilong Cheng[1,*], **Yunxiao Qin**[1,2,*,†], **Yuting Tan**[1,*], **Zhengnan Li**[1], **Ye Wang**[1,2], **Hongjiang Xiao**[1,2,†], **Yuan Zhang**[1,2]

[1] Communication University of China, China
[2] State Key Laboratory of Media Convergence and Communication, China
{chengzhengyu330, qinyunxiao, yutingtan, fmlyd, yewang, xiaohj, yzhang}@cuc.edu.cn

Abstract

Existing LLM-based role-playing methods often rely on superficial textual descriptions or simplistic metrics, inadequately modeling both intrinsic and extrinsic character dimensions. Additionally, they typically simulate character memory with implicit model knowledge or basic retrieval augment generation without explicit memory alignment, compromising memory consistency. The two issues weaken reliability of role-playing LLMs in several applications, such as trustworthy social simulation. To address these limitations, we propose PsyMem, a novel framework integrating fine-grained psychological attributes and explicit memory control for role-playing. PsyMem supplements textual descriptions with 26 psychological indicators to detailed model character. Additionally, PsyMem implements memory alignment training, explicitly trains the model to align character's response with memory, thereby enabling dynamic memory-controlled responding during inference. By training Qwen2.5-7B-Instruct on our specially designed dataset (including 5,414 characters and 38,962 dialogues extracted from novels), the resulting model, termed as PsyMem-Qwen, outperforms baseline models in role-playing, achieving the best performance in human-likeness and character fidelity.

1 Introduction

Recent advancements in large language models (LLMs) (Achiam et al., 2023; Anthropic, 2024a; Anil et al., 2023) have unlocked new possibilities for implementing role-playing systems through language-based behavior simulation (Shanahan et al., 2023). Role-playing systems demonstrate societal value across applications like interactive gaming (Wang et al., 2023a) and AI counseling (Zhang et al., 2024), with particular potential in enabling high-reliable social simulation. Such simulation allows modeling of human interactions and societal dynamics (Park et al., 2023), offering insights into evidence-based policy design and conflict mediation (Zhou et al., 2024). To achieve trustworthy social simulation, LLM-based role-playing must preserve persistent traits, memories, and behavioral logic to ensure agents reliably mirror specified attributes in evolving social contexts.

Recent works in social or group behavior simulation (Kosinski, 2024; Park et al., 2023; Zhou et al., 2024) often rely on general LLMs for role-playing, using context-based learning (Mann et al., 2020) and instruction follow-up (Ouyang et al., 2022). However, these approaches (Yu et al., 2022) commonly struggle to precisely control character attributes due to the unsatisfactory role-playing ability of general LLMs, especially for ordinary characters, as Figure 1 shows. Some recent works (Shao et al., 2023; Zhou et al., 2023; Tao et al., 2023) proposed the LLMs that are dedicated to role-playing to improve character attribute consistency. However, there are still the following two issues: 1) Oversimplified characterization and 2) Weak memory control.

The Oversimplified Characterization Issue. From both neuroscientific and psychological perspectives, human attributes can be described through two broad dimensions: intrinsic attributes and extrinsic attributes (Snyder, 1983; Kahneman, 2011). Intrinsic attributes, such as personality traits and values (Ravlin and Meglino, 1987; Stumpf and Dunbar, 1991), play a significant role in decision-making, particularly in long-term decisions (Borghans et al., 2008; Roberts et al., 2008). On the other hand, extrinsic attributes (Kahneman,

[*]Equal contribution.
[†]Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 510–529, 2026. https://doi.org/10.1162/TACL.a.645
Action Editor: Giuseppe Carenini. Submission batch: 4/2025; Revision batch: 12/2025; Published 4/2026.

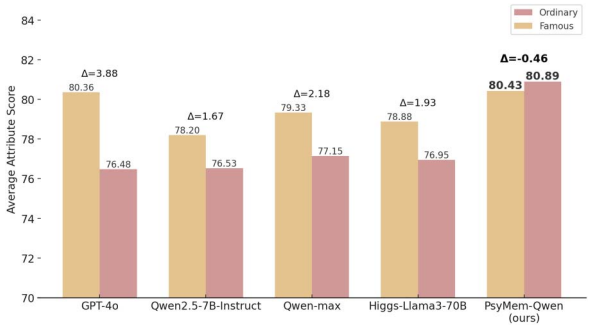

Figure 1: The performance comparison is conducted on two subsets: ‘‘Ordinary,’’ consisting of 20 randomly selected characters from our test set, and ‘‘Famous,’’ consisting of 20 well-known characters. The average attribute scores are calculated as the mean of the quantized scores across the 20 characters in each subset.

2011), such as behavior, language, and body language, reflect an individual's intuitive responses in immediate situations (Bargh et al., 1996).

However, existing role-playing LLMs typically rely on basic textual descriptions or narrow metrics to define the target role, which fails to capture the full complexity and depth of the character. For example, Zhou et al. (2023) define character attributes using textual descriptions based on seven dimensions (including identity, interests, experiences, and other relevant traits). Some recent research (Ran et al., 2024; Yu et al., 2024; Cui et al., 2023) has attempted to use MBTI for personality trait evaluation to improve character control. However, MBTI employs a binary approach to assess personality traits, which has low test-retest reliability (Stein and Swan, 2019; Hunsley et al., 2015) and cannot precisely quantify attributes (Nowack, 1996), such as how introverted a person is.

The Weak Memory Control Issue. Numerous neuroscientific studies (Stumpf and Dunbar, 1991; Ravlin and Meglino, 1987; Pickering and Garrod, 2004) have shown that, during decision-making, the brain relies on not only personality traits (such as emotional responses and values) but also long-term memory, particularly in fast-paced decision-making contexts. However, existing role-playing models (Shao et al., 2023; Zhou et al., 2023) typically rely on the model's internal knowledge to simulate character memory, which works for famous characters but often fails to provide memory for less familiar or unseen ones. Retrieval-Augmented Generation (RAG) (Ram et al., 2023) has been proposed as a promising solution (Zhong et al., 2024), but we observe that simply integrating RAG within Role-Play LLMs' inference stage does not adequately address memory management in the context of role-play (see Table 2).

In this work, we propose a widely recognized modern psychological evaluation framework (Costa and McCrae, 2008; Schwartz, 1992; Thomas, 2008; Zimbardo and Ruch, 1975) and a graph-structured memory mechanism to construct the PsyMem framework, enabling LLMs to achieve exceptional role-playing abilities. Drawing from modern psychology and neuroscience, we enhance the description of character attributes by leveraging both **latent psychological attributes** and **explicit behavioral patterns** (Snyder, 1983; Kahneman, 2011), with personality and values corresponding to latent psychological attributes, and behavioral decision, social interaction, and leadership representing explicit behavioral patterns. Overall, we extract real data from novels and construct a large-scale role-playing dataset, using 26 quantitative indicators and a small amount of textual descriptions to describe characters. As shown in Table 1, this dataset contains 5,414 established characters, 38,962 conversations, and 536,636 utterances, much larger than existing role-playing datasets.

For character memory, inspired by the hippocampus and cognitive mapping mechanisms (Eichenbaum, 2015; Hartley et al., 2014; Tacikowski et al., 2024; Tavares et al., 2015), we transform characters, events, and relationships into a knowledge graph. Unlike previous role playing work (Edge et al., 2024) directly using implicit model knowledge or integrating role memory in inference, **we incorporate memory information into the training process, forcing the model to learn how to respond given retrieved memory**. This approach effectively enhances the model's memory consistency, ensuring that role-playing model generates responses based on not only character attributes but also memory.

In general, our five main contributions are as follows:

- We propose **PsyMem**, a novel LLM-based role-playing framework that establishes precise control over character behaviors through the synergistic integration of fine-grained psychological profiling (e.g., personality, value) and dynamic memory schemata.

Dataset	Character		Conversation			Profile Style			
	Num.	SA	Conv.	Avg. Turns	Auth.	PsyFrm	QI	Text Desc.	Expl. Mem.
HPD	113	✗	1,191	13.2	✓	✗	✗	✓	✗
CharacterGLM	250	✗	1,034	15.8	✗	✗	✗	✓	✗
RoleLLM	100	✗	140,726	2	✗	✗	✗	✓	✗
CharacterLLM	9	✗	14,300	13.2	✗	✗	✗	✓	✗
ChatHaruhi	32	✗	54,726	>2	✗	✗	✗	✓	✗
DITTO	4002	✗	7,186	5.1	✗	✗	✗	✓	✗
Beyond Dialogue	311	✓	3,552	6.5	✓	✗	✓	✓	✗
CoSER	17,966	✗	29,798	13.2	✓	✗	✗	✓	✗
MMRole	85	✗	14,346	4.2	✗	✗	✗	✓	✗
CharacterBench	3956	✗	13,162	11.3	✗	✗	✗	✓	✗
CharacterEval	77	✗	1,785	9.3	✓	✗	✗	✓	✗
Ours	5,414	✓	38,962	13.8	✓	✓	✓	✓	✓

Table 1: Dataset Statistics. Comparison between our dataset and existing open-source role-play datasets. Columns are categorized into: Character (Num.: number of character roles; SA: scene alignment), Conversation (Conv.: number of dialogues; Avg. Turns: average dialogue length; Auth.: whether dialogues are fully sourced from real, non-generated conversations), and Profile Style (PsyFrm: adherence to a psychological framework; QI: presence of quantitative indicators; Text Desc.: availability of textual character descriptions; Expl. Mem.: support for explicit memory retrieval during interaction).

- Beyond textual descriptions, we present a comprehensive attribute set that bridges intrinsic (e.g., values, emotions) and extrinsic (e.g., behaviors, social interactions) character dimensions, operationalized via **26 measurable indicators**. This attribute set captures both the internal motivations and external behavioral manifestations of characters.
- We introduce a memory alignment training paradigm that establishes explicit cognitive anchoring between role-playing responses and character memory. This methodology enhances role-playing through a dual-control mechanism, enabling **precise regulation of generative outputs via both attributes (e.g., personality dimensions) and episodic memory constraints**.
- We introduce the use of role-playing style general supervised fine-tuning data (transforming general SFT data into role-playing style), demonstrating that incorporating such data effectively enhances the performance of Role-Playing LLMs.
- Extensive experiments demonstrate that the proposed PsyMem framework significantly improves the role-playing ability of LLMs. For instance, by training Qwen2.5-7B-Instruct on our dataset (introduced in Section 3) specially designed for PsyMem, the resulting model, termed as PsyMem-Qwen, outperforms baseline models in role-playing, achieving the best performance in human-likeness and character fidelity.

2 Related Work

2.1 LLM-based Role-Playing

In existing role-playing works, the main approaches are Nonparametric Prompting and Parametric Training (Chen et al., 2024). In Nonparametric Prompting, Yu et al. (2022) use carefully designed character descriptions to control general LLMs for role-playing without fine-tuning. However, they face significant challenges in accurately reflecting the intrinsic relationships between character profiles and dialogue content.

In Parametric Training, Chen et al. (2023) focus on mimicking characters from the Harry Potter series, while Wang et al. (2023b) introduced the first role-playing dataset with 100 characters generated using GPT-3.5 prompts. Subsequent works (Li et al., 2023; Zhou et al., 2023; Shao et al., 2023) have expanded and enriched these datasets by improving the number of characters and the diversity of dialogue scenarios, all leveraging various GPT models. Meanwhile, Lu et al. (2024)

Model	Character Fidelity						Character-independent		
	Per.	Val.	SL*	BD*	Mem.	Avg.	H-like*	Cons.	Coh.
General LLMs									
GPT-4o	79.08	80.11	77.43	73.29	94.40	80.86	66.60	89.28	**99.20**
GPT-3.5-Turbo	79.08	79.42	74.15	72.11	84.60	77.87	34.40	88.60	96.20
Claude-3.5-sonnet	80.34	80.29	78.51	74.57	91.80	81.10	75.00	89.50	**99.20**
Qwen-Max	78.67	80.42	76.54	72.37	90.60	79.72	57.60	89.48	97.60
Yi-Large	80.20	80.09	78.27	76.82	92.00	81.48	83.80	88.18	98.40
Deepseek-R1	80.98	81.38	77.88	77.12	**95.00**	82.47	77.20	87.57	96.20
Deepseek-V3	79.82	80.33	79.16	75.08	93.40	81.56	76.40	89.04	99.00
LLaMA3.1-8B-instruct	79.17	80.14	76.76	76.01	83.20	79.06	64.20	88.74	99.00
Qwen2.5-7B-Instruct	77.44	80.02	76.15	76.51	85.60	79.14	64.40	89.58	97.40
Role-play LLMs									
CharacterGLM-6B	80.66	80.67	79.35	71.20	40.20	70.42	52.80	89.61	73.00
Baichuan-NPC-Turbo	76.53	79.05	74.29	72.27	87.20	77.87	61.60	**90.36**	96.40
Hunyuan-Role	79.31	79.71	78.24	78.38	82.00	79.53	87.40	89.16	91.20
Higgs-LLaMA3-70B	80.02	80.61	75.77	74.08	93.20	80.74	60.40	88.40	99.60
CoSER-70B	**81.21**	81.90	78.83	76.35	93.00	82.28	81.80	89.44	97.20
PsyMem-LLaMA (**ours**)	80.47$^{+1.30}$	**81.93**$^{+1.79}$	78.57$^{+1.81}$	74.47$^{-1.54}$	90.20$^{+7.00}$	81.13$^{+2.07}$	85.20$^{+21.00}$	89.93$^{+1.19}$	95.60$^{-3.40}$
PsyMem-Qwen (**ours**)	80.40$^{+2.96}$	81.74$^{+1.72}$	**80.80**$^{+4.65}$	**78.48**$^{+1.97}$	91.80$^{+6.20}$	**82.64**$^{+3.50}$	**87.60**$^{+23.20}$	89.64$^{+0.06}$	96.20$^{-1.20}$

Table 2: Model evaluation across multiple criteria. **Bold** indicates best performance. Superscripts show change from base model. Abbreviations: Per. = Personality, Val. = Values, SL* = Social & Leadership, BD* = Behavioral Decision, Mem. = Memory, H-like* = Human-likeness, Cons. = Consistency, Coh. = Coherence.

adopted a self-alignment approach to dataset construction, moving away from the previous method of cheaply imitating the role-playing capabilities of GPT models. Zhou et al. (2023) leveraged GPT-4 to extract character dialogues from a diverse collection of Chinese novels and scripts. More recently, Yu et al. (2024) focused on addressing the bias between predefined characters and specific scenario dialogues, and constructed a dataset of 311 characters.

Our work belongs to parametric training. We argue that a character's behavior is jointly influenced by their attributes, memory, and the scenario. Drawing from modern psychology, we segment character attributes using contemporary psychological frameworks and construct a dataset containing 38,962 conversations (536,636 utterances) from real novels, leveraging both character attributes and memory to enhance role-playing ability.

2.2 Memory Mechanism in Role-Playing

Existing LLM-based Role-Playing commonly relies on the internal knowledge within the model weights to implicitly mimic character memory (Zhou et al., 2023; Lu et al., 2024; Li et al., 2023). For example, Shao et al. (2023) and others (Lu et al., 2024; Yu et al., 2024) push the boundaries of knowledge to protect character memory, while Lu et al. (2024) and others use a self-alignment approach to distinguish the different memory backgrounds of various characters. Relying on the internal knowledge within the model is effective for famous characters but fails for unfamous and unseen characters. Some works apply RAG (Ram et al., 2023) to dynamically retrieve external contexts for explicitly mimicking character memory, such as Zhong et al. (2024) who adopts memory update mechanisms based on Ebbinghaus's forgetting curve theory. However, as shown in Table 2, existing role-playing models tend to perform better when simulating famous characters, while their performance significantly drops for ordinary characters that rely on external memory.

Unlike previous works, we integrate explicit memory control into the training framework of role-playing models. This approach encourages the model to strictly rely on actual retrieved memories during dialogue generation in inference.

2.3 Role-playing Evaluation

The current evaluation process consists of dialogue generation and dialogue assessment. Due

to its efficiency and completeness, automated dialogue generation is widely adopted. Evaluation methods are generally categorized into three types: Metric-based Evaluation (Li et al., 2023; Wang et al., 2023b), Human Evaluation (Zhou et al., 2023), and ‘‘LLMs as Judges’’ (Shao et al., 2023; Lu et al., 2024). Metric-based evaluation primarily measures the model's alignment with standard responses, such as original text from novels or manually annotated content. While human evaluation is more precise, its high cost and lack of reproducibility limit its scalability. In contrast, ‘‘LLMs as Judges’’, with its efficiency, low cost, and strong scalability, is increasingly becoming the preferred choice. The experiments in Appendix (see Table 8) demonstrate the effectiveness of the ‘‘LLMs as Judges’’ paradigm for role-playing evaluation.

3 Dataset Architecture

An increasing number of role-playing datasets (Ran et al., 2024; Cui et al., 2023) have begun incorporating modern psychological indicators into their characterization frameworks. However, previous studies typically introduced only a limited selection of psychological attributes as supplementary descriptors, without comprehensively integrating psychological systems into character portrayals (Yu et al., 2024; Wang et al., 2024). Moreover, existing datasets predominantly focus on character attributes, overlooking the significant influence that memory has on characters' behaviors, thought processes, and speech styles (Stumpf and Dunbar, 1991; Ravlin and Meglino, 1987; Pickering and Garrod, 2004). This limitation becomes particularly evident when LLMs can hardly access the background knowledge of the role-playing characters during the pretraining stage (see Figure 1).

To address these shortcomings, we design a novel dataset with the architecture formulated as

$$\mathrm{D_{RP}} = \{(R_i, P_i, M_i, C_i, x_i, y_i) | i = 1, 2, \ldots, N\}, \quad (1)$$

where N is the number of data sample. Each sample is a real dialogue turn collected from novels. R_i is the role for response; P_i is the profile of R_i, including basic information and 26 quantitative psychological dimensions; C_i consists of the dialogue turns preceding the current turn, as well as the scene information. x_i is the query to R_i and y_i is the corresponding response. M_i denotes the memory retrieved for role R_i, conditioned on the query x_i and the context C_i, which comprises the preceding dialogue turns and the scene.

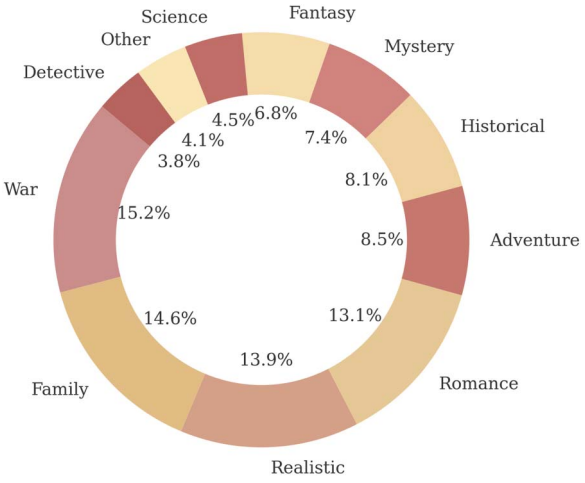

Figure 2: The genre distribution in the dataset.

In this work, the dataset $\mathrm{D_{RP}}$ contains 38,962 samples, extracted from 539 novels characterized by a wide range of genres, rich content, and narrative perspectives of third person. (The genre distribution is presented in Figure 2.)

3.1 Quantifiable Character Attributes

We systematically construct quantifiable character attribute dimensions aligned with contemporary psychological frameworks by categorizing character traits into **latent psychological attributes** and **explicit behavioral patterns** (Kahneman, 2011; Snyder, 1983), as illustrated in Figure 3a. These two dimensions distinctly influence individual behavior, communication styles, decision-making processes, and interpersonal interactions. Besides, a small amount of text describing the social attributes of the characters is also included as a supplement.

Latent psychological attributes reflect internal, often subconscious aspects of character, guiding motivations, preferences, and emotional responses. To describe these latent psychological attributes comprehensively, we utilize the following psychological frameworks:

1) **Personality**: Evaluated using the Big Five Personality Model (Costa and McCrae, 2008), encompassing stable emotional, cognitive, and behavioral tendencies. Traits

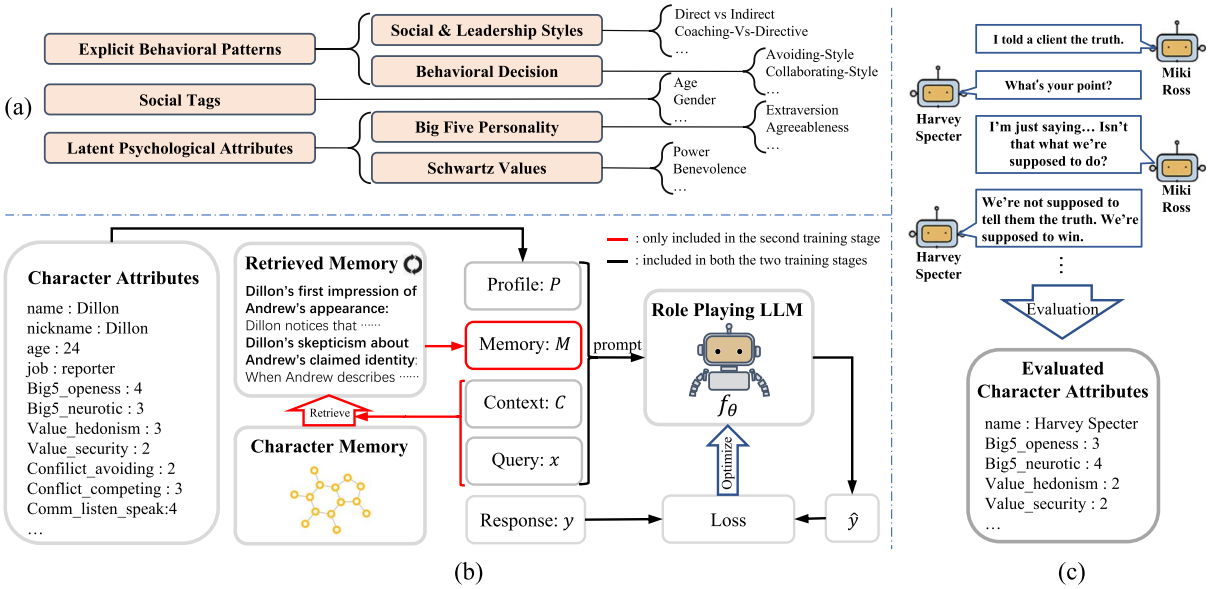

Figure 3: (a): The fine-grained character attributes designed in this work. (b): The two stages training of role-playing LLM. First, the model learns basic role-playing without character memory. In the second stage, we dynamically retrieves memory relevant to the current query and dialogue context from graph-structured character memory. We then integrate character profile, retrieved memories, dialogue history, and current input to enhance role-playing precision, training the model to align responses with both fine-grained character profile and contextual memory. (c): We assess the role-playing LLM by first generating a multi-turn dialogue (up to 15 turns) between two designated roles, followed by three rounds of scoring with GPT-4o based on a quantitative rubric, and report the mean score.

such as openness, conscientiousness, extraversion, agreeableness, and neuroticism directly influence communication styles, social interactions, and general life choices.

2) **Values**: Assessed through Schwartz's Theory of Basic Values (Schwartz, 1992), which classifies core beliefs into ten universally recognized types. These values deeply affect decision-making, priorities, ethical judgments, and interpersonal alignment, enriching the psychological depth and motivational dimensions of individuals.

Explicit behavioral patterns manifest in observable actions, interactions, and communication styles. We capture these explicit patterns using:

1) **Social & Leadership Styles**: Measured through six dimensions based on Zimbardo's psychological principles (Zimbardo and Ruch, 1975), significantly influencing observable interpersonal behaviors, communication effectiveness, leadership approaches, and social dynamics in groups.

2) **Behavioral Decision**: Analyzed using the Thomas-Kilmann Conflict Mode Instrument (TKI) (Thomas, 2008), emphasizing decision-making processes when characters face conflicts or challenges. This explicitly captures how emotional, social, and heuristic influences shape specific decisions, providing precise and actionable insights into behavioral responses.

3.2 Character Memory Integration

Neuroscientific research underscores the significant role memory plays in shaping human behavior, communication, and decision-making processes (Stumpf and Dunbar, 1991; Ravlin and Meglino, 1987; Pickering and Garrod, 2004). Memories influence not only how individuals perceive and interpret events but also guide their interactions and responses within social contexts, highlighting the importance of accurately capturing and utilizing memory in character portrayal.

Therefore, we introduce character memory data M to the role playing dataset to enable the memory simulation capability of role playing model. Specifically, we first build a knowledge graph for

each novel via GraphRAG (Edge et al., 2024), and then retrieve relevant memory context for the current dialogue in local mode (see Appendix B.1 for details). Memory alignment is evaluated on two key dimensions: *correctness* and *rationality*. Correctness assesses whether the model accurately utilizes stored character information in response to queries, while rationality examines logical coherence and contextual appropriateness. Additionally, simple summaries or vague memory references are identified as non-rational uses of memory. To further enhance robustness, irrelevant memory items are deliberately introduced as noise, promoting stable, realistic, and consistent character portrayals reflective of human memory dynamics.

In addition, we define character-specific knowledge boundaries to ensure consistency—for example, Quasimodo from The Hunchback of Notre-Dame would not be aware of torch. Based on the absence of relevant retrievable memories, we used GPT-4o to generate 800 refusal-style QA pairs for 100 characters, grounded in real-world constraints and narrative context. These were integrated into novel-based dialogues to maintain coherence and character fidelity.

3.3 Synthetic Role-playing Data

In addition to $\mathrm{D_{RP}}$, we propose a synthetic role-playing data construction approach that transforms general supervised fine-tuning (SFT) data into role-playing style dialogues, resulting in the dataset $\mathrm{D_{RP}^{synth}}$. The construction pipeline consists of three steps: 1) We randomly sample multi-turn dialogues from the Pure-Dove (PD) dataset (Daniele and Suphavadeeprasit, 2023) and sequentially group every 100 consecutive dialogue turns to form speaker-specific clusters. Each cluster containing approximately 100 dialogue turns and is then assigned to represent a distinct synthetic speaker. 2) For each synthetic speaker, we use GPT-4o to analyze the 100 dialogue turns and annotate the speaker's psychological profile according to our 26 quantitative dimensions. 3) Based on the dialogue context, we employ GPT-4o to generate appropriate scene-specific information to enhance contextual coherence. The resulting dataset $\mathrm{D_{RP}^{synth}}$ is formulated as:

$$\mathrm{D_{RP}^{synth}} = \{(R_i, P_i, C_i, x_i, y_i) | i = 1, 2, \ldots, N\}, \quad (2)$$

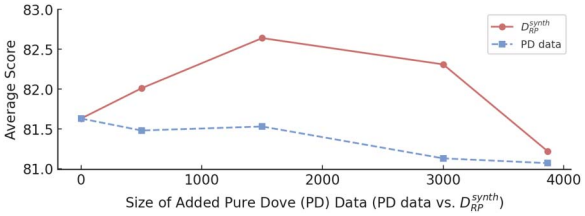

Figure 4: Comparing Original and Synthetic Role-Playing data ($\mathrm{D_{RP}^{synth}}$). PD: Performance by Dataset Size.

where R_i denotes the role responsible for the response, and P_i is the profile of R_i defined by basic information and 26 quantitative psychological dimensions. C_i represents the preceding turns and scene information. x_i and y_i are the current query, and the corresponding response drawn from the Pure-Dove (PD) datasets, respectively.

By incorporating role-playing style general SFT data $\mathrm{D_{RP}^{synth}}$, the proposed PsyMem framework enables the model to develop stronger role-playing capabilities, as Figure 4 shows.

4 Role Playing Training and Evaluation

4.1 Training

In this work, we transfer general LLMs to role-playing models via supervised fine-tuning. Role-playing presents a unique challenge for language models: maintaining both general capabilities and character-specific attributes simultaneously. This challenge exemplifies the fundamental tension in transfer learning between preserving general knowledge and specializing for domain-specific tasks. Traditional fine-tuning approaches (Kotha et al., 2024) often encounter a core multi-task conflict during optimization. In the context of role-playing tasks, this conflict arises from the attempt to optimize two fundamentally distinct objectives—namely, character profile alignment and memory adherence—within a unified framework. Inspired by the Dual-stage Mixed Fine-tuning (DMT) (Dong et al., 2023) method, we introduce a two-stage training strategy for role-playing, effectively balancing the capabilities of character profile alignment and memory adherence.

Specifically, stage 1 establishes core role-playing capabilities by training the model on character profile-based dialogues, enabling the model to learn fundamental character attribute alignment. Stage 2 further trains the model on

memory-augmented role-playing dataset $\mathrm{D_{RP}}$ and role-playing style general fine-tuning dataset $\mathrm{D_{RP}^{synth}}$ to improve role-playing precision, as illustrated in Figure 3(b). As demonstrated in Section 5.3, our two-stage training strategy significantly outperforms vanilla single-stage training for role-playing performance, validating the necessity of this progressive specialization strategy.

Stage 1: Foundational Role-Playing Capacity Development. In this stage, we train LLMs to learn foundational role-playing abilities based on quantifiable character attributes and dialogue contexts without memory augmentation. We fine-tune the model using a subset $\mathrm{D_{RP1}}$ sampled from the overall role-playing dataset $\mathrm{D_{RP}}$ to establish its fundamental role-playing abilities. The objective function is formally defined as:

$$\theta_1^* = \arg\min_\theta \mathbb{E}_{(\cdot)\sim \mathrm{D_{RP1}}} \left[\mathcal{L}(f_\theta(R, P, C, x), y)\right] \tag{3}$$

where $(\cdot)$ in $\mathrm{D_{RP1}}$ denotes the tuple (R, P, C, x, y) and $\mathrm{D_{RP1}}$ excludes the retrieved memory component M from the original dataset $\mathrm{D_{RP}}$.

Stage 2: Memory-Augmented Role-Playing Specialization. In this stage, we enhance the model's role-playing capabilities through two training components: 1) memory-augmented role-playing data $\mathrm{D_{RP}}$, which incorporates long-term memory contexts retrieved via graph-based methods to enable explicit memory alignment, and (2) synthetic role-playing style data $\mathrm{D_{RP}^{synth}}$, which improves general role-playing abilities. This stage employs a weighted combination of two corresponding loss terms to simultaneously enhance memory-controlled response generation while preserving the model's foundational capabilities:

$$\theta_2^* = \arg\min_\theta \left\{\alpha \mathcal{L}_{\mathrm{RP}}(\theta) + \mathcal{L}_{\mathrm{RP}}^{\mathrm{synth}}(\theta)\right\} \tag{4}$$

where $\alpha = 20$.

The two loss terms are individually defined as follows:

$$\mathcal{L}_{\mathrm{RP}}(\theta) = \mathbb{E}_{(\cdot)\sim \mathrm{D_{RP}}} \left[\mathcal{L}\left(f_\theta(R, P, M, C, x), y\right)\right], \tag{5}$$

$$\mathcal{L}_{\mathrm{RP}}^{\mathrm{synth}}(\theta) = \mathbb{E}_{(\cdot)\sim \mathrm{D_{RP}^{synth}}} \left[\mathcal{L}\left(f_\theta(R, P, C, x), y\right)\right], \tag{6}$$

where $(\cdot)$ in $\mathrm{D_{RP}^{synth}}$ represents (R, P, C, x, y) and $(\cdot)$ in $\mathrm{D_{RP}}$ represents the tuple (R, P, M, C, x, y). M denotes the retrieved memory response for role R regarding the query x and the context C (which includes the preceding dialogue turns and scene information), obtained via graph-based retrieval methods and subsequently filtered by LLMs.

4.2 Evaluation

We evaluate role-playing capabilities of LLMs by adopting the widely used ''LLMs as Judges'' approach[1] (Shao et al., 2023; Lu et al., 2024), as illustrated in Figure 3(c). The evaluation process is structured into two main categories: **Character-independent Capabilities** and **Character Fidelity**.

Character-independent Capabilities. Following previous research (Chen et al., 2024; Yu et al., 2024), we measure Character-independent Capabilities using three metrics: Human-likeness, Consistency, and Coherence. In the Human-likeness task, we use GPT-4o with few-shot examples to determine whether a dialogue was generated by a human, ultimately computing the accuracy score. For consistency, we evaluate the coherence between different parts of the conversation to assess whether the role-playing remains consistent throughout long dialogues. For coherence, we assess how logically connected and contextually appropriate the responses are across the entire dialogue, ensuring the conversation flows naturally.

Character Fidelity. Character Fidelity assesses how accurately an LLM portrays specific role-playing characters. As described in Section 3.1, we employ modern psychological quantification frameworks defined in our dataset design, covering five primary dimensions: Personality (Big Five Personality Model), Values (Schwartz's Theory of Basic Values), Social and Leadership Styles (Zimbardo's psychological principles), Behavioral Decision (Thomas-Kilmann Conflict Mode Instrument, TKI), and Memory. The former four dimensions are quantified into numeric scales ranging from 1 to 5 using standardized psychological evaluation methods. The Memory dimension, distinct from the psychological

[1]We validate the reliability of our LLM-based evaluation framework by comparing human-evaluation and GPT-4o judgments in the Appendix (see Table 8).

attributes, is evaluated using accuracy scores from true/false assessments.

Evaluation Process. The evaluation process commences with models generating 15-turn dialogues per scenario (see Section 5.1 for scenario detail), embodying the specified character profile. *For character-independent capabilities evaluation*, we utilize these generated dialogues: **1)** Human-likeness and Coherence are assessed by GPT-4o (true/false based on rules), yielding a positive assessment rate (e.g., the percentage deemed human-like or coherent); **2)** Consistency is measured by the score difference in the four psychological dimensions between dialogue halves, yielding a stability score (e.g., where a smaller difference yields a higher consistency score, scaled from 0 to 1). *For character fidelity evaluation*, GPT-4o serves as the judge, assessing each 15-turn dialogue. Performance is scored across five dimensions: **1)** The first four dimensions (Personality, Values, etc.) are evaluated by GPT-4o on dialogues, using 1-NED (introduced in the following paragraph) for performance; **2)** For the fifth dimension, memory alignment, we use GPT-4o to judge response correctness (true/false) against memory context and reference answer, yielding an accuracy score (i.e., the proportion of correctly aligned responses).

Metrics. We quantify role-playing fidelity for Personality, Values, Social & Leadership, and Behavioral Decision using the 1-NED metric. Derived from Normalized Euclidean Distance (NED), a higher 1-NED score indicates greater similarity to the target profile. The 1-NED score is computed as:

$$\mathbf{1-NED} = 1 - \frac{1}{\Delta X}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_{\text{ref}})^2}, \tag{7}$$

where $\Delta X = X_{\max} - X_{\min}$, with $X_{\max}$ and $X_{\min}$ being the maximum (5) and minimum (1) values of the evaluation scale, respectively. A score of 0, used to denote uncertainty, is excluded from this range. n is the number of subdimensions, X_i represents the measurement for the i-th subdimension, and X_{ref} is the corresponding value contained in role profile P_i (see Eq. (1) and Eq. (2)), with possible values from $\{1, 2, 3, 4, 5\}$. This formulation is robust, as it normalizes the mean squared deviation across different scales, yielding a more interpretable similarity metric.

5 Experiments

In this section, we explain the dataset, baselines, and implementation, and discuss the main results and ablation experiments.

5.1 Experimental Setup

Evaluation Dataset. To mitigate potential knowledge bias introduced during the pretraining phase, we construct an evaluation dataset $\text{D}_{\text{RP}}^{\text{eval}}$ for our main experiments by selecting characters and scenes from novels published after June 2024. This design choice aims to ensure that models have had minimal or no exposure to the selected texts during pretraining, thereby reducing reliance on memorized character-specific knowledge.

We first assess 100 contemporary novels across diverse genres, then select 25 high-quality novels based on character complexity, dialogue richness, expression quality, and narrative perspective. From each selected novel, we identify the two most psychologically complex and representative characters, yielding 50 characters with varied psychological profiles. For each character, we extract ten dialogue-rich scenes directly from the original novels, resulting in 500 evaluation scenarios (50 characters $\times$ 10 scenes), which constitutes our evaluation set $\text{D}_{\text{RP}}^{\text{eval}}$. The character attributes are derived from a detailed analysis of the novel using GPT-4o. This comprehensive evaluation framework enables robust assessment of role-playing capabilities across diverse character types and narrative contexts. For the memory evaluation, each scenario includes not only the character's profile and scene information, but also three key components: a meticulously selected character memory excerpt from the memory text, a memory-dependent question, and its authentic answer serving as the reference answer, allowing us to assess all character fidelity dimensions.

Baselines. We evaluate PsyMem against open-source and proprietary state-of-the-art chatbots. For general-purpose models, we compared the proposed PsyMem with several prominent baselines, including GPT-4o (Hurst et al., 2024), GPT-3.5-turbo (OpenAI, 2023), Claude-3.5-sonnet (Anthropic, 2024b), Qwen-Max (Team, 2024), Yi-Large (Young et al., 2024), Deepseek-R1 (Guo et al., 2025), DeepSeek-V3 (Liu et al., 2024), LLaMA3.1-8B-instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct (Yang et al., 2024). These models are considered some of the best in the

field, offering a broad range of capabilities across various domains. In the category of Role-Playing-focused baselines, we made comparisons with Hunyuan-Role (Tencent Cloud, 2023), Baichuan-NPC-Turbo (Baichuan AI, 2023), and CharacterGLM-6B (Zhou et al., 2023), Higgs-LLaMA3-70B (AI, 2024), and CoSER-70B (Wang et al., 2025), which are specifically designed to handle character-driven interactions and story-based tasks.

Implementations. With the proposed PsyMem framework, we fine-tune LLaMA3.1-8B-Instruct and Qwen2.5-7B-Instruct for three epochs, and denote the resulting models as PsyMem-LLaMA and PsyMem-Qwen, respectively. Additional training settings and implementation details are provided in the Appendix B.2. Evaluations are conducted according to Section 4.2.

5.2 Main Results

Table 2 presents the main results. Within the General Baselines, we observe that proprietary models commonly outperform open-source models. Claude-3.5-sonnet and Yi-Large, as outstanding proprietary models, show impressive performance in Character-independent Capabilities. Additionally, Baichuan-NPC-Turbo, a role-playing baseline, surpasses general baselines in self-consistency, crucial for role-playing. Furthermore, it is worth mentioning that DeepSeek-R1 and CoSER-70B excel in Character Fidelity.

Empirical results under the PsyMem framework reveal that, relative to their original counterparts, PsyMem-LLaMA and PsyMem-Qwen achieve character fidelity improvements of 2.07% and 3.50%, respectively. In addition, PsyMem-Qwen demonstrates strong performance in the Social & Leadership and Behavioral Decision dimensions, achieving scores of 80.80% and 78.48%, respectively. PsyMem-LLaMA, on the other hand, obtains a score of 81.93% in the Value dimension, surpassing all other evaluated models in this category. **Moreover, PsyMem-Qwen, despite having only 7B parameters, surpassed all measured baselines in Character Fidelity.** Additionally, it is worth noting that both models exhibited remarkable improvements in Human-likeness. Overall, the PsyMem framework has demonstrated strong effectiveness in our benchmark tests, enabling precise character control through quantitative character attribute metrics and memory mechanisms.

Model	Per.	Val.	SL*	BD*	Mem.	Avg.
PL (w/o Mem.)	80.11	81.90	**79.23**	**76.36**	44.20	72.36
PL	**80.47**	**81.93**	78.57	74.47	**90.20**	**81.13**
PQ (w/o Mem.)	**81.02**	81.33	**82.10**	**79.02**	53.00	75.29
PQ	80.40	**81.74**	80.80	78.48	**91.80**	**82.64**

Table 3: Ablation studies of memory alignment on Character Fidelity. PL: PsyMem-LLaMA, PQ: PsyMem-Qwen.

5.3 Ablation Study

Memory Alignment Training. To evaluate the effectiveness of memory alignment training, we compare the performance of PsyMem-LLaMA and PsyMem-Qwen with and without memory integration during the second training stage. The results are presented in Table 3. **Our findings indicate that incorporating memory data in the second training stage significantly enhances the model's ability to responding based on retrieved memory, rather than merely summarizing relevant content.** This improves both character attribute consistency and long-term memory retention. However, we also observe that memory alignment training has a slight negative impact on character attribute alignment. We hypothesize that this may be due to the inclusion of a limited number of role-specific dialogue examples in the memory data, which could slightly diminish the model's ability to adhere strictly to predefined character attributes.

Two-stage Training Strategy. To evaluate the DMT strategy, we compare the full PsyMem framework (with DMT) against a single-stage fine-tuning baseline jointly optimizing all datasets and loss functions. To rigorously assess the improvement in Character Fidelity, we designated the ''Average'' score as the primary outcome (Cronbach and Meehl, 1955), treating sub-dimensions as secondary (Feise, 2002; DeVellis and Thorpe, 2021). We used paired two-sided t-tests with Benjamini-Hochberg (BH) correction (Benjamini and Hochberg, 1995) on sub-dimensions to ensure robustness without being overly conservative (Storey, 2002; Yekutieli, 2008) (see verification in Appendix E). Table 4 shows DMT yields statistically significant gains

Model	Per.	Val.	SL*	BD*	Mem.	Avg.
PL (w/o DMT)	79.91±0.34	81.38±0.12	77.88±0.42	74.55±0.34	89.33±0.31	80.61±0.12
PL	**80.42±0.26***	**81.98±0.24***	**78.62±0.14**	**74.64±0.18**	**90.27±0.12***	**81.19±0.06**†
PQ (w/o DMT)	79.60±0.22	81.47±0.30	80.32±0.19	77.93±0.30	90.33±0.12	81.93±0.13
PQ	**80.46±0.18***	**81.75±0.24***	**80.85±0.24***	**78.51±0.29***	**91.80±0.20***	**82.67±0.16**†

Table 4: Ablation studies of DMT on Character Fidelity (Mean ± Std over 3 random seeds). PL: PsyMem-LLaMA, PQ: PsyMem-Qwen. †: Statistical significance on the primary Avg. metric (paired two-sided t-tests, $p < 0.05$). *: Significance on sub-dimensions (paired two-sided t-tests, BH corrected, $p < 0.05$).

Model	Character Fidelity	Human -likeness	Consis- tency	Coher- ence
LLaMA3.1-8B-Instruct	79.06	64.20	88.74	**99.00**
+ PsyMem (PD data)	80.51	70.80	89.49	82.20
+ PsyMem (D_{RP}^{synth})	**81.13**	**85.20**	**89.93**	95.60
Qwen2.5-7B-Instruct	79.14	64.40	89.58	**97.40**
+ PsyMem (PD data)	81.63	86.60	89.45	93.40
+ PsyMem (D_{RP}^{synth})	**82.64**	**87.60**	**89.64**	96.20

Table 5: Ablation study on Synthetic Role-playing style (D_{RP}^{synth}) Pure Dove (PD) dataset in PsyMem.

on the primary metric across both backbones. Specifically, PsyMem-Qwen achieves a 0.74% improvement (82.67% vs. 81.93%), and PsyMem-LLaMA improves by 0.58% (81.19% vs. 80.61%). Furthermore, 8 of the 10 sub-dimension comparisons remained significant after correction. **This provides compelling empirical validation that the two-stage strategy substantially improves character fidelity.**

Synthetic Role-playing Data. We investigate the impact of Synthetic Role-playing style SFT Data D_{RP}^{synth} on model performance through ablation studies presented in Table 5 and Figure 4. Table 5 shows that compared to general SFT data, **D_{RP}^{synth} significantly improves Character Fidelity, Human-likeness, and Consistency for both PsyMem-LLaMA and PsyMem-Qwen**, while preserving coherence.

Figure 4 further examines the effect of data volume, comparing General SFT Data with D_{RP}^{synth} on role-playing performance. The results validate the superiority of our synthetic approach: while General SFT Data consistently degrades character-specific abilities, Synthetic Role-playing style data achieves optimal performance at 1,500 samples before declining at larger volumes. We attribute this performance

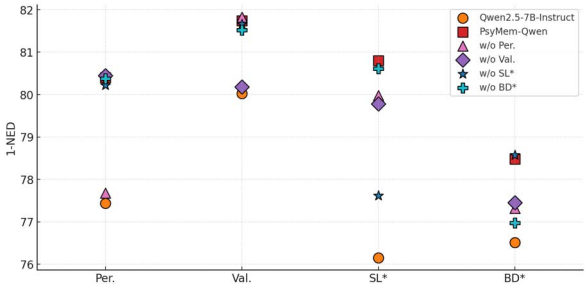

Figure 5: Ablation study of each dimensions. Per.: Personality, Val.: Values, SL*: Social & Leadership, BD*: Behavioral Decision, Mem.: Memory.

drop at 3,000 samples to dataset distribution imbalance, where excessive synthetic content may dilute role-specific attribute diversity.

Character Attribute Dimension. We assess each attribute's contribution by individually removing the four supervision signals (Per., Val., SL*, and BD*) in an ablation study on PsyMem-Qwen (see Figure 5). The results reveal that removing any dimension markedly drops performance on its corresponding evaluation metric. This indicates each dimension is critical to the model's fidelity for that trait, underscoring the necessity of the proposed comprehensive attribute set. We place more ablation experiments on dimensions in Appendix D.

6 Conclusion

We present PsyMem, a novel role-playing framework that synergizes character memory with fine-grained psychological attributes. Evaluations on a large-scale novel-based dataset demonstrate PsyMem's effectiveness. The resulting PsyMem-Qwen, with only 7B parameters, outperforms all baselines in character fidelity while demonstrating strong character-independent capabilities. We also highlight the benefits of integrating Synthetic Role-playing General SFT Data. We believe this work paves the way for psychologically-informed role-playing models, fostering more nuanced, realistic, and consistent character portrayals and expanding LLM utility in trust-worthy social simulation, human-computer interaction, and AI counseling.

Acknowledgments

We sincerely thank the TACL action editor, Giuseppe Carenini, for his careful coordination of the review process and for consolidating the

reviewers' insightful and constructive comments. We are also deeply grateful to all reviewers for their professional and valuable feedback, which has greatly improved the quality and clarity of this paper.

This work is supported in part by the National Natural Science Foundation of China (no. 62206259), in part by the Fundamental Research Funds for the Central Universities (no. CUC25SG005), and in part by the National Key Research and Development Program of China (no. 2024YFF0907200).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Boson AI. 2024. Announcing the higgs family of LLMs.

Gordon Willard Allport. 1937. *Personality: A Psychological Interpretation*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku. Accessed: 2025-02-14.

Anthropic. 2024b. Claude 3.5 sonnet.

Baichuan AI. 2023. Baichuan NPC platform. Accessed: 2025-02-16.

John A. Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype

activation on action. *Journal of Personality and Social Psychology*, 71(2):230–244. https://doi.org/10.1037/0022-3514.71.2.230, PubMed: 8765481

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188. https://doi.org/10.1214/aos/1013699998

BookLikes. 2024. Booklikes – a community for book lovers. Accessed: 2025-06-22.

Lex Borghans, Angela Lee Duckworth, James J. Heckman, and Bas Ter Weel. 2008. The economics and psychology of personality traits. *Journal of Human Resources*, 43(4):972–1059. https://doi.org/10.3386/w13810

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet Harry Potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520. https://doi.org/10.18653/v1/2023.findings-emnlp.570

Paul T. Costa and Robert R. McCrae. 2008. The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*, 2(2):179–198. https://doi.org/10.4135/9781849200479.n9

Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302. https://doi.org/10.1037/h0040957, PubMed: 13245896

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, Yonghong Tian, and Li Yuan. 2023. Machine mindset: An MBTI exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Luigi Daniele and Suphavadeeprasit. 2023. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient LLM training.

DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf. Technical report.

Robert F. DeVellis and Carolyn T. Thorpe. 2021. *Scale development: Theory and applications*. Sage publications.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64. https://doi.org/10.1080/01621459.1961.10482090

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Howard Eichenbaum. 2015. The hippocampus as a cognitive map. . . of social space. *Neuron*, 87(1):9–11. https://doi.org/10.1016/j.neuron.2015.06.013, PubMed: 26139366

Ronald J. Feise. 2002. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, 2(1):8. https://doi.org/10.1186/1471-2288-2-8, PubMed: 12069695

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Goodreads. 2024. Goodreads. Accessed: 2025-06-22.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*. https://doi.org/10.1038/s41586-025-09422-z, PubMed: 40962978

Tom Hartley, Colin Lever, Neil Burgess, and John O'Keefe. 2014. Space in the brain: How the hippocampal formation supports spatial cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):20120510. https://doi.org/10.1098/rstb.2012.0510, PubMed: 24366125

Jan Hauke and Tomasz Kossowski. 2011. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2):87–93. https://doi.org/10.2478/v10117-011-0021-1

John Hunsley, Catherine M. Lee, James M. Wood, and Whitney Taylor. 2015. Controversial and questionable assessment techniques. In S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.), *Science and Pseudoscience in Clinical Psychology* (2nd ed., pp. 42–82). The Guilford Press.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex

Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, et al. 2024. GPT-4o system card. *arXiv preprint arXiv: 2410.21276*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121. https://doi.org/10.1073/pnas.2405460121, PubMed: 39471222

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to its Methodology*. Sage Publications. https://doi.org/10.4135/9781071878781

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.423

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam,

Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv: 2005.14165*, 1.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–Krippendorff's alpha calculator: A user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545. https://doi.org/10.1016/j.mex.2023.102545, PubMed: 39669968

Walter Mischel. 2013. *Personality and Assessment*. Psychology Press. https://doi.org/10.4324/9780203763643

K. Nowack. 1996. Is the myers briggs type indicator the right tool to use. *Performance in Practice*, 6.

OpenAI. 2023. Gpt-3.5 turbo documentation. Accessed: 2025-02-16.

OpenAI. 2025. Introducing openAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744. https://doi.org/10.52202/068431-2011

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. https://doi.org/10.1145/3586183.3606763

Thomas V. Perneger. 1998. What's wrong with Bonferroni adjustments. *BMJ*, 316(7139): 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236, PubMed: 9553006

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190. https://doi.org/10.1017/S0140525X04000056, PubMed: 15595235

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331. https://doi.org/10.1162/tacl_a_00605

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.853

Elizabeth C. Ravlin and Bruce M. Meglino. 1987. Effect of values on perception and decision making: A study of alternative work values measures. *Journal of Applied Psychology*, 72(4):666. https://doi.org/10.1037/0021-9010.72.4.666

Brent W. Roberts, Dustin Wood, and Avshalom Caspi. 2008. The development of personality traits in adulthood. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research* (3rd ed., pp. 375–398). The Guilford Press.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*. Academic Press. https://doi.org/10.1016/S0065-2601(08)60281-6

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

https://doi.org/10.1038/s41586-023-06647-8, PubMed: 37938776

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Mark Snyder. 1983. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of Personality*, 51(3):497–516. https://doi.org/10.1111/j.1467-6494.1983.tb00342.x, PubMed: 28497608

Randy Stein and Alexander B. Swan. 2019. Evaluating the validity of myers-briggs type indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, 13(2):e12434. https://doi.org/10.1111/spc3.12434

John D. Storey. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498. https://doi.org/10.1111/1467-9868.00346

Stephen A. Stumpf and Roger L. M. Dunbar. 1991. The effects of personality type on choices made in strategic decision situations. *Decision Sciences*, 22(5):1047–1072. https://doi.org/10.1111/j.1540-5915.1991.tb01906.x

Pawel Tacikowski, Güldamla Kalender, Davide Ciliberti, and Itzhak Fried. 2024. Human hippocampal and entorhinal neurons encode the temporal structure of experience. *Nature*, 635(8037):160–167. https://doi.org/10.1038/s41586-024-07973-1, PubMed: 39322671

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2023. Rolecraft-GLM: Advancing personalized role-playing in large language models. *arXiv preprint arXiv:2401.09432*.

Rita Morais Tavares, Avi Mendelsohn, Yael Grossman, Christian Hamilton Williams, Matthew Shapiro, Yaacov Trope, and Daniela Schiller. 2015. A map for social navigation in the human brain. *Neuron*, 87(1):231–243. https://doi.org/10.1016/j.neuron.2015.06.011, PubMed: 26139376

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Tencent Cloud. 2023. Document for tencent cloud product 1729. Accessed: 2025-02-16.

The StoryGraph. 2024. The storygraph. Accessed: 2025-06-22.

Kenneth W. Thomas. 2008. Thomas-kilmann conflict mode. *TKI Profile and Interpretive Report*, 1(11).

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. 2025. Coser: Coordinating LLM-based persona simulation of established roles.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2023b. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al.

2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Gu Ye. 2024. nano-graphrag. https://github.com/gusye1234/nano-graphrag. Accessed: 2025-06-22.

Daniel Yekutieli. 2008. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316. https://doi.org/10.1198/016214507000001373

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01. AI. *arXiv preprint arXiv:2403.04652*.

Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. XDAI: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4422–4432. https://doi.org/10.1145/3534678.3539135

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2024. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. *arXiv preprint arXiv:2408.10903*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*. https://doi.org/10.18653/v1/2024.findings-acl.830

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731. https://doi.org/10.1609/aaai.v38i17.29946

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. CharacterGLM: Customizing chinese conversational AI characters with large language models. *arXiv preprint arXiv:2311.16832*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents.

Philip G. Zimbardo and Floyd L. Ruch. 1975. *Psychology and Life*.

Appendix

A Source Dataset

We compiled an initial pool of 1,710 English-language novels from The StoryGraph (The StoryGraph, 2024), Goodreads (Goodreads, 2024), and BookLikes (BookLikes, 2024). Using an LLM, we filtered these down to 539 novels based on four criteria: **1) fluency and reader-friendliness, 2) third-person narration, 3) clear formatting with well-structured layout, and 4) balanced genre coverage for diversity.**

B Implementation Details

B.1 Memory Module

To support contextual memory retrieval during role-playing, we employed Nano-GraphRAG (Ye, 2024), a lightweight implementation of GraphRAG (Edge et al., 2024). Since our test set consists of recently published novels, we utilize this module during inference to enrich context. We configured the local mode with a paragraph length of 8192 tokens, keeping other settings default. In each turn, the system retrieves role-aligned memories based on entities in the prior context and combines them with character attributes to guide generation.

Model	Character Fidelity					
	Per.	Val.	SL*	BD*	Mem.	Avg.
PsyMem-LLaMA (Rewrite)	80.14	81.47	**78.62**	74.35	89.80	80.88
PsyMem-LLaMA (Random)	80.33	81.58	78.37	**74.56**	**90.20**	81.00
PsyMem-LLaMA (Original)	**80.47**	**81.93**	78.57	74.47	**90.20**	**81.13**
PsyMem-Qwen(Rewrite)	**80.70**	81.36	80.20	78.42	91.00	82.34
PsyMem-Qwen(Random)	80.45	81.63	**81.02**	78.39	**92.00**	**82.70**
PsyMem-Qwen(Original)	80.40	**81.74**	80.80	**78.48**	91.80	82.64

Table 6: Evaluation of the model's generalization to input with freer phrasing (Rewrite), shuffled dimension order (Random), and prompts consistent with those used during training (Original).

B.2 Training Setup

We fine-tuned Qwen2.5-7B-Instruct and LLaMA3.1-8B-Instruct on four A800 80G GPUs using LoRA. Training spanned 3 epochs with a global batch size of 128 (per-device batch size 2, 16 gradient accumulation steps) and a learning rate of 1e-4 (cosine schedule, 0.1 warmup). The sequence length was set to 8,196, extended to 16,384 in the second stage. To enhance robustness, we applied data augmentation with a 50% probability of shuffling or masking role attributes for each instance.

C Additional Evaluation

C.1 More Flexible Inputs

To assess PsyMem's robustness to varied prompt styles, we tested three evaluation settings: the original structured prompt ('Original'), prompts with randomly shuffled character dimensions ('Random'), and character profiles rewritten into natural paragraphs by GPT-4o ('Rewrite'). Both PsyMem-LLaMA and PsyMem-Qwen were evaluated across five Character Fidelity dimensions.

As shown in Table 6, the 'Random' variant achieves comparable performance to the 'Original' version, suggesting that the model is robust to prompt order perturbations (likely a result of training-time augmentation). The 'Rewrite' variant results in a slight performance decrease (0.25% for LLaMA, 0.30% for Qwen), potentially due to minor information loss during rephrasing. Nonetheless, consistently high Memory scores across all variants demonstrate the robustness and generalization ability of PsyMem under varying input conditions.

Model	GPT-4o	o3	Gemini-2.5-pro
GPT-4o	80.86	79.62	76.92
+ Concatenated Eval.	81.00	79.39	76.70
Deepseek-R1	82.47	78.40	75.87
+ Concatenated Eval.	82.26	78.66	75.99
CoSER-70B	82.28	80.23	79.16
+ Concatenated Eval.	81.99	80.40	78.99
PsyMem-Qwen	82.64	**80.74**	**79.70**
+ Concatenated Eval.	**82.87**	80.43	79.39

Table 7: Average Character Fidelity scores (Per., Val., SL*, BD*, Mem.) of advanced role-playing models evaluated by different discriminative models and evaluation approaches. ''Concatenated Eval.'' refers to presenting all dialogues across scenarios simultaneously to the discriminative model for evaluation.

C.2 Evaluation Strategies

To verify the robustness of our evaluation methodology, we compared two scoring strategies: one strategy where each character's dialogues across different scenes were first concatenated and then assessed collectively, and another strategy where each character's dialogue, consisting of 15 turns per scene, was evaluated separately before being aggregated. We applied both strategies to four top-performing models, using three discriminative evaluators—GPT-4o (Hurst et al., 2024), OpenAI-o3 (OpenAI, 2025), and Gemini-2.5-pro (DeepMind, 2025). As shown in Table 7, results across all models and evaluators exhibit strong consistency between the two strategies. Notably, for PsyMem-Qwen under GPT-4o evaluation, the average Character Fidelity score under concatenated dialogue input was 82.87%, only 0.23% points higher than the segment-wise evaluation score of 82.64%, indicating that the choice of evaluation strategy does not significantly impact overall conclusions.

C.3 Human Evaluation

To validate our LLM-based evaluation framework, we conducted a large-scale human study involving 40 volunteers who assessed 25 dialogue scenarios (50 characters). The process was highly time-intensive, requiring over 10 hours per rater. For a robust analysis, we employed several complementary metrics. We used Krippendorff's Alpha (α) to measure inter-rater reliability among human annotators, thereby establishing the reliability of our evaluation baseline. Concurrently, we

Metrics	Per.	Val.	SL*	BD*
Cosine (H ↔ L)	0.96	0.95	0.96	0.94
r_s (H)	0.74	0.79	0.75	0.77
ρ (H)	0.72	0.71	0.71	0.70
Mean (Std) (H)	0.59 ± 0.47	0.46 ± 0.42	0.42 ± 0.39	0.54 ± 0.41
α (H)	0.77	0.80	0.74	0.80

Table 8: **Cosine** similarity, Pearson correlation coefficient (ρ), Spearman's correlation coefficient (r_s), and **Mean (Std)** are used to assess the agreement and score distribution between human raters and GPT-4o across different evaluation dimensions. For reference, Krippendorff's Alpha (α) is also reported as a measure of inter-rater reliability among human annotators.

Prompt Set	Per.	Val.	SL*	BD*
All	**80.40**	**81.74**	**80.80**	**78.48**
Per.	<u>78.35</u>	73.33	74.72	73.67
Val.	74.74	<u>79.66</u>	74.35	74.39
SL*	68.12	67.86	<u>79.64</u>	69.35
BD*	69.38	68.77	70.14	<u>77.76</u>
Base	66.01	65.65	68.05	67.19

Table 9: Performance of PsyMem-Qwen with different prompt combinations across four evaluation dimensions. ''All'' indicates inclusion of all prompts for the four dimensions; ''Base'' indicates exclusion of these prompts. **Bold**: Best; Underlined: Second best.

assessed the agreement between GPT-4o's scores and the mean human scores using Cosine Similarity, Pearson's correlation coefficient (ρ), and Spearman's rank correlation coefficient (r_s).

The results, summarized in Table 8, validate our approach. We first confirmed a reliable human baseline, with Krippendorff's Alpha (α) values of 0.74–0.80 indicating substantial inter-rater reliability (Krippendorff, 2018; Marzi et al., 2024). Against this baseline, GPT-4o's scores showed strong agreement with human ratings. Pearson's correlation (ρ) was consistently strong (0.70–0.72), and Spearman's rank correlation (r_s) was similarly high (0.74–0.79) (Fleiss, 1971; Hauke and Kossowski, 2011). These results collectively demonstrate that GPT-4o serves as a valid proxy for human evaluation in our task.

D Evaluation-stage Input Ablation

We investigated the inference-time role of character attributes by comparing three input formats (Table 9): (1) a minimal setting with no character dimensions, (2) inputs with only one active dimension, and (3) full inputs containing all four dimensions.

The model's performance deteriorated markedly when all dimensions were removed, with average scores dropping to 66.01 on Personality and 65.65 on Values. In contrast, activating even a single dimension improved performance on the corresponding axis: the Personality-only prompt recovered the Personality score to 78.35, while Values alone reached 79.66. Additionally, we observed that providing Personality and Values information boosted performance on Social & Leadership and Behavioral Decision dimensions (74.72 and 73.67 under Personality-only input), while the reverse was not true. This suggests a hierarchy where latent traits foundationally guide explicit behavioral patterns.

E Verification of Statistical Significance

As the pre-defined primary outcome, the ''Average'' score represents a single hypothesis test; thus, we apply no correction ($p < 0.05$), aligning with standard practices (Feise, 2002). We use the random seed as the unit of pairing, performing paired two-sided t-tests by matching runs with the same seed across conditions. For the sub-dimensions, treated as secondary outcomes, we address the issue of multiple comparisons by applying the Benjamini-Hochberg (BH) procedure to control the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). We selected BH over strict family-wise error rate corrections (e.g., Bonferroni (Dunn, 1961)) because the latter are known to be overly conservative for inter-dependent psychological metrics (e.g., personality traits influencing behavioral patterns, see Figure 5) (Allport, 1937; Mischel, 2013). In contrast, BH offers a more powerful and appropriate approach for correlated data (Perneger, 1998; Benjamini and Yekutieli, 2001).