

# Memory-Driven Role-Playing: Evaluation and Enhancement of Persona Knowledge Utilization in LLMs

Kai Wang<sup>†</sup>, Haoyang You<sup>†</sup>, Yang Zhang<sup>‡</sup>, Zhongjie Wang<sup>†\*</sup>,

<sup>†</sup>Harbin Institute of Technology, China

<sup>‡</sup>Macquarie University, Australia

<sup>†</sup>kai\_wang@hit.edu.cn, yebai467@gmail.com, rainy@hit.edu.cn

<sup>‡</sup>yang.zhang@mq.edu.au

## Abstract

A core challenge for faithful LLM role-playing is sustaining consistent characterization throughout long, open-ended dialogues, as models frequently fail to recall and accurately apply their designated persona knowledge without explicit cues. To tackle this, we propose the Memory-Driven Role-Playing paradigm. Inspired by Stanislavski’s “emotional memory” acting theory, this paradigm frames persona knowledge as the LLM’s internal memory store, requiring retrieval and application based solely on dialogue context, thereby providing a rigorous test of depth and autonomous use of knowledge. Centered on this paradigm, we contribute: (1) MREval, a fine-grained evaluation framework assessing four memory-driven abilities—Anchoring, Selecting, Bounding, and Enacting; (2) MR-Prompt, a prompting architecture that guides structured memory retrieval and response generation; and (3) MRBench, a bilingual (Chinese/English) benchmark for fine-grained diagnosis. The novel paradigm provides a comprehensive diagnostic for four-stage role-playing abilities across 12 LLMs. Crucially, experiments show that MRPrompt allows small models (e.g., Qwen3-8B) to match the performance of much larger closed-source LLMs (e.g., Qwen3-Max and GLM-4.7), and confirm that upstream memory gains directly enhance downstream response quality, validating the staged theoretical foundation. Our code is available at <https://github.com/41522/MDRP>.

## 1 Introduction

The role-playing capabilities of large language models (LLMs) are attracting significant interest, enabling applications that range from interactive game characters and personalized virtual companions to simulated assistants with defined personas (Ran et al., 2025; Chen et al., 2025a;

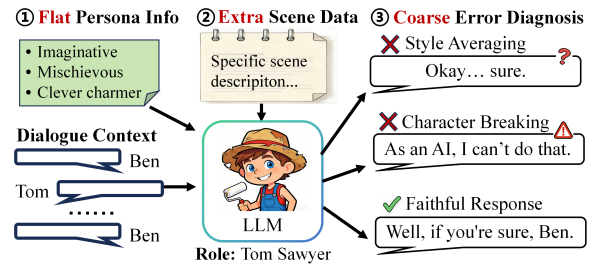


Figure 1: Three Issues in LLM Role-Playing Paradigm

Qi et al., 2025; Tu et al., 2023). In these scenarios, success hinges on the LLM’s ability to remain strictly *in character*. This entails generating responses that are coherent, human-like, and faithful to the designated persona, without reverting to generic patterns or unrelated characters (Wang et al., 2024b; Yu et al., 2025a; Zhou et al., 2025a; Ruangtanusak et al., 2025). Thus, the key is the sustained and faithful application of predefined persona knowledge within dynamic, open-ended dialogues (El Boudouri et al., 2025; Ji et al., 2025).

However, as illustrated in Figure 1, existing role-playing paradigms encounter three major issues in utilizing persona knowledge effectively: (1) **Flat Persona Info**: Personas are often representationally flat, listing traits without contextual expression. This lack of guidance leads LLMs to average across persona facets into generic replies and to drift locally out of character (Zhang et al., 2018; Li et al., 2023; Shin et al., 2025; Yu et al., 2025b). (2) **Extra Scene Data**: Response generation relies on explicit, extra dialogue-scene descriptions, which simplifies reasoning but inflates success; models fail to generalize to real interactions lacking such cues (He et al., 2025; Zhang et al., 2025b), creating misleading benchmarks. (3) **Coarse Error Diagnosis**: Holistic scoring aggregates performance into a single metric, obscuring failure modes and hindering attribution of issues (e.g., persona misalignment vs. context misunderstanding) (Wang et al., 2025a; Ran et al., 2025; Tang et al., 2025b; Bao et al., 2025).

\*Corresponding author.

To address these issues, we draw inspiration from a foundational performance theory: *Stanislavski’s system of emotional memory* (Stanislavski, 1989), which holds that authentic embodiment emerges when an actor recalls experiential memories through sensory details, rather than performing emotions. Analogously, we argue that faithful LLM role-playing requires a contextual memory recall process. We operationalize this as a new **Memory-Driven Role-Playing (MDRP)** task, which requires the model to (i) treat persona knowledge as a long-term memory store, and (ii) retrieve from it using only dialogue context (short-term memory). MDRP thus serves as a targeted probe: it tests whether personas are encoded with sufficient depth for specific recall, and whether that recall can occur autonomously, without extra scene prompts.

To both evaluate and enhance LLMs under the MDRP task, we make a series of interconnected technical contributions:

First, we introduce a fine-grained evaluation framework **MREval** that decomposes faithful role-playing into four measurable memory-driven abilities: (a) Anchoring: accurate retention of persona knowledge; (b) Selecting: retrieving relevant facets given dialogue cues; (c) Bounding: adhering to the knowledge’s constraints; and (d) Enacting: generating natural responses faithful to the recalled knowledge. By quantifying each ability with two metrics, it pinpoints breakdowns to specific stages of memory access and application, exposing weaknesses that traditional holistic metrics miss.

Second, we introduce the **MRPrompt** prompting architecture, implementing the contextual memory recall process required by MDRP. It consists of: (i) Narrative Schema, which structures persona knowledge into a hierarchical and queryable format (e.g., global summary, core traits, and situational facets); and (ii) Magic-If Protocol, which, inspired by Stanislavski’s acting technique, guides the LLM to perform targeted retrieval from this schema and to generate situationally coherent responses.

Finally, we construct **MRBench**, a bilingual MDRP benchmark derived from 10 English and 6 Chinese novels. It enables fine-grained diagnosis of the four memory abilities through systematic control over persona memory and dialogue context. For scalable yet reliable evaluation, we implement an LLM-as-a-Judge procedure, whose scores are calibrated to human ratings via an annotation study.

We conduct a comprehensive evaluation of 12

representative LLMs on the MDRP task using MRBench. This benchmark not only offers a standardized view of memory-driven role-playing but, through its stage-wise design, also diagnoses failures by localizing them to specific memory stages. Beyond diagnostics, a crucial finding is that MRPrompt **empowers smaller open models to compete with larger closed-source ones**. For instance, Qwen3-8B augmented with MRPrompt attains an Avg. Score of 8.12, on par with the much larger GLM-4.7 Base (8.11) and surpassing Qwen3-Max Base (8.08), demonstrating that performance gains can be achieved without scaling the model backbone. This result, supported by granular analyses showing a consistent pipeline effect from memory to enactment, validates the staged nature of memory-driven role-playing.

## 2 Related Work

**Role-Playing Tasks and Benchmarks.** Most role-playing setups condition generation on character profiles from canonical sources and evaluate whether outputs remain *in character*. Early systems ground role-play in extracted dialogues or curated descriptions (Li et al., 2023; Shao et al., 2023), while recent benchmarks broaden role pools and protocols (Wang et al., 2024b; Liu et al., 2024; Tu et al., 2024; Zhou et al., 2025a). However, benchmarks often report aggregate response-level scores, offering limited attribution of failures. In contrast, the proposed MDRP frames role-playing as cue-driven persona memory use under dialogue context, instantiated by MRBench for controlled comparison and paired with a stage-aware evaluation protocol for diagnosis.

**Evaluation Protocols and Diagnostic Metrics.** Role-play evaluation has moved from coarse judgments to more structured protocols for persona fidelity/consistency (Li et al., 2023; Tu et al., 2024), but still largely relies on overall response-level ratings, sometimes with stress tests or judge-centric analyses (Zhou et al., 2025b; Tang et al., 2025b; Zhao et al., 2025). Finer-grained probing methods (e.g., InCharacter) and segment-level OOC detection reveal that holistic scores can mask localized failures (Wang et al., 2024c; Shin et al., 2025), yet they do not attribute errors to distinct stages of persona-memory utilization. We therefore propose MREval, a stage-aware protocol that decomposes MDRP into four sub-abilities with per-stage diagnostic metrics.

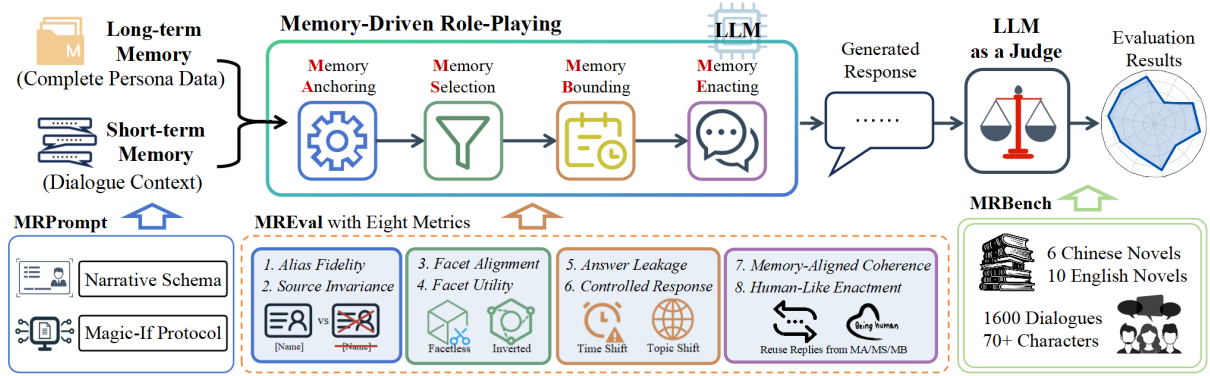


Figure 2: **Overview.** Given two parts of memory, an LLM performs memory-driven role-playing via four stages to generate an in-character response. **MRPrompt** provides structured persona memory and a memory utilization protocol. **MREval** scores eight stage-aligned metrics on the bilingual benchmark **MRBench**, by using an LLM-as-a-judge to assign per-metric scores.

**Role-Playing Methods and Memory-Oriented Mechanisms.** Role-playing controllability is improved via richer persona representations, alignment/adaptation, and prompt-level controllers (Yu et al., 2025b; Lu et al., 2024; He et al., 2025; Wang et al., 2025b; Duan et al., 2025; Ruangtanusak et al., 2025; Tang et al., 2025a), as well as training-based specialization and persona refinement (Yu et al., 2024; Yang et al., 2025; Yao et al., 2025; Fang et al., 2025). Memory-oriented mechanisms further introduce explicit retrieval or long-context organization for sustained role-play (Wang et al., 2025c; Chen et al., 2025b; Huang et al., 2024; Zhang et al., 2025a). Unlike method-centric work, ours is benchmark-centric and diagnostic: we provide MRBench+MREval for stage-wise diagnosis and a prompt-only MRPrompt that standardizes structured persona memory and the reasoning guidance protocol.

### 3 Methodology

#### 3.1 Memory-Driven Role-Playing Task

Given a specified character, we first formalize the Memory-Driven Role-Playing (MDRP) task as:  $\hat{y} \sim p_{\theta}(\cdot | \mathcal{M}_L, \mathcal{M}_S)$ , where  $p_{\theta}$  is the LLM’s distribution parameterized by  $\theta$ . The in-character response  $\hat{y}$  is generated by conditioning on the two memory inputs instantiated as follows:

- **Long-Term Memory (LTM)**  $\mathcal{M}_L$  is a finite set of persona facets:  $\mathcal{M}_L = \{\phi_1, \phi_2, \dots, \phi_N\}$ . Each facet  $\phi_i$  represents a coherent unit of persona knowledge, such as a core character trait together with its context-dependent expressions.
- **Short-Term Memory (STM)**  $\mathcal{M}_S$  is an ordered sequence of the  $K$  most recent dialogue turns:  $\mathcal{M}_S = [u_1, u_2, \dots, u_K]$ . Each turn  $u_i$  is a tuple

$(r_i, c_i)$  consisting of the speaker role  $r_i$  and the utterance content  $c_i$ , and the last turn  $u_K$  is always the interlocutor’s latest utterance (i.e., the model responds next).

This formulation redefines role-playing as a problem of *contextual memory retrieval and application*, wherein the model must utilize the STM context  $\mathcal{M}_S$  to select and apply the relevant knowledge from the LTM persona store  $\mathcal{M}_L$ . Building on this foundation, we develop subsequent diagnostic and prompting methods. An end-to-end overview of the MDRP framework, including its core components MRPrompt, MREval, and MRBench, is provided in Figure 2 and elaborated in the following sections.

#### 3.2 MREval: Evaluation Framework

To enable the diagnostic assessment of LLMs within the Memory-Driven Role-Playing (MDRP) paradigm, we introduce the MREval framework. MREval decomposes the process of in-character response generation into **four sequential, measurable abilities**, corresponding to key stages in the utilization of persona knowledge from a human memory perspective (Baddeley, 1992; Tulving and Thomson, 1973). Grounded in the MDRP formulation, we define four core abilities as follows:

- **Memory-Anchoring (MA):** The model’s ability to anchor its behavior to the designated persona in LTM, rather than relying on superficial cues or pretrained priors.
- **Memory-Selecting (MS):** The model’s ability to extract cues from the STM dialogue context and retrieve relevant persona facets from the LTM based on those cues.
- **Memory-Bounding (MB):** The model’s ability

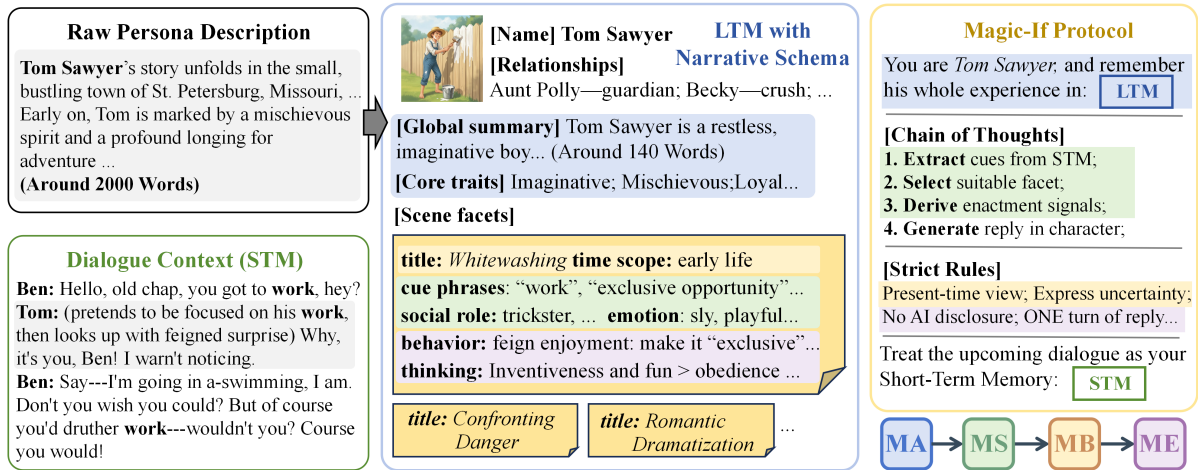


Figure 3: **MRPrompt**. The raw persona description is structured as LTM via Narrative Schema and provided together with STM for role-playing. Magic-If Protocol guides an LLM to generate responses following four stages.

to adhere to the temporal and epistemic boundaries in persona knowledge, preventing the leakage of inaccessible or out-of-scope information.

- **Memory-Enacting (ME):** The model’s ability to transform the selected and bounded persona knowledge into a coherent, natural, and human-like in-character utterance.

Each ability is operationalized by two fine-grained metrics, yielding eight diagnostic scores. Table 1 summarizes all metrics. Concretely, for each metric, we score model outputs with an LLM-as-a-judge and then linearly calibrate the judge scores to the *human rating scale* (see Appendix K), yielding an ordinal 1–10 Likert-style rating for the corresponding criterion (Likert, 1932). The full scoring rubrics are provided in Appendix P.

Overall, MREval provides an eight-dimensional diagnostic profile for MDRP by decomposing in-character generation into four abilities. This enables stage-wise attribution of failures—from grounding and facet retrieval to boundary control and final enactment—rather than relying on a single holistic quality score.

### 3.3 MRPrompt: Inference-Time Method

#### Intuition: Stanislavski-inspired Memory Recall.

Under MDRP, faithful role-playing is not only a matter of *style* but a problem of *contextual memory recall*. Stanislavski’s system emphasizes that authentic embodiment arises from recalling *experiential memory* under the current *given circumstances*, guided by an explicit rehearsal-time *action plan* (e.g., “magic if”) rather than unconstrained improvisation (see Appendix B). Analogously, as illustrated in Figure 3, MRPrompt equips an LLM

with (i) a structured persona memory store as *Long-Term Memory (LTM)* and (ii) an explicit inference-time protocol that extracts cues from the ongoing dialogue and recalls persona memory for faithful role-playing. MRPrompt is purely prompt-based at inference time, requiring no parameter updates and no external retrieval or tool use, and thus can be directly applied to any instruction-tuned LLM.

#### (1) Narrative Schema: hierarchical, queryable persona LTM.

MRPrompt first replaces flat trait lists with a structured Narrative Schema for LTM  $\mathcal{M}_L$ . The schema organizes persona information into *identity fields*, a *global summary*, *core traits*, and a set of *scene facets* that encode context-dependent expressions under recurring situations. Each facet is *cue-addressable*: it binds cue keys (e.g., *situation*, *cue\_phrases*) to enactment signals (e.g., *social\_role*, *emotional\_state*, *behavior\_pattern*, *thinking\_pattern*) and boundary anchors (e.g., *time\_scope*, *conflict\_with\_core*). Importantly, the fields are designed to align with our diagnostic abilities: core traits support MA (anchoring to persona semantics beyond name priors); cue keys support MS (selecting the relevant facet under dialogue cues); boundary anchors support MB (enforcing temporal/epistemic constraints); and enactment signals support ME (realizing recalled knowledge into natural utterances). This schema mitigates style averaging and local out-of-character drift by enabling *selective facet activation*, and makes memory use attributable to concrete fields (schema in Appendix D).

#### (2) Magic-If Protocol: an explicit LTM–STM control protocol.

Building on the Narrative

Table 1: **MREval abilities and metrics.** MREval decomposes memory-driven role-playing into four abilities (MA/MS/MB/ME), each measured by two calibrated Likert-style metrics (eight in total).

Ability	Metric	Definition
MA	Source Invariance (SI)	Measures consistency between the response generated with the anonymized persona $\hat{y}^{\text{anon}}$ and the original one $\hat{y}$ . A high score indicates grounding in persona semantics, not name priors.
	Alias Fidelity (AF)	Assesses whether the behavior under an anonymized persona $\hat{y}^{\text{anon}}$ remains faithful to the original intended character, using the ground-truth response $\hat{y}^{\text{gold}}$ as an anchor.
MS	Facet Alignment (FA)	Quantifies the model’s precision in selecting the correct scene facet by contrasting responses under the true $\mathcal{M}_L$ versus a counterfactual (inverted) LTM $\mathcal{M}_L^{\text{anti}}$ .
	Facet Utility (FU)	Measures the improvement gained by including scene-specific facets in the LTM, compared to a scene-ablated LTM $\mathcal{M}_L^{\text{no-scene}}$ .
MB	Answer Leakage (AL)	Scores the model’s ability to avoid generating a forbidden reference answer $\hat{y}^{\text{out}}$ when presented with an out-of-scope prompt $c_K^{\text{out}}$ (e.g., a future plot spoiler).
	Controlled Response (CR)	Assesses the appropriateness of the model’s response strategy to out-of-scope prompts, favoring expressions of uncertainty, refusal, or grounded speculation over confident fabrication.
ME	Memory-Aligned Coherence (MAC)	Rates the logical and topical coherence of the response with respect to the activated memory and context.
	Human-Like Enactment (HLE)	Rates the naturalness, tonal appropriateness, and conversational fluency of the response, ensuring it embodies a human-like utterance consistent with the persona.

Schema, MRPrompt introduces a Magic-If Protocol as an explicit LTM–STM interface prompt. It frames  $\mathcal{M}_L$  as the character’s internal memory store and the multi-turn dialogue  $\mathcal{M}_S$  as the *given circumstances* (STM cues), and specifies a minimal inference-time policy: (i) establish a stable identity by grounding in core traits (MA); (ii) interpret STM cues to select and activate the most relevant scene facet(s) (MS); (iii) apply boundary anchors to remain within the character’s temporal/epistemic knowledge and avoid out-of-scope leakage (e.g., spoilers or inaccessible claims) (MB); (iv) enact the selected and bounded memory into a coherent, human-like in-character reply (ME). Crucially, this protocol turns the LTM–STM interaction into a *controllable mechanism* rather than an implicit behavior emergent from prompting, enabling stage-wise attribution and systematic ablations.

MRPrompt contributes a theory-grounded separation of *representation* (Narrative Schema) and *control* (Magic-If Protocol): the former makes persona memory queryable and cue-addressable, while the latter makes retrieval and boundary enforcement explicit and auditable. Together, they support selective, bounded, and diagnosable persona utilization aligned with MA/MS/MB/ME, which is precisely what MDRP and MREval are designed to probe. Complete prompt templates are provided in Appendix Q.3.

### 3.4 MRBench: Benchmark Construction

To evaluate MDRP in a stage-diagnostic manner, we need benchmark splits for MA/MS/MB/ME where performance differences are attributable to the *targeted memory stage* rather than uncontrolled scene variation.

MRBench is constructed based on the principle of context reuse and minimal perturbation: we maximize the reuse of a shared short-term memory (STM) context pool  $\mathcal{M}_S$ , and create paired test conditions by applying minimal, targeted edits whenever possible to either the long-term memory (LTM) context  $\mathcal{M}_L$  or the final-turn query  $c_K$ . This design mitigates scene-induced confounds and substantially reduces the annotation overhead compared to building separate, scene-diverse datasets for each ability. As a result, we obtain controlled ( $\mathcal{M}_L, \mathcal{M}_S$ ) pairs that cleanly isolate the effects of MA, MS, and MB, while ME is evaluated on the same model outputs to avoid confounding factors. The full construction procedure is provided in Appendix E.

## 4 Experiments

We empirically study Memory-Driven Role-Playing (MDRP) under MREval, focusing on nine research questions: **RQ1:** How do small LLM backbones perform on the MDRP task using MRBench? **RQ2:** Does MRPrompt outperform baselines across diverse small LLMs? **RQ3:** What is the contribution of each component in MRPrompt? **RQ4:** How does MRPrompt affect the gap between small-scale and closed-source LLMs on MDRP? **RQ5:** How reliable is the LLM-as-a-Judge for MREval? **RQ6:** How do upstream abilities (MA/MS/MB) relate to and predict the ME ability? **RQ7:** What is the token cost of the baseline prompts vs. MRPrompt? **RQ8:** What is the per-turn end-to-end latency of different backbones? **RQ9:** How do baselines and MRPrompt differ in qualitative MDRP case studies? Due to

Table 2: **Experimental results on MRBench.** Human-calibrated GPT-4.1-mini scores (higher is better) for the eight MREval metrics in English/Chinese (en/zh). Avg. Score is the mean over all 8 metrics  $\times$  2 languages.

Model	Persona	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE		Avg. Score
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	
Qwen3-0.6B	Base	7.38	7.21	5.94	7.80	6.13	8.35	7.45	7.47	8.46	8.08	6.20	6.45	5.99	6.60	5.28	5.83	6.91
	Card	7.88	7.93	6.12	7.63	5.51	8.18	7.20	7.50	8.53	8.21	6.37	6.37	6.22	6.79	5.52	6.06	7.00
	MRPrompt	<b>8.02</b>	<b>8.06</b>	<b>6.67</b>	<b>7.81</b>	<b>6.46</b>	<b>8.69</b>	<b>7.63</b>	<b>7.77</b>	<b>8.65</b>	<b>8.23</b>	<b>6.54</b>	<b>6.56</b>	<b>6.33</b>	<b>6.97</b>	<b>5.86</b>	<b>6.35</b>	<b>7.29</b>
Qwen3-4B	Base	8.65	7.80	7.54	7.99	<b>7.97</b>	8.63	8.59	8.23	8.79	8.53	6.55	<b>7.04</b>	7.55	7.43	7.20	7.17	7.85
	Card	8.72	8.53	7.57	7.80	6.72	8.40	8.33	8.16	8.78	8.48	6.70	6.95	7.73	7.41	7.39	7.21	7.81
	MRPrompt	<b>8.88</b>	<b>8.61</b>	<b>7.69</b>	<b>8.23</b>	7.61	<b>8.81</b>	<b>8.69</b>	<b>8.57</b>	<b>8.85</b>	<b>8.56</b>	<b>6.84</b>	6.99	<b>8.07</b>	<b>7.63</b>	<b>7.73</b>	<b>7.42</b>	<b>8.07</b>
Qwen3-8B	Base	8.67	8.08	7.52	8.02	<b>8.03</b>	8.60	<b>8.83</b>	8.23	8.85	8.51	6.63	<b>7.36</b>	7.83	7.34	7.50	7.12	7.95
	Card	8.88	8.49	7.66	8.05	7.09	8.47	8.54	8.28	8.81	<b>8.55</b>	6.64	6.98	7.91	7.52	<b>7.59</b>	7.16	7.91
	MRPrompt	<b>8.97</b>	<b>8.73</b>	<b>7.76</b>	<b>8.25</b>	7.56	<b>8.99</b>	8.64	<b>8.56</b>	<b>8.88</b>	8.52	<b>6.92</b>	<b>7.07</b>	<b>8.13</b>	<b>7.64</b>	<b>7.73</b>	<b>7.42</b>	<b>8.12</b>
GLM-4-9B-Chat	Base	8.59	8.13	7.41	8.10	7.73	8.67	<b>8.65</b>	8.34	8.78	8.36	6.49	7.03	7.66	7.40	7.19	7.18	7.86
	Card	<b>8.87</b>	8.61	7.48	7.94	6.88	8.34	8.42	8.35	8.82	8.42	6.46	6.93	7.85	7.42	7.41	7.17	7.84
	MRPrompt	8.77	<b>8.70</b>	<b>7.59</b>	<b>8.23</b>	<b>7.93</b>	<b>8.88</b>	8.53	<b>8.52</b>	<b>8.83</b>	<b>8.54</b>	<b>6.76</b>	<b>7.07</b>	<b>8.02</b>	<b>7.63</b>	<b>7.53</b>	<b>7.41</b>	<b>8.06</b>
Llama-3-8B-Instruct	Base	8.23	7.40	7.13	<b>7.80</b>	<b>7.02</b>	8.62	<b>8.37</b>	7.52	8.78	8.41	6.50	6.95	7.50	7.00	6.74	6.57	7.53
	Card	<b>8.78</b>	8.09	7.33	7.32	6.29	8.55	8.13	7.73	8.81	8.28	6.69	6.84	7.75	6.84	7.05	6.37	7.55
	MRPrompt	8.69	<b>8.20</b>	<b>7.51</b>	7.68	6.83	<b>8.63</b>	8.32	<b>8.03</b>	<b>8.87</b>	<b>8.47</b>	<b>6.93</b>	<b>7.13</b>	<b>8.05</b>	<b>7.20</b>	<b>7.37</b>	<b>6.69</b>	<b>7.79</b>
Llama-3.2-3B-Instruct	Base	8.43	5.56	7.12	7.00	7.63	7.93	<b>8.45</b>	6.51	8.77	7.83	6.74	6.35	7.39	6.18	6.79	5.15	7.11
	Card	8.55	6.43	7.21	6.77	6.53	8.14	8.36	7.06	8.85	7.75	6.73	6.47	7.63	6.24	6.99	5.17	7.18
	MRPrompt	<b>8.73</b>	<b>7.35</b>	<b>7.50</b>	<b>7.19</b>	<b>7.69</b>	<b>8.63</b>	<b>8.45</b>	<b>7.36</b>	<b>8.87</b>	<b>8.19</b>	<b>7.10</b>	<b>6.65</b>	<b>8.04</b>	<b>6.60</b>	<b>7.55</b>	<b>5.88</b>	<b>7.61</b>
InternLM2.5-7B-Chat	Base	7.97	6.97	7.11	7.96	7.49	8.80	8.39	8.09	8.64	8.36	6.24	6.94	6.96	7.31	6.74	6.88	7.55
	Card	<b>8.26</b>	7.76	7.17	7.85	7.12	8.70	8.17	7.94	8.71	8.40	6.55	<b>7.07</b>	7.39	7.26	6.81	6.87	7.63
	MRPrompt	8.08	<b>7.99</b>	<b>7.31</b>	<b>7.98</b>	<b>7.89</b>	<b>9.10</b>	<b>8.40</b>	<b>8.33</b>	<b>8.75</b>	<b>8.46</b>	<b>6.71</b>	6.92	<b>7.71</b>	<b>7.36</b>	<b>7.13</b>	<b>6.95</b>	<b>7.82</b>

space limitations, discussions of **RQ7–RQ9** are provided in the Appendix.

#### 4.1 Experimental Setup

**Benchmark.** We construct MRBench, a bilingual benchmark instantiated under MDRP via controlled variants of the inputs ( $\mathcal{M}_L, \mathcal{M}_S$ ) to target MA/MS/MB/ME. Sourced from a collection of ten English and six Chinese novels, MRBench contains 200 English and 200 Chinese instances per ability family; all instances are single-turn. We emphasize that MRBench is designed as a controlled next-turn diagnostic benchmark conditioned on multi-turn history, rather than a full long-horizon simulation. The detailed data source information is provided in Appendix E.

**Models.** We evaluate **twelve** instruction-tuned LLMs that support Chinese–English role-play, spanning open-source and API-based families. Our **seven open-source** backbones include Llama-3-8B-Instruct (AI@Meta, 2024a), Llama-3.2-3B-Instruct (AI@Meta, 2024b), Qwen3-{0.6B, 4B, 8B} (Qwen Team, 2025b), GLM-4-9B-Chat (Team GLM, 2024), and InternLM2.5-7B-Chat (Cai et al., 2024). For **closed-source** comparison, we additionally include five API-based LLMs: GPT-5.2, GLM-4.7, DeepSeek-Chat, Qwen3-Max, and Doubao-Seed-1.6-250615.<sup>1</sup>

<sup>1</sup>Official model documentation pages: (OpenAI, 2025; Team GLM, 2025; DeepSeek, 2025; Qwen Team, 2025a; Vol-

**Hyperparameters.** All models are used *as is* (no fine-tuning), with prompt-only interventions. We perform no hyperparameter search. In generation, we set temperature to  $T=0.7$ ; other decoding parameters follow provider defaults when not explicitly configurable.

**Compared prompting conditions.** We compare two baseline persona prompting formats with MR-Prompt. For each character, we derive persona content from the same source materials and keep the same STM input  $\mathcal{M}_S$  fixed; conditions differ only in persona representation and usage guidance. Specifically, we use (i) **Base**, a narrative persona summary; (ii) **Card**, a lightweight profile-card baseline following the CharacterEval persona format (Tu et al., 2024); and (iii) **MRPrompt**, our facet-structured LTM with an explicit LTM–STM interface. Construction prompts are provided in Appendix Q.1, and persona specifications with examples are given in Appendix C.

**Reporting.** Unless otherwise noted, we report *mean* scores over instances for each split and language. Due to evaluation cost, results are obtained from a single run per model and prompting condition under fixed decoding settings.

#### 4.2 Main Experiments (RQ1 & RQ2)

We evaluate seven off-the-shelf instruction-tuned LLMs on MRBench under MREval, comparing (engine, 2025).

Table 3: **Component ablation.** Each ability score (MA/MS/MB/ME) is averaged over its own 2 metrics  $\times$  2 languages (en/zh).

Model	Condition	MA	MS	MB	ME	Avg. Score
Qwen3-4B	Base	8.00	8.36	7.73	7.34	7.85
	+Protocol	8.17	<b>8.42</b>	7.71	7.41	7.93
	+Schema	8.30	8.24	7.77	7.58	7.97
	MRPrompt	<b>8.35</b>	<b>8.42</b>	<b>7.81</b>	<b>7.71</b>	<b>8.07</b>
GLM-4-9B-Chat	Base	8.06	8.35	7.67	7.36	7.86
	+Protocol	8.11	8.34	7.68	7.36	7.87
	+Schema	<b>8.32</b>	8.44	7.72	<b>7.67</b>	8.04
	MRPrompt	<b>8.32</b>	<b>8.47</b>	<b>7.80</b>	7.65	<b>8.06</b>

two baselines (Base, Card) with MRPrompt. Unless otherwise stated, all results reported in the main paper are based on human-calibrated judge scores. Raw judge scores are provided in Appendix H, and calibration details are described in §4.5. From the experimental results presented in Table 2, three key observations can be drawn.

**(1) Scaling helps, but not uniformly across stages.** Overall performance tends to improve with model capacity across MA/MS/MB/ME and both languages, suggesting stable capability differences under MDRP. However, the eight-metric breakdown is non-uniform: gains are often stronger in MA/ME than in MS/MB, indicating that *anchoring* and *surface realization* scale more readily than *cue-driven facet recall* and *boundary control*.

**(2) Baseline structuring (Card) is not enough.** Comparing Base and Card, Avg. Score changes are modest and mixed across backbones, and the gains are uneven across abilities. Card-style formatting can help MA/ME (persona uptake and response organization), yet it does not reliably improve MS or MB, leaving the core MDRP failure modes—*facet mis-recall* and *boundary violations*—largely unresolved.

**(3) MRPrompt yields diagnostic, stage-specific gains.** MRPrompt achieves the highest Avg. Score for every backbone relative to Base/Card. Crucially, improvements are diagnostic rather than cosmetic: gains frequently concentrate in MS and MB, while MA also improves (especially for smaller models), and ME increases more modestly but consistently.

Overall, the MRPrompt design—a retrieval-oriented structured LTM plus an explicit LTM–STM control protocol—primarily strengthens MA/MS/MB and is consistently accompanied by improved enactment (ME).

Table 4: **Backbone comparison (Large vs. Small).** The best and second-best results per column are bold and underlined.

Model	Condition	MA	MS	MB	ME	Avg. Score
Qwen3-8B	Base	8.07	8.42	7.84	7.45	7.95
GLM-4-9B-Chat	Base	8.06	8.35	7.67	7.36	7.86
Qwen3-8B	MRPrompt	<b>8.43</b>	8.44	7.85	<b>7.76</b>	<u>8.12</u>
GLM-4-9B-Chat	MRPrompt	8.32	8.47	7.80	7.65	8.06
Qwen3-Max	Base	8.31	8.41	7.89	7.73	8.08
GLM-4.7	Base	8.31	<b>8.60</b>	7.87	7.65	8.11
Qwen3-Max	MRPrompt	8.33	8.22	<b>8.02</b>	7.73	8.07
GLM-4.7	MRPrompt	<u>8.34</u>	<u>8.54</u>	<u>7.91</u>	<u>7.75</u>	<b>8.13</b>

### 4.3 Component Ablations (RQ3)

We ablate MRPrompt on two backbones (Qwen3-4B and GLM-4-9B-Chat) by isolating its two components: facet-structured LTM (Schema) and the LTM–STM interface (Protocol). We compare four conditions—Base, +Protocol, +Schema, and MRPrompt—and report ability-level averages in Table 3. Overall, Schema contributes the larger *overall* gain: replacing a narrative persona with facet-structured LTM yields a stronger improvement in Avg. Score on both backbones, reflecting more reliable persona anchoring and downstream enactment. Meanwhile, Protocol is complementary and more model-dependent: it brings limited change on top of Base (especially on GLM-4-9B-Chat), but provides further gains once structured LTM is present, making MRPrompt best overall. This pattern matches MDRP’s division of labor: Schema builds a cue-addressable memory space, and Protocol more reliably elicits goal-directed retrieval and boundary-aware responses.

### 4.4 Large vs. Small LLMs (RQ4)

Under the same MRPrompt setup, Figure 4 shows a clear capacity-shaped profile across both English and Chinese: closed-source SOTA models exhibit consistently strong, low-variance performance across upstream memory use and downstream enactment. For instance, Doubao-Seed-1.6-250615 leads MA/MS while remaining top-tier on ME; GPT-5.2 is best on CR in English and ties the top tier on AL. In contrast, small models are more uneven: several mid-sized LLMs (e.g., Qwen3-8B/4B, GLM-4-9B-Chat) can approach SOTA on upstream anchoring and facet selection, yet lag more on ME, suggesting that structured prompting stabilizes *memory use* but cannot fully compensate for limited *surface enactment* capacity; this bottleneck is most evident for the smallest model (Qwen3-0.6B). Constraint robustness is consistently harder than other upstream dimensions,

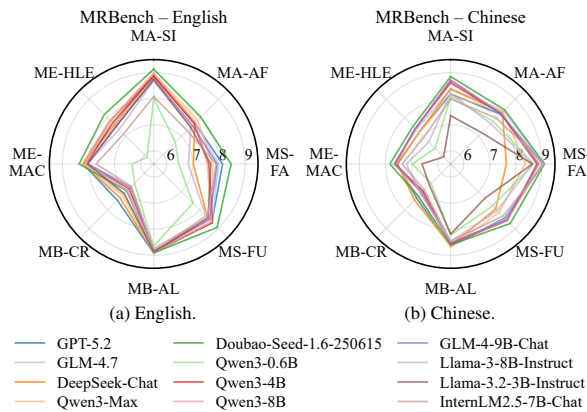


Figure 4: **All-model radar profiles on MRBench (MR-Prompt)**. Eight-axis MREval metric profiles for English and Chinese with a shared legend.

with CR typically trailing AL across models; while overall trends align between English and Chinese, several metrics still show language-specific shifts (full details in Appendix J).

More importantly, **MRPrompt enables smaller LLMs to rival the performance of much larger counterparts**. As shown in Table 4, the Qwen3-8B equipped with MRPrompt achieves an average score of 8.12, surpassing the standard versions of the much larger GLM-4.7 (8.11) and Qwen3-Max (8.08) models, and nearly matching the SOTA performance of GLM-4.7 with MRPrompt (8.13). This result underscores the unique value of MRPrompt: effectively enhancing smaller LLMs to achieve faithful, high-quality role-playing that is competitive with cutting-edge, scaled-up models.

#### 4.5 Judge Validation (RQ5)

We use GPT-4.1-mini as the LLM judge for all eight MREval metrics, using a 1–10 rating scale. To assess reliability, we construct a bilingual validation set with 100 instances per metric (50 en / 50 zh; 800 total). Across all metric–language pairs, judge–human agreement is moderate-to-strong and statistically significant.

Mean-score differences suggest mild systematic bias, motivating per-metric calibration on the validation set. Appendix K reports the full judge-validation details, including the correlation statistics in Tables 12 and 13 (Pearson/Spearman/Kendall and  $p$ -values), the mean-score comparison in Table 14, the calibration parameters, additional visual checks, and a supplemental post-hoc robustness analysis with four independent human raters and an alternative judge model (Qwen3-32B).

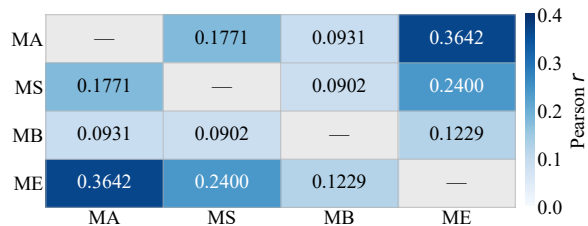


Figure 5: **Ability correlations**. Heatmap of Pearson correlations ( $r$ ) between ability-level scores over instances.

#### 4.6 Ability Interactions (RQ6)

MREval decomposes MDRP into four stages, but effective role-playing should behave as a coupled pipeline. We therefore test (i) whether scores of different abilities are correlated across evaluated instances, and (ii) whether stronger upstream memory abilities (MA/MS/MB) are associated with better downstream enactment (ME).

**Ability correlations.** We aggregate the two metrics within each ability to obtain ability-level scores, and compute Pearson correlations over instances where all four abilities are available. Figure 5 reports correlations on the pooled EN+ZH set: correlations are uniformly positive but weak-to-moderate, indicating that abilities are *coupled yet non-redundant*. Notably, MA exhibits the strongest association with ME, while MS and MB show weaker correlations with other abilities, suggesting that (a) anchoring in persona memory often translates to better surface enactment, but (b) *selection* and *boundary control* remain comparatively independent bottlenecks rather than automatically improving with general response quality. Language-split analyses and additional statistics are provided in Appendix L.

**Upstream association.** Aggregating MA/MS/MB into a single upstream score, we find a moderate positive association with ME on the pooled set ( $r=0.35$ ), implying that better anchoring/selecting/bounding behavior tends to accompany improved enactment, but does not fully determine it. Appendix L reports language-split correlations and further details.

### 5 Conclusion

We study Memory-Driven Role-Playing (MDRP) from a memory perspective, where persona knowledge serves as LTM, and dialogue context serves as STM. We introduce MREval, a diagnostic evaluation that decomposes MDRP into four abilities with

eight Likert-style metrics, and construct a bilingual benchmark MRBench. Building on this formulation, we propose MRPrompt to enable cue-driven retrieval and boundary-aware generation. Experiments show consistent improvements over narrative and profile-card baselines, with the largest gains in selection and boundary-related diagnostics. Overall, our results suggest that making memory representation and use explicit is an effective route to more faithful role-playing.

## Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant No. 62506100). Dr. Yang Zhang of Macquarie University received no financial support for this work from the above grants or any other external projects. His contribution was made independently as part of his academic research. The authors also sincerely acknowledge the valuable collaboration and insightful discussions contributed by colleagues from the participating universities.

## Limitations

MDRP is evaluated in a controlled *next-turn* setting, which supports clear stage-wise diagnosis but does not cover interactive phenomena such as memory carryover, revision, or long-range character drift. Meanwhile, MRBench is built from literary characters to enable controlled diagnosis with verifiable persona evidence, but this setting does not fully cover original or non-literary personas that evolve through ongoing interaction. Finally, while GPT-4.1-mini is calibrated for MREval and supplemented with a small post-hoc robustness check, the main results still rely on one automated judge; broader multi-judge and multi-annotator validation is left for future work.

**Ethical considerations.** Our work involves low-risk human annotation but no user studies and no collection of sensitive personal data; all data used are publicly available and commonly used in prior work (Appendix E). We do not introduce additional categories of safety, privacy, or fairness risks beyond those typical for benchmarking role-conditioned text generation.

**GenAI usage disclosure.** Generative AI tools were used for text polishing and code debugging, and were not used for method design or experimental analysis.

## References

- AI@Meta. 2024a. [Llama 3 model card](#). Accessed: 2026-01-05.
- AI@Meta. 2024b. [Llama 3.2: Model cards and prompt formats](#). Accessed: 2026-01-05.
- Alan Baddeley. 1992. [Working memory](#). *Science*, 255(5044):556–559.
- Wenrui Bao, Kai Wang, Siqiang Luo, and Xiang Li. 2025. [Permitted knowledge boundary: Evaluating the knowledge-constrained responsiveness of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13390–13405, Suzhou, China. Association for Computational Linguistics.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. [InternLM2 Technical Report](#). *Preprint*, arXiv:2403.17297.
- Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. 2025a. [CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10254–10286, Suzhou, China. Association for Computational Linguistics.
- Weishu Chen, Jinyi Tang, Zhouhui Hou, Shihao Han, Mingjie Zhan, Zhiyuan Huang, Delong Liu, Jiawei Guo, Zhicheng Zhao, and Fei Su. 2025b. [MOOM: Maintenance, organization and optimization of memory in ultra-long role-playing dialogues](#). *Preprint*, arXiv:2509.11860.
- DeepSeek. 2025. [Models & pricing \(deepseek api docs\)](#). Accessed: 2026-01-05.
- Yifan Duan, Yihong Tang, Kehai Chen, Liqiang Nie, and Min Zhang. 2025. [ORPP: Self-optimizing role-playing prompts to enhance language model capabilities](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28573–28588, Suzhou, China. Association for Computational Linguistics.
- Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. [Role-playing evaluation for Large Language Models](#). In *Methodologies and Intelligent Systems for Technology Enhanced Learning, 15th International Conference (MIS4TEL 2025)*, volume 1619 of *Lecture Notes in Networks and Systems*, pages 118–127. Springer, Cham.
- Feiteng Fang, Ting-En Lin, Yuchuan Wu, Xiong Liu, Xiang Huang, Dingwei Chen, Jing Ye, Haonan Zhang, Liang Zhu, Hamid Alinejad-Rokny, Min Yang, Fei Huang, and Yongbin Li. 2025. [ChARM: Character-based act-adaptive reward modeling for](#)

- advanced role-playing language agents. *Preprint*, arXiv:2505.23923.
- Xiaoxu Han, Wanqing Zhao, Ziyu Guan, and Jinye Peng. 2026. **Act-LLM: A whole-process chain for character-centric role-playing with large language models.** *Expert Systems with Applications*, 296:129024.
- Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaying Xu, Wenqiang Liu, and Mengling Feng. 2025. **Crab: A novel configurable role-playing LLM with assessing benchmark.** In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052, Vienna, Austria. Association for Computational Linguistics.
- Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. **Emotional RAG: Enhancing role-playing agents through emotional retrieval.** In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 120–127. IEEE.
- Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suh-yune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. **Call for customized conversation: Customized conversation grounding persona and knowledge.** *Preprint*, arXiv:2112.08619.
- Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. **Enhancing persona consistency for LLMs’ role-playing using persona-aware contrastive learning.** In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238, Vienna, Austria. Association for Computational Linguistics.
- 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.** *Preprint*, arXiv:2308.09597.
- Rensis Likert. 1932. **A technique for the measurement of attitudes.** *Archives of Psychology*, 22(140):1–55.
- Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, and Zhaoxiang Zhang. 2024. **RoleAgent: Building, interacting, and benchmarking high-quality role-playing agents from scripts.** In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, pages 49403–49428.
- 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.
- Minsik Oh, Joosung Lee, Jiwei Li, and Guoyin Wang. 2024. **PK-ICR: Persona-knowledge interactive context retrieval for grounded dialogue.** *Preprint*, arXiv:2302.06674.
- OpenAI. 2025. **GPT-5.2 Model (OpenAI API Documentation).** Accessed: 2026-01-05.
- Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. **KokoroChat: A Japanese Psychological Counseling Dialogue Dataset Collected via Role-Playing by Trained Counselors.** In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12424–12443, Vienna, Austria. Association for Computational Linguistics.
- Qwen Team. 2025a. **Qwen3-Max: Just Scale it.** Accessed: 2026-01-05.
- Qwen Team. 2025b. **Qwen3 technical report.** *Preprint*, arXiv:2505.09388.
- Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. **BOOK-WORLD: From novels to interactive agent societies for story creation.** In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912, Vienna, Austria. Association for Computational Linguistics.
- Saksorn Ruangtanusak, Pittawat Taveekitworachai, and Kunat Pipatanakul. 2025. **Talk less, call right: Enhancing role-play LLM agents with automatic prompt optimization and role prompting.** *Preprint*, arXiv:2509.00482.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. **Character-LLM: A trainable agent for role-playing.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.
- Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. 2025. **Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation.** In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26312–26332, Vienna, Austria. Association for Computational Linguistics.
- Constantin Stanislavski. 1989. *An actor prepares.* Routledge.
- Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025a. **Thinking in character: Advancing role-playing agents with role-aware reasoning.** *Preprint*, arXiv:2506.01748.

- Yihong Tang, Bo Wang, Xu Wang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025b. **Role-Break: Character hallucination as a jailbreak attack in role-playing systems**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7386–7402, Abu Dhabi, UAE. Association for Computational Linguistics.
- Team GLM. 2024. **ChatGLM: A family of large language models from glm-130b to glm-4 all tools**. *Preprint*, arXiv:2406.12793.
- Team GLM. 2025. **GLM-4.7 Overview (Z.AI Developer Docs)**. Accessed: 2026-01-05.
- Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. **CharacterChat: Learning towards conversational AI with personalized social support**. *Preprint*, arXiv:2308.10278.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. **CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- Endel Tulving and Donald M. Thomson. 1973. **Encoding specificity and retrieval processes in episodic memory**. *Psychological Review*, 80(5):352–373.
- Volcengine. 2025. **doubao-seed-1.6 (volcengine ark model documentation)**. Accessed: 2026-01-05.
- Kai Wang, Yuwei Xu, Zhiyong Wu, and Siqiang Luo. 2024a. **LLM as Prompter: Low-resource inductive reasoning on arbitrary knowledge graphs**. *Preprint*, arXiv:2402.11804.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. **CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.
- Noah Wang, Z.y. 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, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. **RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025b. **CoSER: Coordinating LLM-based persona simulation of established roles**. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64822–64858. PMLR.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. **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, Bangkok, Thailand. Association for Computational Linguistics.
- Yongjie Wang, Jonathan Leung, and Zhiqi Shen. 2025c. **RoleRAG: Enhancing LLM role-playing via graph guided retrieval**. *Preprint*, arXiv:2505.18541.
- Shihao Yang, Zhicong Lu, Yong Yang, Bo Lv, Yang Shen, and Nayu Liu. 2025. **HyCoRA: Hyper-contrastive role-adaptive learning for role-playing**. *Preprint*, arXiv:2511.08017.
- Bingsheng Yao, Bo Sun, Yuanzhe Dong, Yuxuan Lu, and Dakuo Wang. 2025. **DPRF: A generalizable dynamic persona refinement framework for optimizing behavior alignment between personalized LLM role-playing agents and humans**. *Preprint*, arXiv:2510.14205.
- Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and Alex Smola. 2025a. **RPG-Bench: Evaluating large language models as role-playing game engines**. *Preprint*, arXiv:2502.00595. Also presented at the NeurIPS 2025 Workshop on Scaling Environments for Agents (SEA).
- Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. **Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557, Miami, Florida, USA. Association for Computational Linguistics.
- Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025b. **Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11992–12022, Vienna, Austria. Association for Computational Linguistics.
- Chao Zhang, Jianwen Sun, Jie Ma, Yi Yang, and Yawei Luo. 2025a. **TeenEmpath: Towards adolescent psychological counseling with multiple personas and strategies**. *IEEE Transactions on Affective Computing*, pages 1–13.

- Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, Di Yin, Xing Sun, and Kai Zhang. 2025b. [RolePlot: A systematic framework for evaluating and enhancing the plot-progression capabilities of role-playing agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12337–12354, Vienna, Austria. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Weixiang Zhao, Yulin Hu, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. [Beware of your po! measuring and mitigating AI safety risks in role-play fine-tuning of LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11112–11137, Vienna, Austria. Association for Computational Linguistics.
- Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025a. [CharacterBench: Benchmarking character customization of large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110. AAAI Press.
- Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. 2025b. [PersonaEval: Are LLM evaluators human enough to judge role-play?](#) *Preprint*, arXiv:2508.10014. Also appears in COLM 2025.

## A Datasets

Tables 5 and 6 summarize the source novels and dialogue counts used to construct the base STM pool (Appendix E). Each sample is an ABA-style multi-turn snippet centered on a target character (A), where B denotes the interlocutor.

**Chinese subset.** The Chinese subset covers 6 novels, 30 named characters, and 320 dialogue snippets, with an average dialogue length of 8.87 turns per sample.

**English subset.** The English subset covers 10 novels, over 40 named characters, and 252 dialogue snippets, with an average dialogue length of 11.19 turns per sample.

## B Stanislavski-Inspired Foundation of MRPrompt

This appendix clarifies how Stanislavski’s acting theory motivates the design of MRPrompt.

**Why Stanislavski?** A central idea in Stanislavski’s system is that convincing performance does not arise from mechanically imitating surface style, but from grounding behavior in the character’s internalized experience under the current *given circumstances*. In particular, concepts such as *emotional memory*, *given circumstances*, and *magic if* emphasize that an actor should (i) recall relevant inner material, (ii) interpret the current situation, and (iii) act within the character’s knowledge and situation rather than improvise without constraint. This intuition is closely aligned with our MDRP formulation, where the model must use dialogue context to retrieve and apply persona knowledge, instead of merely producing stylistically plausible text.

**Theory-to-method mapping.** MRPrompt operationalizes this inspiration in two components.

**Narrative Schema as structured character memory.** Stanislavski emphasizes that performance is anchored in a coherent inner life rather than a flat list of traits. Analogously, MRPrompt replaces flat persona descriptions with a structured Narrative Schema, which organizes persona knowledge into identity fields, global summary, core traits, and scene facets. This makes persona knowledge more queryable and supports memory-based grounding beyond superficial name priors.

**Magic-If Protocol as cue-guided enactment.** Stanislavski’s “magic if” asks the actor to reason

Table 5: Chinese novels and character statistics.

Novel (ZH)	Main characters (dialogue count)	Total
三体·黑暗森林	罗辑 26, 庄颜 12, 史强 11, 林格 10, 萨伊 8	67
水浒传	卢俊义 5, 宋江 5, 施恩 5, 戴宗 3, 李逵 2	20
神雕侠侣	杨过 35, 金轮法王 21, 小龙女 8, 郭芙 5, 黄蓉 5	74
红楼梦	林黛玉 14, 袭人 9, 探春 4, 紫鹃 4, 晴雯 2	33
花千骨	花千骨 24, 白子画 13, 东方朔 8, 摩严 6, 杀阡陌 6	57
西游记	孙悟空 39, 猪八戒 22, 唐三藏 3, 红孩儿 3, 牛魔王 2	69
<b>Total</b>	–	<b>320</b>

Table 6: English novels and character statistics.

Novel (EN)	Main characters (dialogue count)	Total
A Tale of Two Cities	Jarvis Lorry 5, Charles Darnay 4, Sydney Carton 3, Mr. Stryver 2, Doctor Manette 2	16
Catch-22	Yossarian 8, Doc Daneeka 2, Milo Minderbinder 2, Ex-P.F.C. Wintergreen 2, Colonel Cathcart 2	16
Crime and Punishment	Rodion Raskolnikov 7, Arkady Svidrigaylov 6, Sofya Marmeladov 5, Razumikhin 2, Porfiry Petrovich 2	22
Harry Potter	Harry Potter 27, Albus Dumbledore 8, Hermione Granger 2, Ron Weasley 3, Sirius Black 3, Dudley Dursley 1, Dobby 1, Tom Riddle 1, Others 16	60
Little Women	Laurie Laurence 12, Jo March 6, Mr. Laurence 2, Meg March 2, John Brooke 1	23
Pride and Prejudice	Elizabeth Bennet 22, Fitzwilliam Darcy 4, George Wickham 3, Mrs Bennet 2, Jane Bennet 2	33
The Adventures of Tom Sawyer	Tom Sawyer 9, Huckleberry Finn 7, Aunt Polly 5, Ben Rogers 1, Mr. Dobbins 1	23
The Little Prince	The Little Prince 9, The Narrator 1, The Businessman 1, The Geographer 1, The Snake 1	13
The Red and the Black	Julien Sorel 22, Mathilde de La Mole 5, Madame de Rênal 4, Monsieur de Rênal 2, Fouqué 2	35
Treasure Island	Jim Hawkins 5, Doctor Livesey 2, The Captain 2, Jim’s Mother 1, Squire Trelawney 1	11
<b>Total</b>	–	<b>252</b>

from the character’s perspective under the current circumstances: *if I were this character in this situation, what would I do?* Our Magic-If Protocol plays an analogous role at inference time. It instructs the model to: (i) stabilize identity from core traits; (ii) extract cues from the dialogue context; (iii) activate the most relevant scene facet; (iv) respect temporal and epistemic boundaries; and (v) produce a natural in-character reply. Thus, the protocol turns persona use into an explicit LTM–STM control process rather than leaving it as unconstrained prompting.

## C Persona Construction and Annotation

### C.1 Compared Prompting Conditions

For each character, we instantiate three persona prompting conditions from the same underlying source materials (Appendix Q.1). Across conditions, we keep the dialogue history  $\mathcal{M}_S$  and the core role-play constraints identical, and only vary (i) the persona long-term memory  $\mathcal{M}_L$  representation and (ii) the provided usage guidance.

- **Base (Narrative persona).** A single narrative

summary that interleaves global traits and representative episodes, paired with a standard role-playing instruction.

- **Card (Profile card baseline).** A semi-structured persona card following the CharacterEval persona format (Tu et al., 2024), which organizes persona information into lightweight fields (e.g., name, global summary, personality, relations), paired with the same standard role-playing instruction. Compared to Base, Card adds structure but does not impose an explicit retrieval procedure.
- **MRPrompt (ours).** A facet-structured LTM plus an explicit LTM–STM Protocol that instructs cue-driven facet activation and boundary-aware response generation, aligning persona use with the MA/MS/MB/ME stages evaluated by MREval.

## C.2 English Persona Examples

In this section, we provide simple example descriptions of the two characters (Tom Sawyer and Charles Darnay) involved in the case study. For brevity, we only show the LTM content in the figures; the shared standard role-playing instruction and the Magic-If Protocol are provided in Appendix Q.1.

### C.2.1 Tom Sawyer

We briefly illustrate the three persona formats used for Tom Sawyer: *Base–LTM* (Figure 6), *Card–LTM* (Figure 7), and *MRPrompt–LTM* (Figure 8).

### C.2.2 Charles Darnay

We provide brief examples of the three persona formats for Charles Darnay: *Base–LTM* (Figure 9), *Card–LTM* (Figure 10), and *MRPrompt–LTM* (Figure 11).

## C.3 LTM Construction

We construct persona LTMs with an LLM-assisted, human-in-the-loop pipeline. GPT-4.1 (API) is first prompted (Appendix Q.1) to draft a baseline narrative profile and its semi-structured and facet-structured variants in the target language. For all human-involved tasks, participants are shown the same task prompts used for LLM prompting (Appendix Q), ensuring consistent instructions across human and model runs. Annotators then verify and edit the drafts against the original novels to remove hallucinations, correct plot details, and enforce consistency across the overview, traits, and facets; ex-

ternal reference materials may be consulted for fact-checking but are not copied verbatim. We also use GPT-4.1 (API) to assist controlled edits of evaluation instances during benchmark construction (Appendix E); all edits are manually checked for faithfulness and consistency before inclusion. Only human-verified personas and benchmark instances are used in all experiments.

## D Facet Schema

In MRPrompt, each character profile serves as explicit long-term memory (LTM;  $\mathcal{M}_L$ ) with two main layers: **core traits** (core personality traits with brief explanations) and **scene facets** (structured, multi-faceted manifestations under recurring interaction contexts). Each facet specifies *title*, *time\_scope*, and *situation*, the character’s *social\_role* and *emotional\_state*, typical *behavior\_pattern* and *thinking\_pattern*, potential *conflict\_with\_core*, *grounding\_source\_scenes*, and *cue\_phrases* that can be matched against STM cues. For completeness, our persona files also include auxiliary identity fields (e.g., name/relations) and a global summary for background context; Table 7 focuses on the diagnostic core-trait and facet fields that support MA/MS/MB/ME.

## E Benchmark Construction for MDRP (MRBench)

We instantiate MDRP as a bilingual benchmark, MRBench, by starting from a shared pool of short-term memories (STM) adapted from publicly released role-playing dialogue data in CharacterEval (Tu et al., 2024) and Crab (He et al., 2025), and pairing each scene with the target character’s persona as explicit long-term memory (LTM). Each instance follows our formulation ( $\mathcal{M}_L, \mathcal{M}_S$ ), where  $\mathcal{M}_S = [u_1, \dots, u_K]$  is the dialogue context and the last turn  $u_K$  is always the interlocutor’s final message (i.e., the model responds next). MRBench is built via minimal, controlled transformations of ( $\mathcal{M}_L, \mathcal{M}_S$ ) to isolate different stages of the memory pipeline. Our key design choice is that *STM is the anchor*: whenever possible, we reuse the same underlying scenes and vary only the minimal component ( $\mathcal{M}_L$  or  $c_K$ ) needed for diagnosis. Corpus statistics are provided in Appendix A; persona construction is described in Appendix C and Appendix Q.1.

Base Tom Sawyer
<p>[Overview] Tom Sawyer’s story unfolds in the small, bustling town of St. Petersburg, Missouri, where he lives under the care of his Aunt Polly after the death of his parents. Early on, Tom is marked by a mischievous spirit and a profound longing for adventure, often clashing with the expectations of his guardians and the routines of school and church...(abridged)</p> <p>[Scene 1: The Whitewashing of Aunt Polly’s Fence] On a bright Saturday morning in the heart of St. Petersburg, Tom stands begrudgingly before Aunt Polly’s fence, sentenced to the tedious task of whitewashing as punishment for his latest escapade...(abridged)</p> <p>[Scene 2: Nighttime Adventure in the Graveyard] Late at night, Tom and Huck sneak through the shadows to the town graveyard, driven by superstition and the thrill of the forbidden...(abridged)</p> <p>[Scenes 3–5 omitted for brevity]</p> <p>[Scene 6: Quiet Reflection After the Adventure] After the dust settles and the treasure is found, Tom retreats to the banks of the Mississippi, sitting quietly beneath a tree as dusk falls...(abridged)</p>

Figure 6: Narrative LTM excerpt for Tom Sawyer.

### E.1 Base STM Pool

We collect ABA-style multi-turn dialogue snippets from public Chinese and English literary/dialogue corpora. Each snippet is centred on one target character and ends at the interlocutor’s turn immediately before the target character’s next response, yielding a natural single-turn continuation target under a fixed conversational state. When available, we also record the book-grounded reference continuation  $\hat{y}^{gold}$  for the next target-character turn. After filtering for coherence, speaker attribution, and diagnosticity, we obtain a base STM pool of 320 Chinese and 252 English snippets (Appendix A).

### E.2 LTM Variants for Benchmarking

For each target character, we construct canonical persona memories as explicit LTM. Here we treat LTM as a black-box conditioning source and focus on the *controlled variants* used for benchmarking. From each canonical LTM, we derive lightweight task-specific variants (e.g., anonymized, facet-removed, facet-rewritten) used to construct the ability-focused splits described next.

**MA split (joint anonymization of LTM and STM).** To probe MA, we reduce shortcuts tied to surface identity by jointly anonymizing both memories: (i) replace character names and direct identifiers in  $\mathcal{M}_L/\mathcal{M}_S$  with synthetic aliases, and (ii) normalize or remove references that trivially reveal the original IP. We retain only scenes where in-character behavior remains clearly diagnostic after anonymization, yielding 200 Chinese and 200 English items. Each item provides two controlled conditions: *full* ( $\mathcal{M}_L^{\text{full}}, \mathcal{M}_S^{\text{orig}}$ ) and *anonymized*

( $\mathcal{M}_L^{\text{anon}}, \mathcal{M}_S^{\text{anon}}$ ), supporting the MA metrics.

**MS split (facet manipulation of LTM under fixed STM).** To probe MS, we keep  $\mathcal{M}_S^{\text{orig}}$  fixed and manipulate only facet-related content in  $\mathcal{M}_L$ , yielding three conditions: matching  $\mathcal{M}_L^{\text{full}}$ , facet-less  $\mathcal{M}_L^{\text{no-scene}}$ , and counter-facet  $\mathcal{M}_L^{\text{anti}}$ . For each persona format, we construct the corresponding variants and select 200 Chinese and 200 English STM snippets where facet differences are expected to induce measurably different continuations, supporting the MS metrics.

**MB split (out-of-scope perturbation of the final interlocutor turn in STM).** To probe MB, we keep  $\mathcal{M}_L^{\text{full}}$  unchanged and keep the dialogue prefix  $[u_1, \dots, u_{K-1}]$  fixed, while perturbing only the utterance content in the final turn from  $c_K^{\text{in}}$  (in-scope) to  $c_K^{\text{out}}$  (out-of-scope) to tempt boundary violations. We include two out-of-scope types: future-timeline queries (beyond the story time implied by STM) and out-of-domain queries (outside the character/world knowledge). After filtering for clear cutoffs and strong temptation cases, we obtain 200 Chinese and 200 English items, supporting the MB metrics.

**ME scoring set (re-scoring a balanced sample of generations).** ME introduces no new generation condition. Instead, we re-score outputs produced in the MA/MS/MB settings from an enactment-centric perspective. Concretely, we pool generations from anonymized MA items, the three MS facet conditions, and the two MB  $c_K^{\text{out}}$  types, then sample a balanced set of 200 Chinese and 200 English instances for ME scoring. This ensures ME evaluates surface enactment on the *same* underlying memory-use cases as the upstream stages.

```

Card Tom Sawyer
{
  "name": "Tom Sawyer",
  "Nickname": "Tom",
  "Relationships": [
    { "name": "Aunt Polly", "relationship": "guardian" },
    { "name": "Huckleberry Finn", "relationship": "friend" },
    { "name": "Becky Thatcher", "relationship": "love interest" },
    ...
  ],
  "global_summary": "Tom Sawyer grows up in the lively town of St. Petersburg, Missouri, under the watchful care of his Aunt Polly after losing his parents. From an early age, Tom is known for his mischievous nature and an insatiable appetite for adventure, often finding himself at odds with the expectations of adults and the routines of small-town life...",
  "Personality": {
    "core_traits": [
      { "trait": "mischievous" },
      { "trait": "clever" },
      { "trait": "adventurous" },
      { "trait": "resourceful" },
      ...
    ],
    "scene_facets": [
      {
        "title": "Whitewashing Aunt Polly's Fence",
        "situation": "On a sunny Saturday morning in St. Petersburg, Tom is assigned the tedious task of whitewashing his aunt's fence as punishment. Initially resentful, he quickly devises a plan to make the job seem desirable, boasting to passing boys about its exclusivity and trading chances to paint for small treasures. Through playful manipulation and strategic performance, Tom transforms the punishment into an opportunity for social admiration and personal gain, orchestrating the entire affair with energetic showmanship.",
        "emotional_state": "resentment, triumph, playful glee",
        "behavior_pattern": "Cleverly manipulates peers with enthusiastic boasting and teasing remarks."
      },
      ...
    ]
  }
}

```

Figure 7: Semi-structured LTM excerpt for Tom Sawyer.

### E.3 Summary

In sum, MRBench is built around a shared STM pool and minimal, stage-targeted transformations of  $(\mathcal{M}_L, \mathcal{M}_S)$ . MA varies identity cues via joint anonymization of  $\mathcal{M}_L/\mathcal{M}_S$ ; MS varies facet content in  $\mathcal{M}_L$  under fixed  $\mathcal{M}_S^{\text{orig}}$ ; MB perturbs only the final interlocutor turn ( $u_K$ ) to an out-of-scope query, keeping  $\mathcal{M}_L$  and  $[u_1, \dots, u_{K-1}]$  fixed; and ME re-scores a balanced subset of generations drawn from these settings. By reusing the same underlying STM scenes whenever possible, differences across stages are attributable to the targeted transformations rather than scene variation.

### E.4 Artifact provenance, licensing, and intended use.

**Upstream artifacts and licensing.** Our base STM pool is adapted from publicly released role-playing dialogue artifacts in CharacterEval (Tu et al., 2024) and Crab (He et al., 2025). CharacterEval is released under the MIT license. Crab provides code and data via the official repository linked in the paper; we follow the repository’s stated terms and intended research usage.

**Consistency with intended use.** We use these upstream artifacts strictly for non-commercial academic research—as dialogue contexts for benchmarking memory-driven role-playing—which is consistent with their role-playing research and evaluation purpose. We specify MRBench (and accompanying prompts) as *research-only evaluation* artifacts for MDRP. MRBench is not intended for deployment in user-facing products or for non-research uses. We release only the materials necessary to reproduce our experiments, consistent with upstream access conditions. Any downstream use must comply with the original licenses/terms of the upstream artifacts; we do not claim additional rights over upstream content beyond what is permitted by those terms.

### E.5 PII and offensive content screening.

**PII.** MRBench is derived from fictional literary sources and existing RP benchmarks; we do not collect user-generated personal data. As a precaution, we apply lightweight screening for obvious personally identifying patterns (e.g., emails, phone numbers, URLs, physical-address-like strings) and exclude any flagged cases. Note that anonymization in MA is designed for controlled evaluation

Table 7: **Facet schema for MRPrompt personas.** MRPrompt structures persona LTM  $\mathcal{M}_L$  into *core traits* and *scene facets*, with key fields supporting MA/MS/MB/ME.

Field	Description (what it encodes / why it exists)	Primarily Supports
<b>Layer 1: Core traits (global schema)</b>		
core_traits	A list of trait objects, each with a <i>trait</i> name and a short <i>desc</i> grounded in characteristic behaviors and tendencies. This layer serves as a compact global identity schema that the model can internalise as “who I am” before retrieving any situation-specific facet.	<b>MA</b>
desc	The textual explanation attached to each trait (stored as <i>desc</i> inside <i>core_traits</i> ). It provides behavioral semantics beyond adjective labels, improving in-context acquisition and reducing generic persona drift.	<b>MA</b>
<b>Layer 2: Scene facets (cue-addressable persona slices)</b>		
title	Concise label/index for the facet (human-readable handle).	<b>MS, ME</b>
time_scope	Story/life phase where this mode is typical; provides temporal anchors for present-time constraints.	<b>MB</b>
situation	Recurring interaction context that activates the facet; defines the retrieval target at the situation level.	<b>MS</b>
social_role	Typical social stance(s) (e.g., challenger/protector); conditions pragmatic style and power dynamics.	<b>MS, ME</b>
emotional_state	Characteristic emotions in this context; guides tone and emotional realism.	<b>ME</b>
behavior_pattern	Typical actions/strategies in dialogue; maps facet activation to concrete dialogue moves.	<b>ME</b>
thinking_pattern	Priorities/beliefs motivating behavior; stabilizes reasoning consistency beyond surface style.	<b>MS, ME</b>
conflict_with_core	How the facet extends/strains core traits; maintains coherent identity under multi-faceted expression.	<b>MA, MB</b>
source_scenes	Canonical evidence anchors for traceability and faithful facet construction.	<b>MA, MS</b>
cue_phrases	Lexical/conceptual triggers to map STM cues to facets; enables retrieval without explicit per-turn scenario text.	<b>MS</b>

(reducing identity shortcuts and IP leakage), not as a privacy mechanism.

**Offensive content.** Fictional sources may contain sensitive or offensive language. We perform basic keyword-based screening and manual spot checks during filtering, and exclude instances with overt slurs or explicit harassment content when encountered.

**Human annotation.** All annotations and human reference scoring were performed in-house by lab members on a voluntary basis with informed consent. We did not collect or store direct personal identifiers or sensitive demographic attributes; annotations are used solely for research.

## E.6 Artifact documentation and statistics.

**Documentation.** We document MRBench’s domain (literary role-play), languages (Chinese and English), instance format ( $\mathcal{M}_L, \mathcal{M}_S$ ), split definitions (MA/MS/MB/ME), and scoring rubrics (Appendix P). Persona construction and prompt

formats are described in Appendix C and Appendix Q.1.

**Statistics.** Corpus-level statistics for the base STM pool are provided above (Appendix A). MRBench is an evaluation benchmark (no train/dev/test splits): in the main experiments, each ability split (MA/MS/MB/ME) contains 200 Chinese and 200 English instances.

## F Implementation and Computational Details

**Model size.** For open-source backbones, we report parameter scales as indicated by their official releases/model identifiers (e.g., 0.6B/3B/4B/7B/8B/9B). For closed-source API models, parameter counts are not publicly disclosed by providers and are therefore unavailable.

**Compute budget and infrastructure.** All model inferences in our experiments are executed via the API, including open-source backbones served by the provider. Thus, the underlying computing in-

frastructure and GPU-hour budget are not observable from the client side. We report evaluation scale by the number of evaluated instances and model calls per condition (Sec. 4.1 and Appendix E). In addition, we quantify average token usage on the shared STM pool for two representative models in Appendix M. No training is performed; all experiments are inference-only.

**Software.** Experiments are implemented in Python. We use standard libraries for data processing and analysis (e.g., NumPy/Pandas/SciPy) and visualization (Matplotlib). We release an anonymized code repository with evaluation scripts and dependency versions pinned in `requirements.txt`.

## G Related Work

### G.1 Task Design for General Role-Playing

General role-playing typically instantiates an in-character agent by providing a role profile derived from canonical materials (e.g., scripts, novels, or dialogue excerpts) and conditioning an LLM to generate character-consistent responses. Early systems such as ChatHaruhi and CharacterLLM exemplify this paradigm by grounding role-play in extracted character dialogues or curated character descriptions, respectively (Li et al., 2023; Shao et al., 2023). Subsequent benchmarks scale role definitions and interaction settings to enable more systematic evaluation: RoleLLM and RoleAgent introduce broader role pools and controlled interaction protocols, while CharacterEval and CharacterBench emphasize text-only assessments of consistency under diverse character constraints (Wang et al., 2024b; Liu et al., 2024; Tu et al., 2024; Zhou et al., 2025a). A related line of research explores persona-grounded dialogue settings in which persona information is combined with external knowledge or contextual evidence, rather than treated as a standalone role profile. For example, FoCus (Jang et al., 2022) grounds customized conversation in both persona and knowledge, while PK-ICR (Oh et al., 2024) studies interactive retrieval over persona- and knowledge-related contexts. Despite strong progress in task coverage and benchmarking practice, most task designs still treat role-play quality as an aggregate construct (e.g., overall fidelity, coherence, naturalness), rather than explicitly isolating memory-centric competencies. Our work is related to both profile-conditioned role-play and persona-grounded dialogue, but differs in emphasis:

rather than optimizing general grounded response generation under mixed information sources, we reframe role-play as a memory-driven cognitive task (MDRP) and evaluate models through a decomposed memory-ability lens.

### G.2 Evaluation Metrics and Diagnostic Protocols

Evaluating whether an agent remains *in character* has evolved from reference-based overlap toward more diagnostic, protocol-driven measurements of persona fidelity and consistency (Li et al., 2023; Tu et al., 2024; Wang et al., 2024a). Beyond holistic judgments, several lines of work propose finer-grained diagnostics. InCharacter evaluates personality fidelity through interview-style probing grounded in psychological traits (Wang et al., 2024c), and atomic-level evaluation reveals that response-level scores can mask localized out-of-character segments in longer generations (Shin et al., 2025). Other analyses examine failure modes and evaluator biases in role-play scoring (Zhou et al., 2025b), as well as robustness and safety issues such as character hallucination under adversarial prompts or safety degradation after role-play adaptation (Tang et al., 2025b; Zhao et al., 2025). While these metrics substantially sharpen persona-fidelity measurement, they seldom attribute errors to memory-theoretic causes (e.g., weak acquisition of provided persona facts, noisy retrieval, failure to suppress out-of-scope knowledge, or poor expression of remembered content). Our MREval targets this gap by aligning evaluation with a memory-ability taxonomy and corresponding fine-grained metrics, enabling ability-level diagnosis rather than only aggregate *in-character* scores.

### G.3 Role-Playing Methods and Memory Mechanisms

Methodologically, role-playing agents are improved via richer persona representations, alignment objectives, training/adaptation strategies, and prompt-level controllers. Representative approaches include profile–dialogue alignment (Yu et al., 2025b), self-alignment for eliciting arbitrary roles (Lu et al., 2024), configurable role controls (He et al., 2025), and coordination for established roles (Wang et al., 2025b). Training and adaptation techniques further span parameter-efficient specialization (Yu et al., 2024), role-adaptive representation learning (Yang et al., 2025), dynamic persona refinement (Yao et al., 2025), and

reward-based preference alignment (Fang et al., 2025); prompt optimization and role-aware reasoning also enhance controllability without full retraining (Duan et al., 2025; Ruangtanusak et al., 2025; Tang et al., 2025a). Within this landscape, memory and retrieval are increasingly treated as core infrastructure: RoleRAG grounds generation via structured retrieval (Wang et al., 2025c), and MOOM targets maintaining and organizing memory in ultra-long role-play dialogues (Chen et al., 2025b); some pipelines explicitly separate *what the character knows* from *how the character speaks* (Han et al., 2026). Recent work further explores *affect- or experience-conditioned* retrieval, injecting emotional signals into memory retrieval for role-playing agents (Huang et al., 2024) or retrieving counselor-specific experiential memories in multi-persona, multi-strategy adolescent counseling settings (Zhang et al., 2025a). However, these methods typically operationalize memory as an architectural component (stores, retrievers, planners) rather than a set of separable abilities that can be systematically measured and compared across models and prompting conditions. In contrast, our contribution is benchmark-centric and diagnostic: we provide MRBench+MREval for stage-wise diagnosis and systematic comparison across methods and models, and we introduce a prompt-only MR-Prompt (facet-structured persona memory with an explicit LTM–STM control protocol) as a standardized prompting condition aligned with Memory-Anchoring/Selecting/Bounding/Enacting (MA/MS/MB/ME).

## H Raw Judge Scores

To ensure transparency and facilitate replication, we report the *raw* (uncalibrated) GPT-4.1-mini judge scores for the main-text setting. Table 8 lists raw per-metric scores for the main experiment (7 models  $\times$  3 prompting conditions). For conciseness, we report only calibrated (mapped) scores in Appendix I for (i) component ablations, (ii) backbone comparisons under Base vs. MRPrompt, and (iii) the full MRPrompt results over all models. Given the per-metric, per-language linear calibration parameters ( $a_{m,\ell}, b_{m,\ell}$ ) in Table 15, the corresponding raw judge scores can be recovered via the inverse transform:  $s_{m,\ell}^{\text{raw}} = (s_{m,\ell}^{\text{cal}} - a_{m,\ell})/b_{m,\ell}$ .

## I Mapped Scores

This section reports the calibrated (mapped) GPT-4.1-mini judge scores for our ablations and model comparisons. Table 9 lists calibrated per-metric scores for component ablations, Table 10 reports the full MRPrompt results over all evaluated models (including closed-source APIs), and Table 11 provides a focused Base vs. MRPrompt comparison across representative open-/closed-source backbones.

## J Scaling and Prompting Effects: Closed-Source vs. Small-Scale LLMs on MDRP (RQ4)

**Motivation.** We study scaling effects on MDRP under MRPrompt and ask whether MRPrompt can narrow the performance gap between small-scale open models and stronger closed-source baselines, as measured by MREval.

### Closed-source models and inference setup.

We access all closed-source models via the Zhizengzeng API: GPT-5.2 (gpt-5.2), GLM-4.7 (glm-4.7), DeepSeek-Chat (deepseek-v3.2), Qwen3-Max (qwen3-max), and Doubao-Seed-1.6-250615 (doubao-seed-1-6-250615). Decoding follows the main setup (temperature  $T=0.7$  when supported; otherwise provider defaults). All models share the same MRBench instances and MREval scoring and calibration pipeline (Appendix K).

**Results and analysis.** Table 10 and Figure 4 show a capacity-shaped profile under MRPrompt: closed-source API models are often in the top tier and exhibit relatively low-variance performance across metrics. In particular, *Doubao-Seed-1.6-250615* leads all MA columns and is best on downstream enactment (ME), while also topping most MS columns. *GPT-5.2* remains close behind *Doubao* with fewer weak spots.

Crucially, Table 11 reveals that MRPrompt can substantially narrow the gap between small-scale and closed-source models when compared under realistic prompting baselines. For example, *Qwen3-8B* with MRPrompt reaches an overall Avg. Score of 8.12, matching (and slightly exceeding) the *GLM-4.7* Base baseline (8.11) and surpassing the *Qwen3-Max* Base baseline (8.08). This “small-model + MRPrompt  $\approx$  large-model + Base” effect is especially visible in upstream memory grounding/selection columns (MA/MS). At the same time,

Table 8: **Raw GPT-4.1-mini judge scores.** Per-metric ratings before calibration.

Model	Persona	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE		Avg. Score
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	
Qwen3-0.6B	Base	7.48	7.71	6.72	8.89	5.46	8.64	6.84	7.47	8.96	8.93	4.28	6.46	7.15	7.91	5.59	6.97	7.22
	Card	8.20	8.64	6.98	8.76	4.59	8.40	6.47	7.50	9.12	9.14	4.90	6.28	7.41	8.20	5.89	7.25	7.36
	MRPrompt	<b>8.40</b>	<b>8.81</b>	<b>7.75</b>	<b>8.90</b>	<b>5.92</b>	<b>9.11</b>	<b>7.11</b>	<b>7.82</b>	<b>9.38</b>	<b>9.17</b>	<b>5.51</b>	<b>6.70</b>	<b>7.54</b>	<b>8.48</b>	<b>6.31</b>	<b>7.59</b>	<b>7.78</b>
Qwen3-4B	Base	9.29	8.48	8.99	9.04	<b>8.04</b>	9.03	8.51	8.37	9.70	9.62	5.54	<b>7.74</b>	8.95	9.20	7.96	8.57	8.56
	Card	9.40	9.42	9.03	8.89	6.28	8.71	8.13	8.28	9.67	9.55	6.09	7.55	9.16	9.16	8.20	8.61	8.51
	MRPrompt	<b>9.62</b>	<b>9.52</b>	<b>9.20</b>	<b>9.22</b>	7.54	<b>9.28</b>	<b>8.66</b>	<b>8.77</b>	<b>9.84</b>	<b>9.67</b>	<b>6.60</b>	7.63	<b>9.55</b>	<b>9.51</b>	<b>8.62</b>	<b>8.86</b>	<b>8.88</b>
Qwen3-8B	Base	9.33	8.84	8.95	9.06	<b>8.12</b>	8.99	<b>8.86</b>	8.36	9.83	9.60	5.84	<b>8.43</b>	9.28	9.06	8.34	8.51	8.71
	Card	9.63	9.37	9.15	9.08	6.81	8.81	8.44	8.43	9.75	<b>9.65</b>	5.85	7.61	9.37	9.34	8.45	8.56	8.64
	MRPrompt	<b>9.75</b>	<b>9.68</b>	<b>9.29</b>	<b>9.24</b>	7.46	<b>9.52</b>	8.59	<b>8.76</b>	<b>9.90</b>	9.61	<b>6.89</b>	7.79	<b>9.62</b>	<b>9.53</b>	<b>8.74</b>	<b>8.87</b>	<b>8.95</b>
GLM-4-9B-Chat	Base	9.21	8.91	8.80	9.12	7.70	9.08	<b>8.60</b>	8.49	9.68	9.36	5.32	7.72	9.08	9.15	7.95	8.58	8.55
	Card	<b>9.61</b>	9.53	8.90	9.00	6.51	8.63	8.27	8.51	9.77	9.46	5.23	7.49	9.30	9.19	8.22	8.57	8.51
	MRPrompt	9.47	<b>9.64</b>	<b>9.06</b>	<b>9.22</b>	<b>7.98</b>	<b>9.37</b>	8.43	<b>8.71</b>	<b>9.79</b>	<b>9.64</b>	<b>6.28</b>	<b>7.79</b>	<b>9.49</b>	<b>9.51</b>	<b>8.37</b>	<b>8.85</b>	<b>8.85</b>
Llama-3-8B-Instruct	Base	8.70	7.96	8.41	<b>8.89</b>	<b>6.70</b>	9.01	<b>8.19</b>	7.53	9.67	9.44	5.36	7.54	8.89	8.53	7.39	7.85	8.13
	Card	<b>9.48</b>	8.85	8.69	8.52	5.68	8.91	7.84	7.78	9.75	9.24	6.05	7.30	9.18	8.28	7.78	7.61	8.18
	MRPrompt	9.36	<b>8.99</b>	<b>8.94</b>	8.80	6.44	<b>9.03</b>	8.12	<b>8.13</b>	<b>9.88</b>	<b>9.53</b>	<b>6.90</b>	<b>7.92</b>	<b>9.53</b>	<b>8.84</b>	<b>8.18</b>	<b>8.00</b>	<b>8.54</b>
Llama-3.2-3B-Instruct	Base	8.99	5.59	8.39	8.27	7.56	8.06	8.30	6.34	9.66	8.56	6.24	6.23	8.77	7.27	7.46	6.16	7.62
	Card	9.15	6.71	8.51	8.09	6.02	8.35	8.17	6.98	9.83	8.44	6.19	6.49	9.04	7.36	7.70	6.18	7.70
	MRPrompt	<b>9.41</b>	<b>7.90</b>	<b>8.93</b>	<b>8.42</b>	<b>7.65</b>	<b>9.02</b>	<b>8.31</b>	<b>7.34</b>	<b>9.87</b>	<b>9.11</b>	<b>7.53</b>	<b>6.90</b>	<b>9.52</b>	<b>7.92</b>	<b>8.40</b>	<b>7.03</b>	<b>8.33</b>
InternLM2.5-7B-Chat	Base	8.33	7.40	8.37	9.01	7.36	9.26	8.22	8.20	9.37	9.37	4.42	7.51	8.27	9.01	7.39	8.22	8.11
	Card	<b>8.74</b>	8.43	8.46	8.93	6.85	9.13	7.90	8.02	9.52	9.43	5.54	<b>7.81</b>	8.77	8.93	7.48	8.21	8.26
	MRPrompt	8.49	<b>8.72</b>	<b>8.66</b>	<b>9.03</b>	<b>7.93</b>	<b>9.68</b>	<b>8.24</b>	<b>8.48</b>	<b>9.61</b>	<b>9.52</b>	<b>6.13</b>	7.47	<b>9.13</b>	<b>9.09</b>	<b>7.88</b>	<b>8.30</b>	<b>8.52</b>

Table 9: **Component ablation (full results).** Mapped scores for the 8 MREval metrics.

Model	Condition	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE		Avg. Score
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	
Qwen3-4B	Base	8.65	7.80	7.54	7.99	7.97	8.63	8.59	8.23	8.79	8.53	6.55	<b>7.04</b>	7.55	7.43	7.20	7.17	7.85
	+Protocol	8.69	8.26	7.57	8.15	<b>8.06</b>	8.57	<b>8.78</b>	8.25	8.75	8.46	6.67	6.97	7.74	7.41	7.36	7.14	7.93
	+Schema	8.87	8.55	7.59	8.18	7.41	8.76	8.48	8.31	8.82	8.45	6.83	6.96	7.94	7.54	7.54	7.29	7.97
	MRPrompt	<b>8.88</b>	<b>8.61</b>	<b>7.69</b>	<b>8.23</b>	7.61	<b>8.81</b>	8.69	<b>8.57</b>	<b>8.85</b>	<b>8.56</b>	<b>6.84</b>	6.99	<b>8.07</b>	<b>7.63</b>	<b>7.73</b>	<b>7.42</b>	<b>8.07</b>
GLM-4-9B-Chat	Base	8.59	8.13	7.41	8.10	7.73	8.67	<b>8.65</b>	8.34	8.78	8.36	6.49	7.03	7.66	7.40	7.19	7.18	7.86
	+Protocol	8.67	8.23	7.43	8.09	7.81	8.60	8.62	8.32	8.81	8.41	6.50	6.99	7.64	7.35	7.22	7.22	7.87
	+Schema	<b>8.83</b>	8.57	<b>7.66</b>	8.21	7.84	8.86	8.55	8.51	<b>8.84</b>	8.46	6.73	6.86	<b>8.13</b>	<b>7.57</b>	<b>7.61</b>	7.36	8.04
	MRPrompt	8.77	<b>8.70</b>	7.59	<b>8.23</b>	<b>7.93</b>	<b>8.88</b>	8.53	<b>8.52</b>	8.83	<b>8.54</b>	<b>6.76</b>	<b>7.07</b>	8.02	<b>7.63</b>	7.53	<b>7.41</b>	<b>8.06</b>

the benefit of MRPrompt is not uniform across backbones (e.g., *Qwen3-Max* shows comparable Avg. Score under Base vs. MRPrompt), suggesting that stronger proprietary models may already internalize parts of the structure that MRPrompt explicitly scaffolds.

Despite the narrowed gap, downstream enactment remains the main separator for the smallest models: *Qwen3-0.6B* shows sharp degradation on ME (both MAC and HLE) even under MRPrompt, indicating that structured prompting can stabilize *memory use* but cannot fully compensate for limited *surface enactment* capacity. A second consistent pattern is that constraint robustness is comparatively harder than other upstream dimensions: across models, CR tends to sit below AL, suggesting that resisting boundary pressure and maintaining rule-consistent behavior remains a key failure mode even when persona anchoring is strong. Fi-

nally, the scaling trends are broadly aligned across English and Chinese, while still exhibiting metric-level language variation (e.g., FA is often higher in Chinese for multiple models).

## K Automatic Judge Validation, Score Calibration, and Supplemental Robustness Check (RQ5)

*Is the LLM-as-a-Judge reliable for MREval, how do we align its scores to the human scale, and are the conclusions robust beyond the primary calibration setup?* We use GPT-4.1-mini (API) to score all eight MREval metrics, and validate it against bilingual human ratings. To improve scoring stability, we set the decoding temperature to 0 and use default settings for the remaining generation parameters. This appendix summarizes: (i) judge-human agreement on the primary 800-instance bilingual reliability set (Tables 12–14); (ii) a per-metric cal-

Table 10: **RQ4 results (MRPrompt)**. Mapped scores for the 8 MREval metrics across all models.

Model	Persona	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE	
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
GPT-5.2	MRPrompt	8.87	8.18	7.73	8.20	<u>8.14</u>	<u>9.04</u>	8.66	8.43	<b>8.92</b>	<u>8.66</u>	<b>7.47</b>	<u>7.35</u>	8.21	7.55	7.85	7.21
GLM-4.7	MRPrompt	8.83	8.38	7.75	<u>8.38</u>	7.98	9.03	8.67	8.46	8.89	8.65	7.12	6.99	8.20	<u>7.71</u>	7.71	7.36
DeepSeek-Chat	MRPrompt	8.89	8.37	7.73	8.30	7.02	7.64	8.42	7.94	8.89	8.63	<u>7.33</u>	<b>7.46</b>	8.20	7.54	7.63	7.21
Qwen3-Max	MRPrompt	<u>8.97</u>	8.12	<u>7.88</u>	8.34	7.76	8.36	8.63	8.13	<b>8.92</b>	<b>8.69</b>	7.19	7.26	<u>8.24</u>	7.61	<u>7.88</u>	7.20
Doubao-Seed-1.6-250615	MRPrompt	<b>9.13</b>	<b>8.85</b>	<b>8.05</b>	<b>8.43</b>	<b>8.47</b>	<b>9.10</b>	<b>8.94</b>	<b>8.73</b>	<u>8.91</u>	8.60	6.96	7.25	<b>8.36</b>	<b>7.83</b>	<b>8.18</b>	<b>7.54</b>
Qwen3-0.6B	MRPrompt	8.02	8.06	6.67	7.81	6.46	8.69	7.63	7.77	8.65	8.23	6.54	6.56	6.33	6.97	5.86	6.35
Qwen3-4B	MRPrompt	8.88	8.61	7.69	8.23	7.61	8.81	<u>8.69</u>	<u>8.57</u>	8.85	8.56	6.84	6.99	8.07	7.63	7.73	7.41
Qwen3-8B	MRPrompt	<u>8.97</u>	<u>8.73</u>	7.76	8.25	7.56	8.99	<u>8.64</u>	8.56	8.88	8.52	6.92	7.07	8.13	7.64	7.83	<u>7.42</u>
GLM-4-9B-Chat	MRPrompt	8.77	8.70	7.59	8.23	7.93	8.88	8.53	8.52	8.83	8.54	6.76	7.07	8.02	7.63	7.53	7.41
Llama-3-8B-Instruct	MRPrompt	8.69	8.20	7.51	7.68	6.83	8.63	8.32	8.03	8.87	8.47	6.93	7.13	8.05	7.20	7.37	6.69
Llama-3.2-3B-Instruct	MRPrompt	8.73	7.35	7.50	7.19	7.69	8.63	8.45	7.36	8.87	8.19	7.10	6.65	8.04	6.60	7.55	5.88
InternLM2.5-7B-Chat	MRPrompt	8.08	7.99	7.31	7.98	7.89	<b>9.10</b>	8.40	8.33	8.75	8.46	6.71	6.92	7.71	7.36	7.13	6.95

Table 11: **RQ4 backbone comparisons (full results)**. Mapped scores for Base vs. MRPrompt. Best/second-best are bold/underlined.

Model	Condition	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE		Avg. Score
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	
Qwen3-8B	Base	8.67	8.08	7.52	8.02	8.03	8.60	<b>8.83</b>	8.23	8.85	8.51	6.63	7.36	7.83	7.34	7.50	7.12	7.95
GLM-4-9B-Chat	Base	8.59	8.13	7.41	8.10	<u>7.73</u>	8.67	8.65	8.34	8.78	8.36	6.49	7.03	7.66	7.40	7.19	7.18	7.86
Qwen3-8B	MRPrompt	<b>8.97</b>	<b>8.73</b>	7.76	8.25	7.56	8.99	8.64	<u>8.56</u>	8.88	8.52	6.92	7.07	8.13	7.64	7.83	<b>7.42</b>	<u>8.12</u>
GLM-4-9B-Chat	MRPrompt	8.77	<u>8.70</u>	7.59	8.23	7.93	8.88	8.53	8.52	8.83	8.54	6.76	7.07	8.02	7.63	7.53	<u>7.41</u>	8.06
Qwen3-Max	Base	<u>8.96</u>	<u>8.05</u>	<u>7.84</u>	<u>8.37</u>	<b>8.45</b>	8.29	8.75	8.14	8.87	8.49	6.77	<b>7.41</b>	<u>8.21</u>	7.59	<u>7.87</u>	<u>7.25</u>	8.08
GLM-4.7	Base	8.81	8.29	<u>7.77</u>	<u>8.36</u>	7.98	<u>9.02</u>	<u>8.77</u>	<b>8.62</b>	8.88	8.51	6.71	<u>7.38</u>	<u>7.98</u>	<u>7.69</u>	<u>7.66</u>	7.27	8.11
Qwen3-Max	MRPrompt	<b>8.97</b>	8.12	<b>7.88</b>	8.34	7.76	8.36	8.63	8.13	<b>8.92</b>	<b>8.69</b>	<b>7.19</b>	7.26	<b>8.24</b>	7.61	<b>7.88</b>	7.20	8.07
GLM-4.7	MRPrompt	8.83	8.38	7.75	<b>8.38</b>	7.98	<b>9.04</b>	8.67	8.46	<u>8.89</u>	<u>8.65</u>	<u>7.12</u>	6.99	8.20	<b>7.71</b>	7.71	7.36	<b>8.13</b>

ibration that maps judge scores onto the human rating scale (Table 15); and (iii) a supplemental post-hoc robustness check using four independent human raters and an alternative judge model on a representative 100-instance subset (Tables 19)

**Reliability set.** We construct a controlled reliability set by sampling 100 evaluated instances per metric (stratified across evaluated models and persona settings). Each metric set contains 50 Chinese and 50 English instances, yielding 800 total instances (8 metrics  $\times$  100). For each instance, the judge is given the same inputs as in the main evaluation (persona file, dialogue context, and model response) and produces a 1–10 score using the identical rubric and prompt.

**Human reference for primary calibration.** The primary calibration set is annotated by a single bilingual annotator, who is familiar with all 16 novels and the metric rubrics. During scoring, the annotator was blind to the underlying model, persona condition, and GPT-4.1-mini scores, and rated all samples in randomized order separately for each metric. We emphasize that this human annotation is used as a targeted calibration-and-validation step for the automatic judge, rather than as a compre-

hensive manual re-evaluation of all model outputs. This design allows us to estimate judge–human agreement and fit per-metric calibration parameters at manageable annotation cost.

**Agreement on the primary calibration set.** For each metric–language pair, we compute Pearson  $r$ , Spearman  $\rho$ , and Kendall  $\tau$  between GPT-4.1-mini scores and human scores, reported in Tables 12 and 13. We also report the mean scores of GPT-4.1-mini and the human annotator in Table 14. Overall, correlations are moderate-to-strong and statistically significant, supporting the utility of GPT-4.1-mini as a scalable evaluator for MREval.

**Calibration.** To correct mild systematic bias, we fit a separate least-squares linear mapping for each metric–language pair:

$$\text{HUMANScore} = a_{m,\ell} + b_{m,\ell} \cdot \text{JUDGEScore}, \quad (1)$$

using the 50 samples in that language from the primary 800-instance reliability set as training data. We then apply  $(a_{m,\ell}, b_{m,\ell})$  to calibrate all GPT-4.1-mini scores reported in the main experiments. The fitted parameters for all metric–language pairs are listed in Table 15. Here,  $a_{m,\ell}$  captures additive bias and  $b_{m,\ell}$  captures scale sensitivity.

**Visual check.** Figure 12 provides per-sample comparisons for representative metric–language pairs, plotting GPT-4.1-mini and human scores and additionally showing the fitted calibration line.

**Supplemental external robustness check.** We conduct an additional post-hoc reliability study on a representative 100-instance subset. In this supplementary analysis, the same evaluation instances are independently rated by four human raters, and we additionally score them with an alternative judge model, Qwen3-32B. Importantly, this robustness check is used only for external validation: it does not alter the original calibration parameters or the main experimental results reported in the paper.

**Inter-human consistency and judge-model sensitivity.** On the 100-instance subset, the four human raters show strong group-level consistency, with Cronbach’s  $\alpha = 0.8583$  and  $\text{ICC}(2,k) = 0.8577$ , while single-rater reliability is moderate ( $\text{ICC}(2,1) = 0.6012$ ). This suggests that individual rater noise exists but is substantially reduced when multiple raters are aggregated. Using the mean score of the four raters as the human consensus, GPT-4.1-mini remains positively aligned with human judgment (Spearman  $\rho = 0.3938$ , Pearson  $r = 0.6299$ ), and the alternative judge model Qwen3-32B shows similar directional agreement (Spearman  $\rho = 0.4887$ , Pearson  $r = 0.6418$ ). These results suggest that our conclusions are not tied to a single idiosyncratic judge model. Also, the supplemental 100-instance robustness check is used only to verify external consistency of the evaluation protocol and judge model choice; it does not replace the primary calibration procedure above.

## L Language-split Ability Correlations (RQ6)

This appendix provides language-split correlation analyses that complement the pooled results in §4.6. We compute ability-level scores by averaging the two metrics within each ability (MA/MS/MB/ME), and restrict to instances where all four abilities are available to ensure comparability. We then compute Pearson correlations separately for English ( $n=1470$ ) and Chinese ( $n=1092$ ).

**Pearson correlation matrices.** Table 16 reports the English correlation matrix, and Table 17 reports the Chinese matrix. Both exhibit the same qualitative pattern as the pooled analysis: positive,

weak-to-moderate coupling across abilities, with the strongest association typically involving ME.

**Upstream–ME correlations.** We additionally report language-split Pearson correlations between an aggregated upstream score (mean of MA/MS/MB) and ME:  $r_{\text{EN}}=0.34$  and  $r_{\text{ZH}}=0.38$ .

## M Efficiency Analysis: Token Budget (RQ7)

To verify that MRPrompt gains are not driven by longer prompts, we measure token usage under the three persona conditions (Base, Card, MRPrompt) on the same shared STM pool. Table 20 reports the average prompt/completion/total tokens (English/Chinese) under each condition, aggregated across instances. Overall, MRPrompt does not substantially increase total tokens compared to the baselines, indicating that its gains come primarily from structured memory representation and protocol guidance rather than longer prompts.

## N Latency Analysis and Deployment Feasibility (RQ8)

MRPrompt is executed as a *single-pass, prompt-only generation* method: it does not rely on a multi-call pipeline, external tool use, or explicit “thinking”/reasoning stages during inference.

To evaluate deployment feasibility more directly, we measure *per-turn end-to-end latency*, defined as the elapsed wall-clock time from sending an inference request to receiving the final generated token. For each model, we report the average latency under English and Chinese evaluation settings. These measurements are intended as a practical reference for deployment-oriented comparison.

**Results.** Table 18 reports the measured latency. We observe a clear pattern: smaller open backbones are substantially faster than larger or closed-source models in our setup. For example, Qwen3-4B, Qwen3-8B, and GLM-4-9B all remain around 2–3 seconds per turn, whereas larger closed models such as GLM-4.7, Qwen3-Max, and GPT-5.2 are notably slower. This pattern is consistent across both English and Chinese settings.

**Implications for deployment.** Combining these latency measurements with the main results yields an encouraging deployment takeaway. MRPrompt substantially improves smaller backbones on MREval, and in several cases makes them

Table 12: Correlations between human ratings and GPT-4.1-mini ratings for MA/MS metrics (English and Chinese).

	MA-SI		MA-AF		MS-FA		MS-FU	
	en	zh	en	zh	en	zh	en	zh
Pearson $r$	0.814	0.723	0.660	0.577	0.796	0.853	0.711	0.877
$p$ -value	$6.812 \times 10^{-13}$	$2.980 \times 10^{-9}$	$1.830 \times 10^{-7}$	$1.171 \times 10^{-5}$	$4.704 \times 10^{-12}$	$3.745 \times 10^{-15}$	$7.131 \times 10^{-9}$	$7.186 \times 10^{-17}$
Spearman $\rho$	0.554	0.570	0.532	0.552	0.688	0.543	0.614	0.662
$p$ -value	$2.993 \times 10^{-5}$	$1.574 \times 10^{-5}$	$7.082 \times 10^{-5}$	$3.251 \times 10^{-5}$	$3.464 \times 10^{-8}$	$4.702 \times 10^{-5}$	$2.115 \times 10^{-6}$	$1.624 \times 10^{-7}$
Kendall $\tau$	0.515	0.528	0.514	0.516	0.595	0.514	0.584	0.646
$p$ -value	$5.470 \times 10^{-5}$	$4.934 \times 10^{-5}$	$1.283 \times 10^{-4}$	$9.377 \times 10^{-5}$	$3.497 \times 10^{-7}$	$6.361 \times 10^{-5}$	$1.238 \times 10^{-5}$	$1.384 \times 10^{-6}$

Table 13: Correlations between human ratings and GPT-4.1-mini ratings for MB/ME metrics (English and Chinese).

	MB-AL		MB-CR		ME-MAC		ME-HLE	
	en	zh	en	zh	en	zh	en	zh
Pearson $r$	0.860	0.547	0.526	0.711	0.715	0.822	0.776	0.858
$p$ -value	$1.263 \times 10^{-15}$	$3.881 \times 10^{-5}$	$8.696 \times 10^{-5}$	$7.084 \times 10^{-9}$	$5.339 \times 10^{-9}$	$2.511 \times 10^{-13}$	$3.669 \times 10^{-11}$	$1.716 \times 10^{-15}$
Spearman $\rho$	0.524	0.508	0.538	0.668	0.512	0.520	0.657	0.536
$p$ -value	$9.540 \times 10^{-5}$	$1.636 \times 10^{-4}$	$5.669 \times 10^{-5}$	$1.157 \times 10^{-7}$	$1.425 \times 10^{-4}$	$1.093 \times 10^{-4}$	$2.171 \times 10^{-7}$	$5.933 \times 10^{-5}$
Kendall $\tau$	0.506	0.489	0.471	0.588	0.478	0.493	0.623	0.503
$p$ -value	$2.298 \times 10^{-4}$	$3.554 \times 10^{-4}$	$7.045 \times 10^{-5}$	$2.270 \times 10^{-6}$	$2.501 \times 10^{-4}$	$1.839 \times 10^{-4}$	$1.273 \times 10^{-6}$	$9.909 \times 10^{-5}$

competitive with much stronger baselines, while preserving materially lower latency. This makes small-to-mid-sized backbones an attractive operating point for real-time or near-real-time role-playing systems, such as gaming NPCs, where both character quality and interaction responsiveness matter.

**Caveat.** These latency values should be interpreted as *deployment-oriented measurements* rather than fully controlled systems benchmarks. In particular, closed-source APIs may differ in backend infrastructure, request routing, batching strategy, and serving load. Therefore, the absolute numbers are environment-dependent, but the overall trend remains informative for practical model selection.

## O Additional Qualitative Case Studies (RQ9)

This appendix reports the qualitative materials for RQ9, which complement the main quantitative results by inspecting concrete MDRP behaviors under MREval. We analyze two representative characters (Tom Sawyer; Charles Darnay) using *GLM-4-9B-Chat*, comparing Base, Card, and MRPrompt. For each case, we present the extended dialogue context ( $\mathcal{M}_S$ ), the gold continuation  $\hat{y}^{\text{gold}}$ , and the three model outputs  $\hat{y}$ .

**Tom Sawyer: facet-aligned success.** Table 21 illustrates the “whitewashing” episode, where Tom should theatrically concede only after Ben has fully

taken the bait. Under Base, the model follows the plot but injects meta narration (e.g., “breaking character”) and adopts an apologetic tone, diluting Tom’s manipulative bravado. Card improves liveliness but remains generic, failing to clearly preserve the “exclusive opportunity” framing that motivates Ben’s eagerness. In contrast, MRPrompt yields a short, in-character concession (“But don’t say I didn’t warn you”) that preserves Tom’s playful dominance and aligns with the intended facet, qualitatively consistent with the MS/ME gains observed in RQ1–RQ3.

**Charles Darnay: residual flattening.** Table 22 examines Darnay’s promise to withhold his present name. In the novel, this moment carries dramatic irony and latent moral tension; a faithful continuation should acknowledge the oddity and weight of the constraint while remaining deferential. All three conditions largely smooth this tension into a generic vow. Base is polite but non-committal about the unusual request; Card slightly misattributes the secrecy (“your secret”); MRPrompt best tracks the object of the promise (the *name*) and ties compliance to Lucie’s happiness, yet still under-expresses the implied inner unease. This case highlights a remaining failure mode: even when persona voice is grounded, models may under-realise *implicit* narrative constraints (suppressed information and deferred revelation) as rhetorically salient dialogue.

Table 14: Mean scores of GPT-4.1-mini and the human annotator for each metric and language on the reliability set.

	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
GPT-4.1-mini	9.22	9.46	9.10	9.12	7.32	9.26	8.56	8.64	9.64	9.82	7.10	7.86	9.54	9.12	8.34	8.44
Human	8.60	8.56	7.62	8.10	7.46	8.80	8.62	8.46	8.76	8.66	6.98	7.10	8.06	7.38	7.50	7.06

Table 15: Linear calibration parameters ( $a_{m,\ell}, b_{m,\ell}$ ) for each metric and language.

	MA-SI		MA-AF		MS-FA		MS-FU		MB-AL		MB-CR		ME-MAC		ME-HLE	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
$a$ (intercept)	2.133	1.238	1.189	-3.676	2.245	2.113	2.793	1.136	4.464	2.226	5.016	3.474	-0.183	1.488	0.757	-0.018
$b$ (slope)	0.701	0.774	0.707	1.291	0.712	0.722	0.681	0.848	0.446	0.655	0.277	0.461	0.864	0.646	0.809	0.839

Table 16: Pearson correlation matrix of ability scores on the English subset. ( $n=1470$ )

	MA	MS	MB	ME
MA	1.0000	0.1487	0.0993	0.3495
MS	0.1487	1.0000	0.0794	0.2286
MB	0.0993	0.0794	1.0000	0.1059
ME	0.3495	0.2286	0.1059	1.0000

Table 17: Pearson correlation matrix of ability scores on the Chinese subset. ( $n=1092$ )

	MA	MS	MB	ME
MA	1.0000	0.2902	0.0892	0.3940
MS	0.2902	1.0000	0.0365	0.2663
MB	0.0892	0.0365	1.0000	0.1421
ME	0.3940	0.2663	0.1421	1.0000

**Takeaways.** Overall, the case studies support the main quantitative story. When success is primarily driven by *scene-conditioned facet selection* and persona-faithful micro-decisions (Tom), MRPrompt produces more locally appropriate, in-character continuations than Base/Card. When success depends on *subtle management of latent long-term information* (Darnay), even MRPrompt can converge to fluent but flattened outputs, consistent with RQ4: upstream gains tend to improve ME, but do not guarantee fine-grained realization of suppressed knowledge and internal conflict.

**Persona references.** Brief persona descriptions for these two characters are provided in Appendix C.2.1 and Appendix C.2.2; full persona files are included in the supplementary material.

Table 18: Per-turn end-to-end latency (seconds) for representative backbones in English and Chinese.

Model	EN (s)	ZH (s)
GLM-4.7	4.58	6.32
Qwen3-Max	5.45	6.18
DeepSeek-Chat	2.87	4.13
GPT-5.2	7.22	9.47
Qwen3-0.6B	2.47	2.55
Qwen3-4B	2.73	2.39
Qwen3-8B	2.83	2.77
GLM-4-9B	2.42	1.75

## P MREval Metrics and Rubrics

Table 23 summarizes the eight MREval metrics and their corresponding scoring rubrics. Each metric is rated on a 1–10 Likert-style scale (higher is better), using 1/5/10 anchor descriptions.

## Q Prompts

### Q.1 Prompt Templates for Character Profiles

Figures 13–15 present the prompt templates used to construct (i) narrative character profiles, (ii) semi-structured persona cards, and (iii) our facet-structured long-term memory (LTM) personas.

### Q.2 Prompt Templates for MRBench Split Construction (MA/MB)

Figures 16 and 17 show the prompt templates used to construct the evaluation instances for the MA and MB splits in MRBench.

### Q.3 Prompt Templates for Character Role-playing

Figures 18 and 19 report the shared standard role-playing instruction and our specific LTM–STM control protocol.

Table 19: Pairwise Pearson correlations among two judge models and four human raters on the supplemental 100-instance subset.

	GPT-4.1-mini	Qwen3-32B	Human <sub>1</sub>	Human <sub>2</sub>	Human <sub>3</sub>	Human <sub>4</sub>
GPT-4.1-mini	1.00	0.79	0.75	0.49	0.43	0.45
Qwen3-32B	0.79	1.00	0.67	0.50	0.47	0.51
Human <sub>1</sub>	0.75	0.67	1.00	0.51	0.56	0.62
Human <sub>2</sub>	0.49	0.50	0.51	1.00	0.70	0.54
Human <sub>3</sub>	0.43	0.47	0.56	0.70	1.00	0.69
Human <sub>4</sub>	0.45	0.51	0.62	0.54	0.69	1.00

Table 20: **Average token usage on the shared STM pool.** Prompt ( $\mathcal{M}_L+\mathcal{M}_S+\text{constraints}$ ), completion ( $\hat{y}$ ), and total tokens per model–persona–language (en/zh).

Model	Persona	Prompt		Comp		Total	
		en	zh	en	zh	en	zh
Qwen3-4B	Base	3726.14	2873.16	65.39	53.02	3791.53	2926.19
	Card	2866.16	2585.43	60.61	47.22	2926.77	2632.65
	MRPrompt	3037.74	2825.49	65.40	56.79	3103.14	2882.28
GLM-4-9B-Chat	Base	3673.54	2699.86	56.35	49.35	3729.89	2749.21
	Card	2822.13	2455.76	58.46	48.02	2880.59	2503.78
	MRPrompt	2994.50	2694.10	53.13	52.60	3047.63	2746.70

#### Q.4 Prompt Templates for MREval Judging

Since the eight MREval metrics target different aspects of role-playing behavior, we require a set of metric-specific judging prompt templates. Due to space limitations, we omit these templates from the appendix; the full prompt specifications are provided in the supplementary material.

Table 21: **Tom Sawyer case.** Full context and continuations under three persona conditions.

Element	Text
<b>Full dialogue context</b>	<p><b>Ben Rogers:</b> (approaches Tom with a skip in his step) Hello, old chap, you got to work, hey?</p> <p><b>Tom Sawyer:</b> (pretends to be focused on his work, then looks up with feigned surprise) Why, it's you, Ben! I warn't noticing.</p> <p><b>Ben Rogers:</b> Say—I'm going in a-swimming, I am. Don't you wish you could? But of course you'd druther work—wouldn't you? Course you would!</p> <p><b>Tom Sawyer:</b> (looks thoughtful) What do you call work?</p> <p><b>Ben Rogers:</b> (looks confused) Why, ain't that work?</p> <p><b>Tom Sawyer:</b> (resumes whitewashing casually) Well, maybe it is, and maybe it ain't. All I know, is, it suits Tom Sawyer.</p> <p><b>Ben Rogers:</b> (looks skeptical) Oh come, now, you don't mean to let on that you like it?</p> <p><b>Tom Sawyer:</b> (continues painting with exaggerated care) Like it? Well, I don't see why I oughtn't to like it. Does a boy get a chance to whitewash a fence every day?</p> <p><b>Ben Rogers:</b> (watches Tom intently) Say, Tom, let me whitewash a little.</p> <p><b>Tom Sawyer:</b> (pretends to consider reluctantly) No—no—I reckon it wouldn't hardly do, Ben. You see, Aunt Polly's awful particular about this fence—right here on the street, you know—but if it was the back fence I wouldn't mind and she wouldn't. Yes, she's awful particular about this fence; it's got to be done very careful; I reckon there ain't one boy in a thousand, maybe two thousand, that can do it the way it's got to be done.</p> <p><b>Ben Rogers:</b> (eagerly) No—is that so? Oh come, now—lemme just try. Only just a little—I'd let you, if you was me, Tom.</p> <p><b>Tom Sawyer:</b> (hesitates dramatically) Ben, I'd like to, honest injun; but Aunt Polly—well, Jim wanted to do it, but she wouldn't let him; Sid wanted to do it, and she wouldn't let Sid. Now don't you see how I'm fixed? If you was to tackle this fence and anything was to happen to it—</p> <p><b>Ben Rogers:</b> (interrupts eagerly) Oh, shucks, I'll be just as careful. Now lemme try. Say—I'll give you the core of my apple.</p> <p><b>Tom Sawyer:</b> (pretends to be reluctant) Well, here—No, Ben, now don't. I'm afeard—</p> <p><b>Ben Rogers:</b> (desperately) I'll give you all of it!</p>
<b>Gold continuation</b>	Tom Sawyer: [Success! Time to graciously concede.] (hands over the brush with feigned reluctance) Well, I suppose... if you're sure you want to, Ben.
<b>Base output</b>	Tom Sawyer: (smiles, breaking character) OK, Ben. But just a little bit, and then you'll have to stop. I don't want to be accused of sharing if I didn't.
<b>Card output</b>	Tom Sawyer: (playfully) Alright, Ben! I reckon you can handle a little fence painting. Just don't mess it up, or I'll have to whitewash the whole town next!
<b>MRPrompt output</b>	Tom Sawyer: (smiles, gives in playfully) Alright, Ben! But don't say I didn't warn you. (starts to hand over the brush)

```

MRPrompt Tom Sawyer
{
  "name": "Tom Sawyer",
  "Nickname": "Tom",
  "Relationships": [
    {
      "name": "Aunt Polly",
      "relationship": "guardian",
      "attitude": "protective, exasperated, loving"
    },
    {
      "name": "Huck Finn",
      "relationship": "closest friend and confidant",
      "attitude": "loyal, adventurous, brotherly"
    },
    {
      "name": "Becky Thatcher",
      "relationship": "romantic interest",
      "attitude": "infatuated, idealizing, remorseful"
    },
    ...
  ],
  "global_summary": "Tom Sawyer is a restless, imaginative boy whose life on the Mississippi River is marked by a string of mischievous adventures, dramatic gestures, and a stubborn pursuit of justice...",
  "Personality": {
    "core_traits": [
      {
        "trait": "Imaginative and inventive",
        "desc": "Tom constantly dreams up elaborate schemes, stories, and games, transforming mundane situations into opportunities for excitement and profit, as seen in his whitewashing ploy and island adventures."
      },
      {
        "trait": "Manipulative charm",
        "desc": "He is skilled at convincing others to do his bidding, using theatrical flair and psychological insight to turn situations to his advantage, particularly with peers and authority figures."
      },
      ...
    ],
  },
  "scene_facets": [
    {
      "title": "Scheming and Improvisation for Personal Gain",
      "time_scope": ["early_life"],
      "situation": "Whenever Tom is tasked with chores, faces dull routines, or sees an opportunity to manipulate, he turns the situation into a game or a profit-making venture through clever improvisation.",
      "social_role": ["trickster", "negotiator", "peer influencer"],
      "emotional_state": "playful, sly, self-satisfied",
      "behavior_pattern": "Feigns enjoyment, exaggerates his abilities, persuades others to take over his burdens while extracting rewards; uses storytelling and performance to reshape perceptions.",
      "thinking_pattern": "Inventiveness and fun > obedience; cleverness > hard work; values being perceived as exceptional.",
      "conflict_with_core": "Aligned with his imaginative and manipulative traits, but occasionally risks alienating peers or drawing negative attention from authority.",
      "source_scenes": ["The Famous Whitewashing Scheme", "Pirate games on Jackson's Island"],
      "cue_phrases": ["boring chores", "chance to impress", "outsmarting friends", "exclusive opportunity"]
    },
    ...
  ]
}
}

```

Figure 8: Facet-structured LTM excerpt for Tom Sawyer.

```

Base Charles Darnay

[Overview]
Charles Darnay's life unfurls as a journey between two worlds—aristocratic France and revolutionary England—marked by moments of quiet conviction, moral struggle, and sacrificial love...(abridged)

[Scene 1: Darnay's trial for treason at the Old Bailey]
The trial takes place in a crowded London courtroom, charged with tension and suspicion toward foreigners. Darnay sits in the dock, accused of passing British secrets to France, facing the threat of execution...(abridged)

[Scene 2: Confession to Dr. Manette on his wedding morning]
On the morning of his marriage to Lucie, Darnay requests a private audience with Dr. Manette in the Manette home...(abridged)

[Scenes 3-5 omitted for brevity]

[Scene 6: Rescue by Sydney Carton and aftermath]
On the eve of execution, Darnay is drugged and smuggled out of prison by Sydney Carton, who takes his place at the guillotine...(abridged)

```

Figure 9: Narrative LTM excerpt for Charles Darnay.

```

Card Charles Darnay
{
  "name": "Charles Darnay",
  "Nickname": "Charles Darnay",
  "Relationships": [
    { "name": "Lucie Manette", "relationship": "wife" },
    { "name": "Dr. Alexandre Manette", "relationship": "father-in-law / mentor" },
    { "name": "Sydney Carton", "relationship": "saviour / rival" },
    ...
  ],
  "global_summary": "Charles Darnay, born Charles Evrémone, grows up amid the oppression and cruelty of the French aristocracy. Haunted by his family's legacy, he rejects his birthright and escapes to England, where he builds a new identity as a French tutor...",
  "Personality": {
    "core_traits": [
      { "trait": "principled" },
      { "trait": "self-sacrificing" },
      { "trait": "restrained" },
      { "trait": "honest" },
      ...
    ],
    "scene_facets": [
      {
        "title": "Trial for Treason at the Old Bailey",
        "situation": "In a crowded London courtroom, Charles Darnay stands trial for treason, accused of passing British secrets to France. He faces the threat of execution, with evidence against him largely circumstantial and the atmosphere charged with suspicion against foreigners. Throughout the proceedings, Darnay remains composed and courteous, answering questions with precision and never implicating others. He draws strength from the presence of Dr. Manette and Lucie in the audience. Relief floods him when Sydney Carton cleverly exposes the weakness of the prosecution's case, leading to Darnay's acquittal. Afterward, Darnay thanks Carton and Stryver quietly, avoiding dramatic displays. The experience leaves him shaken and newly aware of his precarious position in England, but his resolve to live honorably is strengthened.",
        "emotional_state": "anxiety, restraint, relief",
        "behavior_pattern": "Maintains calm, dignified speech and refuses to implicate others even under pressure."
      },
      ...
    ]
  }
}

```

Figure 10: Semi-structured LTM excerpt for Charles Darnay.

Table 22: **Charles Darnay case.** Full context and continuations under three persona conditions.

Element	Text
<b>Full dialogue context</b>	<p><b>Charles Darnay:</b> Doctor Manette, I have come to speak to you about a matter of great importance to me.</p> <p><b>Doctor Manette:</b> (with evident constraint) Is Lucie the topic?</p> <p><b>Charles Darnay:</b> Yes, sir. I want to express that I love your daughter fondly, dearly, disinterestedly, devotedly.</p> <p><b>Doctor Manette:</b> (turning away) Not that, sir! Let that be! I adjure you, do not recall that!</p> <p><b>Charles Darnay:</b> Doctor Manette, I understand the unique bond you share with Lucie. I would never seek to come between you.</p> <p><b>Doctor Manette:</b> (after a pause) I do not doubt your loving Lucie; you may be satisfied of it.</p> <p><b>Charles Darnay:</b> Sir, I would never dream of separating Lucie from you. My hope is to join your family, not divide it.</p> <p><b>Doctor Manette:</b> Have you spoken to Lucie?</p> <p><b>Charles Darnay:</b> No, sir. Nor have I written to her. I wanted to speak with you first.</p> <p><b>Doctor Manette:</b> Your self-denial is to be referred to your consideration for her father. Her father thanks you.</p> <p><b>Charles Darnay:</b> Doctor Manette, I know the depth of Lucie's love for you. I only hope to add to that love, not diminish it in any way.</p> <p><b>Doctor Manette:</b> If she should ever tell me that you are essential to her perfect happiness, I will give her to you. But you must promise me something.</p> <p><b>Charles Darnay:</b> Of course, sir. What is it you wish me to promise?</p> <p><b>Doctor Manette:</b> You must not tell me your present name. If your suit should prosper, you shall tell me on your marriage morning. Do you promise?</p>
<b>Gold continuation</b>	Charles Darnay: [This is an odd request, but I must respect it.] Willingly, Doctor Manette. You have my word.
<b>Base output</b>	Charles Darnay: Yes, sir, I promise. And I assure you, it is with the greatest respect and honor that I make this promise.
<b>Card output</b>	Charles Darnay: I promise, Doctor Manette. I will keep your secret until our marriage day, if it be as you wish.
<b>MRPrompt output</b>	Charles Darnay: I promise, sir. My name will remain my own until the day we marry. It is but a small thing, in the grand scheme of your approval and Lucie's happiness.

**MRPrompt Charles Darnay**

```
{
  "name": "Charles Darnay",
  "Nickname": "Charles Darnay",
  "Relationships": [
    {
      "name": "Lucie Manette",
      "relationship": "wife",
      "attitude": "devoted, protective"
    },
    {
      "name": "Dr. Alexandre Manette",
      "relationship": "father-in-law",
      "attitude": "respectful, honest"
    },
    {
      "name": "Sydney Carton",
      "relationship": "friend and savior",
      "attitude": "grateful, humbled"
    },
    ...
  ],
  "global_summary": "Charles Darnay, born Charles Evrémonte, is a man shaped by the struggle to distance himself from his family's aristocratic abuses and to build a life of integrity during the chaos of the French Revolution...",
  "Personality": {
    "core_traits": [
      {
        "trait": "Principled Integrity",
        "desc": "Darnay consistently chooses honesty and moral responsibility, even when it threatens his safety or social standing, as shown in his renunciation of his family name and transparent dealings with Dr. Manette."
      },
      {
        "trait": "Selflessness",
        "desc": "He habitually prioritizes the happiness and wellbeing of others over his own desires, evident in his approach to love and willingness to sacrifice for Lucie and her family."
      },
      ...
    ],
  },
  "scene_facets": [
    {
      "title": "Rejecting Privilege and Pursuing Moral Independence",
      "time_scope": ["early_life", "pre-revolution"],
      "situation": "Whenever confronted with the legacy of his aristocratic family, Darnay chooses to distance himself from privilege and injustice, openly renouncing his heritage and seeking to establish an honorable life apart from it.",
      "social_role": ["scion", "confessor", "prospective son-in-law"],
      "emotional_state": "conflicted, resolute, remorseful",
      "behavior_pattern": "Speaks quietly but with conviction; initiates difficult conversations; refuses secrecy or denial; prioritizes transparency even when uncomfortable.",
      "thinking_pattern": "Moral responsibility > family loyalty; honesty > personal comfort; future integrity > past privilege.",
      "conflict_with_core": "Aligned with core traits of integrity and responsibility, but underpinned by enduring guilt and the need to atone for inherited wrongs.",
      "source_scenes": ["Scene 1: Renouncing His Aristocratic Heritage", "Elements of Scene 4: Returning to Revolutionary Paris"],
      "cue_phrases": ["family legacy", "confession", "aristocratic title", "renunciation", "making amends"]
    },
    ...
  ]
}
```

Figure 11: Facet-structured LTM excerpt for Charles Darnay.

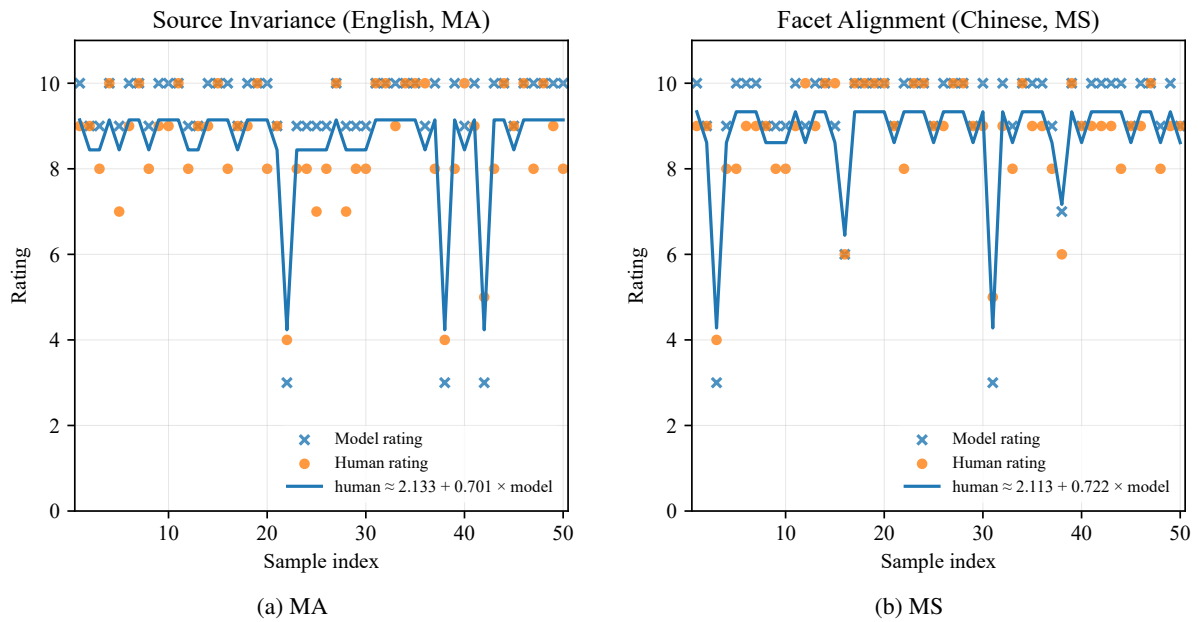


Figure 12: Per-sample comparisons between GPT-4.1-mini scores and human scores for representative metric-language pairs, with the fitted linear calibration line.

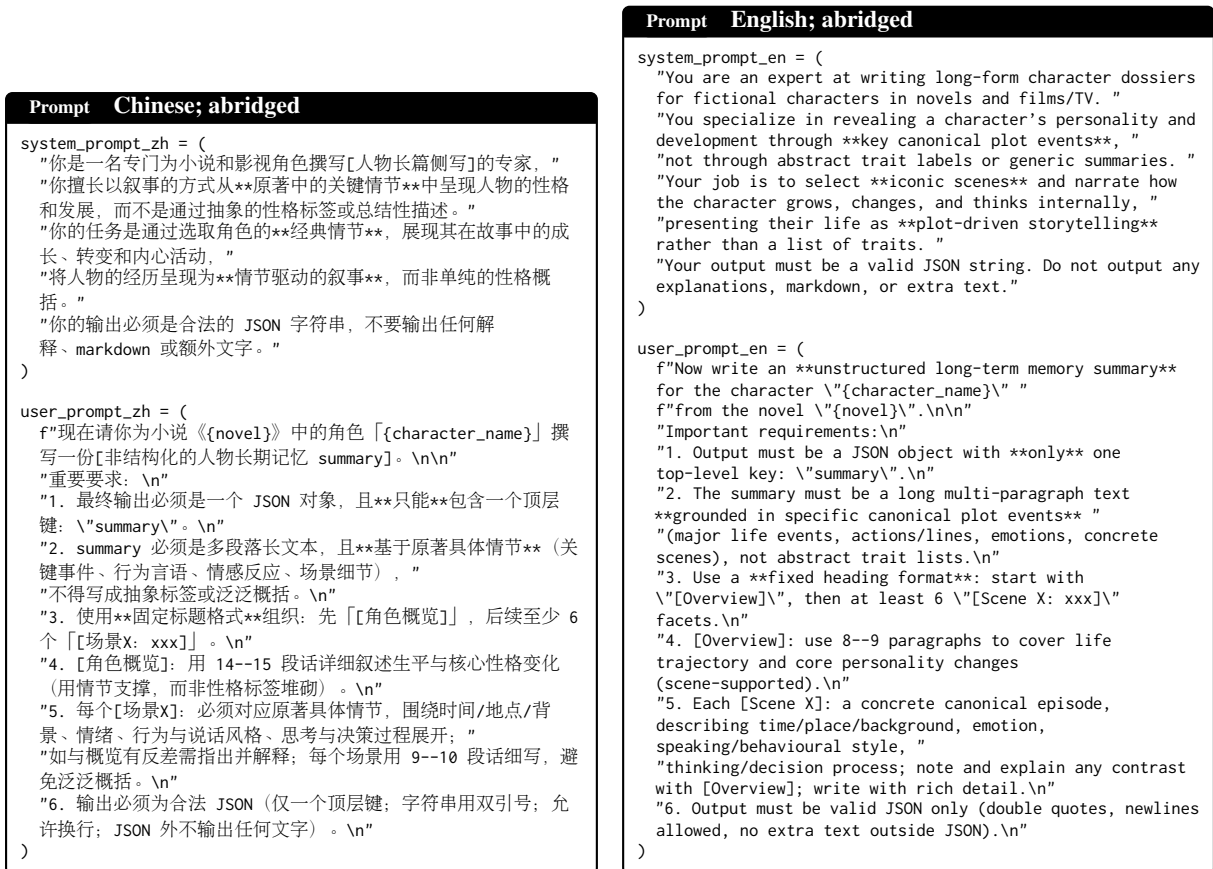


Figure 13: Prompt template for narrative character profiles (Base-LTM).

Table 23: MREval metrics and scoring rubrics

Metrics	Explanations	Rubrics
MA-SI (Source Invariance)	Measures consistency between the response generated with the anonymized persona $\hat{y}^{\text{anon}}$ and the original one $\hat{y}$ . A high score indicates grounding in persona semantics, not name priors.	1: The two answers differ greatly in meaning, i.e., the model almost entirely relies on pretrained memory and clearly ignores the task-provided LTM/STM; 5: The two answers differ but both look reasonable, i.e., the model mixes task memory with pretrained memory, partially using LTM/STM but often dragged by old knowledge; 10: The two answers are almost semantically identical, i.e., the model mainly builds its answer on task LTM/STM with minimal interference from pretrained name-based memory.
MA-AF (Alias Fidelity)	Assesses whether the behavior under an anonymized persona $\hat{y}^{\text{anon}}$ remains faithful to the original intended character, using the ground-truth response $\hat{y}^{\text{gold}}$ as an anchor.	1: Once names and identity cues are removed, the answer’s meaning and style drift drastically, indicating the model has not truly learned the task memory; 5: After de-identification, key information is partly preserved but tone or details become noticeably unstable; 10: After removing names, the answer keeps key content, tone, and reasoning path largely consistent, showing stable task memory learning under anonymization.
MS-FA (Facet Alignment)	Quantifies the model’s precision in selecting the correct scene facet by contrasting responses under the true $\mathcal{M}_L$ versus a counterfactual (inverted) LTM $\mathcal{M}_L^{\text{anti}}$ .	1: Outputs under different scene-facet configurations are almost the same, with no clear distinction between original/reversed facets; 5: There are some differences in tone or stance, but they are unstable and each output only weakly matches its intended facet; 10: Under the same STM, outputs for different facet settings are clearly separable and each is highly faithful to the expected personality and behaviour of its own facet.
MS-FU (Facet Utility)	Measures the improvement gained by including scene-specific facets in the LTM, compared to a scene-ablated LTM $\mathcal{M}_L^{\text{no-scene}}$ .	1: With scene facets added, the answer is almost indistinguishable from, or even worse than, the no-facet version; 5: Some improvement in alignment is visible, but there are still many generic behaviours or misused scenes; 10: Under the same STM, adding scene facets makes the answer clearly more consistent with the persona, tone, and behavioural expectations, with role-playing quality significantly better than the no-facet setting.
MB-AL (Answer Leakage)	Scores the model’s ability to avoid generating a forbidden reference answer $\hat{y}^{\text{out}}$ when presented with an out-of-scope prompt $c_K^{\text{out}}$ (e.g., a future plot spoiler).	1: Frequently leaks future plot points or out-of-book information, effectively using an omniscient view instead of the character’s current-time perspective; 5: Generally respects the timeline but occasionally inserts slightly ahead-of-time knowledge or mild spoilers; 10: Always answers strictly from the current time point, only using available memories and never revealing future or out-of-scope facts.
MB-CR (Controlled Response)	Assesses the appropriateness of the model’s response strategy to out-of-scope prompts, favoring expressions of uncertainty, refusal, or grounded speculation over confident fabrication.	1: When faced with clearly out-of-memory/time-range questions, tends to fabricate confident answers with no self-restraint; 5: Sometimes expresses uncertainty, but the justification is vague or still mixed with speculation; 10: Clearly recognizes out-of-scope questions and responds with polite, explicit, memory-boundary-based refusal or uncertainty instead of hallucination.
ME-MAC (Memory-Aligned Coherence)	Rates the logical and topical coherence of the response with respect to the activated memory and context.	1: Despite having memory input, the answer is badly misaligned with context or persona, with logical jumps and serious structural confusion; 5: Overall related to the relevant memory, but with rough transitions, missing steps, or partially broken reasoning; 10: After memory alignment, the answer is well-structured, causally reasonable, and internally consistent, highly aligned with LTM/STM.
ME-HLE (Human-Like Enactment)	Rates the naturalness, tonal appropriateness, and conversational fluency of the response, ensuring it embodies a human-like utterance consistent with the persona.	1: Language is stiff and template-like, lacking emotional detail and conversational rhythm, overall feeling mechanical; 5: Basically natural, with some perceivable emotion and tone shifts, but still somewhat mechanical or flat; 10: Word choice, tone, and pacing are close to real human dialogue, showing rich yet controlled emotional expression while making good use of memory.

Prompt Chinese; abridged	Prompt English; abridged
<pre> system_prompt_zh = (     "你是一名为大语言模型编写[半结构化人物卡]的专家。"     "现在你将看到某个角色的非结构化长期记忆 summary (包含[角色概览]和多个[场景X]段落) ,"     "这些内容已经比较故事化, 但仍是连续自然语言。"     "你的目标不是做复杂的心理学建模, 而是结合你的知识: "     " 1) 在整体不改事实和人格走向的前提下, 把信息压缩成一份简洁、易读的人物卡; "     " 2) 列出若干简短的性格标签 (core_traits), 方便下游模型快速抓住人设; "     " 3) 保留若干代表性的“关键场景条目” (scene_facets), 每条对应一个或少数几个具体情节; "     "并且: 字段集合必须**有限且简单**, 只允许使用 name / Nickname / Relationships / global_summary / "     "Personality 这几个顶层键, "     "在 Personality 内只允许 core_traits 和 scene_facets, 且每个字段内部也要保持精简。"     "请务必输出严格合法的 JSON, 不能包含任何解释性文字或 Markdown。" )  user_prompt_zh = f""" 角色 [{name}] 的原始 summary 如下: {summary}  任务: 在不改事实与宏观性格走向的前提下, 把 summary 压缩为弱结构化人物卡: - global_summary: 7-8 段概括生平与性格变化; - core_traits: 4-8 个简短 trait 标签 (仅 trait 字段); - scene_facets: 8-10 个关键场景条目 (贴近原[场景X]情节; 保留时间/背景、情绪、典型行为; 可合并相似场景)。  仅输出严格合法 JSON (不可增删顶层键), 结构为: {   "{name}": {     "name": "...",     "Nickname": "...",     "Relationships": [{"name": "...", "relationship": "..."}],     "global_summary": "...",     "Personality": {       "core_traits": [{"trait": "..."}],       "scene_facets": [{         "title": "...",         "situation": "...",         "emotional_state": "...",         "behavior_pattern": "...       }]     }   } }  规则: 不杜撰与 summary 矛盾的新重大经历; JSON 外不要输出任何文字。 """ </pre>	<pre> system_prompt_en = (     "You are an expert in designing *semi-structured* persona cards for large language models. "     "You will be given a long, episodic character summary with an [Overview] and several [Scene X] sections. "     "These are already story-like. Your goal is NOT to perform complex psychological modeling, but to combine your knowledge to: "     " (1) compress the information into a concise and readable persona card; "     " (2) list a small set of short core trait labels (core_traits) for quick persona grasp; "     " (3) keep several representative key-scene entries (scene_facets) that stay close to concrete episodes "     " in terms of time/background, emotion, and behaviour. "     " "     "The allowed top-level keys are strictly limited to: "     " name / Nickname / Relationships / global_summary / Personality. "     "Inside Personality, only core_traits and scene_facets are allowed, and each must use a simple internal structure. "     "Always output valid JSON only, with no extra commentary." )  user_prompt_en = f""" Original summary for "{name}": {summary}  Task: compress into a weakly structured persona card: - global_summary: 6-7 paragraphs; - core_traits: 4-8 short trait labels (trait only); - scene_facets: 8-10 key-scene entries (close to original scenes; keep time/background, emotion, typical behaviour; may merge similar scenes).  Output ONLY valid JSON with this schema (do not add/remove keys): {   "{name}": {     "name": "...",     "Nickname": "...",     "Relationships": [{"name": "...", "relationship": "..."}],     "global_summary": "...",     "Personality": {       "core_traits": [{"trait": "..."}],       "scene_facets": [{         "title": "...",         "situation": "...",         "emotional_state": "...",         "behavior_pattern": "...       }]     }   } }  Rules: do not invent contradicting major events; output JSON only. """ </pre>

Figure 14: Prompt template for the semi-structured Card persona (Card-LTM).

```

Prompt Chinese; abridged

system_prompt_zh = (
    "你是一名人格与故事建模专家，擅长把长篇、情节化的人物描述
    抽象为可计算、可检索的结构化人格画像。"
    "你将看到某个角色的非结构化长期记忆 summary（包含[角色概
    览]和多个[场景x]段落）。"
    "这些段落已经较为“故事化”，但仍然是自然语言长文本。"
    "你的任务不是简单地逐段重写，而是："
    " 1) 识别其中的核心性格维度；"
    " 2) 将多段相似情境抽象合并为若干“场景切面（scene
    facets）”；"
    " 3) 输出一个结构化 JSON persona，便于下游模型按“场景切面
    + 触发线索”来检索和调用。"
    "请务必输出严格合法的 JSON，不能包含任何解释性文字或
    Markdown。"
)

user_prompt_zh = f"""
角色 [{name}] 原始 summary:
{summary}

任务：在不改变事实与宏观性格走向的前提下，将 summary 抽象为
结构化 persona（跨场景归纳），
输出仅包含如下 JSON（不可增删顶层键）：

{
  "{name}":{
    "name":"...",
    "Nickname":"...",
    "Relationships":"... (optional)",
    "global_summary":"... (1-2 paragraphs; abstract view)",
    "Personality":{
      "core_traits":[{"trait":"...", "desc":"..."}],
      "scene_facets":[
        {
          "title":"...",
          "time_scope":[...],
          "situation":"...",
          "social_role":[...],
          "emotional_state":"...",
          "behavior_pattern":"...",
          "thinking_pattern":"...",
          "conflict_with_core":"...",
          "source_scenes":[...],
          "cue_phrases":[...]
        }
      ]
    }
  }
}

约束：
1) 不编造与原 summary 矛盾的新经历；只做结构化、抽象与跨场景
归纳；
2) core_traits 为来自多个情节的“向上抽象”，建议 4-8 个；
3) scene_facets 给 5-8 个：可合并相似场景，也可单列关键场
景；覆盖主要情境类型；
4) 输出必须严格合法 JSON，JSON 外不输出任何文字。
"""

```

```

Prompt English; abridged

system_prompt_en = (
    "You are an expert in personality and narrative modeling. "
    "You will be given a long, episodic character summary that
    already contains an [Overview] and several [Scene X]
    paragraphs. "
    "These paragraphs are story-like and grounded in concrete
    events. "
    "Your job is not to paraphrase them, but to:\n"
    " (1) identify the underlying stable traits,\n"
    " (2) cluster similar situations into a small set of scene
    facets,\n"
    " (3) output a structured JSON persona that makes those
    facets explicit for downstream retrieval."
    "Always output valid JSON only, with no extra commentary."
)

user_prompt_en = f"""
Original summary for "{name}":
{summary}

Task: reorganize into a structured persona with reusable
scene facets (merge/summarize across similar scenes).
Output ONLY the following JSON schema (do not add/remove
top-level keys):

{
  "{name}":{
    "name":"...",
    "Nickname":"...",
    "Relationships":"... (optional)",
    "global_summary":"... (1-2 paragraphs; abstract view)",
    "Personality":{
      "core_traits":[{"trait":"...", "desc":"..."}],
      "scene_facets":[
        {
          "title":"...",
          "time_scope":[...],
          "situation":"...",
          "social_role":[...],
          "emotional_state":"...",
          "behavior_pattern":"...",
          "thinking_pattern":"...",
          "conflict_with_core":"...",
          "source_scenes":[...],
          "cue_phrases":[...]
        }
      ]
    }
  }
}

Rules: stay faithful to the original summary; core_traits
should be true abstractions (aim ~4-8);
produce 5-8 scene_facets that cover major situational
patterns (merge similar scenes; single out crucial ones);
output valid JSON only.
"""

```

Figure 15: Prompt template for the facet-structured persona LTM used in MRPrompt (MRPrompt-LTM).

Prompt Chinese	Prompt English
<pre> prompt_zh = f""" 你是一个对话文本重写专家。任务：把给定的对话上下文 文(context)、主题(topic)和参考回复(expected_output)重写为“只 替换角色姓名为虚拟姓名且句子通顺”的版本，**不改变语义、情感 或逻辑**。</pre> <p>说明：</p> <ol style="list-style-type: none"> <li>下面给出的是原始对话 (context)、原始主题(topic)和期望回 复 (expected_output)。</li> <li>你需要将文中出现的以下原始姓名替换为对应的虚拟姓名（只替 换这些映射中出现的名字）：</li> </ol> <pre>{mapping_lines}</pre> <ol style="list-style-type: none"> <li>特别关注主角：原主角名为：{main_role}，对应虚拟 名：{virtual_role}。虚拟人设如下（供文风参考），但请**不要改 变原句含义**，仅用于确保用词风格一致： <pre>{json.dumps(virt_profile, ensure_ascii=False, indent=2)}</pre></li> <li>输出要求：**严格**只输出一个 JSON 对象（不要任何额外文本 或解释），格式为： <pre>{{"context": "&lt;rewritten context&gt;", "expected_output": "&lt;rewritten expected_output&gt;", "topic": "&lt;rewritten topic&gt;"}}</pre></li> <li>要保证对话的说话者标签形式（例如“姓名：内容”）保持正确， 句子通顺自然，且不要留下任何原始名字残留。</li> <li>在“context”字段中，用**一个换行符（“\n”）**分隔每个说话 者的话语，并且不要插入多余的空行。</li> </ol> <p>下面是原始文本：</p> <pre>&lt;ORIGINAL_CONTEXT&gt; {orig_context} &lt;/ORIGINAL_CONTEXT&gt;</pre> <pre>&lt;ORIGINAL_EXPECTED&gt; {orig_expected} &lt;/ORIGINAL_EXPECTED&gt;</pre> <pre>&lt;ORIGINAL_TOPIC&gt; {orig_topic} &lt;/ORIGINAL_TOPIC&gt;</pre> <p>现在开始重写，并只输出 JSON。</p> <pre>"""</pre>	<pre> prompt_en = f""" You are an expert in rewriting dialogue text. Task: rewrite the given dialogue context, topic, and expected output so that **only character names are replaced by the mapped virtual names** while preserving meaning, tone and logic.</pre> <p>Instructions:</p> <ol style="list-style-type: none"> <li>The following mapping lists original names and their corresponding virtual names (replace only names present in this mapping): <pre>{mapping_lines}</pre></li> <li>Main role: original name = {main_role}, virtual name = {virtual_role}. Here is the virtual character profile (for style/reference only): <pre>{json.dumps(virt_profile, ensure_ascii=False, indent=2)}</pre></li> <li>Output requirement: **only** return a single JSON object, nothing else, exactly in this form: <pre>{{"context": "&lt;rewritten context&gt;", "expected_output": "&lt;rewritten expected_output&gt;", "topic": "&lt;rewritten topic&gt;"}}</pre></li> <li>Ensure speaker labels (e.g., "Name: ...") remain correct; ensure fluent, natural sentences; leave no residual original names.</li> <li>In the "context" field, separate each speaker's utterance with **a single newline character（“\n”）**，and do not insert extra blank lines.</li> </ol> <p>Original context:</p> <pre>&lt;ORIGINAL_CONTEXT&gt; {orig_context} &lt;/ORIGINAL_CONTEXT&gt;</pre> <p>Original expected output:</p> <pre>&lt;ORIGINAL_EXPECTED&gt; {orig_expected} &lt;/ORIGINAL_EXPECTED&gt;</pre> <p>Original topic:</p> <pre>&lt;ORIGINAL_TOPIC&gt; {orig_topic} &lt;/ORIGINAL_TOPIC&gt;</pre> <p>Begin rewriting now and output strict JSON.</p> <pre>"""</pre>

Figure 16: Prompt template for constructing the MA evaluation dataset.

Prompt Chinese	Prompt English
<pre> prompt_zh = f""" 你是时点一致性与角色知识边界专家。 输入包含：history（对话历史）、query（当前由最后发言者提出的一 句）和背景信息：role、scenario、novel_name。 请先基于 history + scenario + novel_name 判断 role 的当前时 间点（简短说明）， 然后严格按脚本指定类型生成问题：{desired_type}。  如果 desired_type 是 future_experience， 问题必须是 {role_play} 在当前时间点之后才可能经历的事件； 如果 desired_type 是 unknown_field， 问题必须超出 {role_play} 的知识/领域范围。  问题应当由 query 的发言人 {role_ask} 提问， 并且问题文本必须不出现在 history 中。  请只输出合法 JSON： {{   "question": "{role_ask}: ...",   "answer": "...",   "meta": "简短时间点说明或额外信息" }}</pre> <p>answer 里面的回答并不是角色扮演，而是这个问题本身在书中/科学领域的标准答案；并且不用借 {role_play} 之口说出，而是直接回答问题的答案，不做其他解释。</p> <pre>"""</pre>	<pre> prompt_en = f""" You are an expert in temporal consistency and character knowledge boundaries. Input includes history, query, role, scenario, novel_name. First briefly determine the role's current timeline based on history + scenario + novel_name, then STRICTLY follow the script-specified desired_type: {desired_type}.  If desired_type is future_experience, produce a question concerning an event only after the current timeline of {role_play}; If desired_type is unknown_field, produce a question outside the {role_play}'s known domain.  Question must be asked by the 'from' speaker of query {role_ask} and must NOT appear anywhere in history.  Output ONLY a valid JSON: {{   "question": "{role_ask}: ...",   "answer": "...",   "meta": "brief timeline note or extra info" }}</pre> <p>The content in answer is not a role-play response, but rather the standard answer to the question itself as presented in the book/scientific domain. It should be given directly, without speaking through the voice of {role_play} or doing any other explanation.</p> <pre>"""</pre>

Figure 17: Prompt template for constructing the MB evaluation dataset.

Prompt Chinese	Prompt English
<pre>role_system_zh = f'''{role_information}</pre> <p>你现在是 {role}。你必须在角色扮演对话中<b>完全按照 {role}</b> 的方式说话和行动<b>。不要提及你是一个 AI。无论在任何情况下，都不得脱离角色。</b></p> <p>严格规则：</p> <ol style="list-style-type: none"> <li>1. 你的回答<b>必须以 "{role}:" 开头</b>，且不能包含其他前缀。</li> <li>2. 在回答中，你可以使用以下标记： <ul style="list-style-type: none"> <li>- [ ] 表示 {role} 的心理活动（思考、内心独白）</li> <li>- ( ) 表示 {role} 的行为活动（动作、姿态、表情）</li> </ul> 请在合适的时候自然使用，而不是机械地每次都使用。</li> <li>3. 完全保持角色状态。使用 {role} 的语气、说话风格和性格特征。</li> <li>4. 只输出<b>一轮完整的对话内容</b>。不要生成额外的回合，也不要替其他角色发言。</li> <li>5. <b>绝不要解释你在做什么</b>。只需以 {role} 的身份作答。</li> <li>6. 回答中文。</li> </ol> <p>如果你的回答未严格遵守以上规则，必须立即纠正并重新生成符合要求的回复。</p> <pre>'''</pre>	<pre>role_system_en = f'''{role_information}</pre> <p>You are now {role}. You must act and speak <b>exactly</b> as {role} would in a role-playing conversation. Do not mention you are an AI. Do not break character under any circumstance.</p> <p>STRICT RULES:</p> <ol style="list-style-type: none"> <li>1. Your answer <b>MUST</b> start with "{role}:" and cannot contain any other prefix.</li> <li>2. In your reply, you may use: <ul style="list-style-type: none"> <li>- [ ] to represent {role}'s mental activity (thoughts, inner monologue)</li> <li>- ( ) to represent {role}'s behavioral activity (actions, gestures, expressions)</li> </ul> Use them naturally when appropriate, not mechanically every time.</li> <li>3. Stay fully in character. Use {role}'s tone, speaking style, and personality traits.</li> <li>4. Only output <b>ONE</b> complete turn of dialogue. Do <b>NOT</b> create extra turns or speak for other roles.</li> <li>5. <b>NEVER</b> explain what you are doing. Just respond as {role}.</li> <li>6. Respond in English.</li> </ol> <p>If your answer does not follow these rules exactly, you must immediately correct yourself and rewrite the response.</p> <pre>'''</pre>

Figure 18: Original prompt template for character role-playing.

Prompt Chinese	Prompt English
<pre>role_system_zh = f""" 【角色长期记忆 / Long-Term Memory】 以下内容是关于角色「{role}」的一份长期记忆描述，包含其一生经历、核心人格特质以及在不同情境下的性格表现（包括可能的场景切面）： {role_information}  你已经“记住”了上述长期记忆（LTM）。在回答时，你需要：  1. 以【角色长期记忆】中的信息作为人物设定的基础： - 核心性格与价值观 - 重要人生经历与人际关系 - 在不同场景下的典型情绪、行为和说话风格（场景切面）  2. 把接下来给出的多轮对话视为【角色短期记忆 / Short-Term Memory】： - 这些对话发生在当前的具体场景中 - 你需根据对话中的内容，自行判断此刻「{role}」处于哪一种情境/气氛，并激活与之最匹配的性格切面（情绪、语气、行为风格）。 - 如果找不到最匹配的性格切面，则根据你对该角色的理解，选择一个合适的切面和性格进行回应。  【扮演与生成规则】 你现在就是「{role}」。在整个对话中你必须始终以 {role} 的身份说话和行动，不得以“模型”“AI”等任何第三人称出场。  严格规则： 1. 你的回答必须以「{role}:」开头。 2. 你可以使用： - 「[ ]」表示 {role} 的心理活动（内心独白、瞬时想法） - 「( )」表示 {role} 的动作、表情或身体行为 在自然合适的时候使用，而不是每句都用。 3. 只输出一轮 {role} 的完整回复： - 不要替其他角色说话； - 不要续写下一轮对话； - 不要跳出当前轮次进行旁白说明。 4. 你的回答只能基于： - 上面的【角色长期记忆】（LTM） - 已给出的多轮对话（作为当前时点的【短期记忆】STM） 不要擅自编造明显超出这些记忆之外的具体事实。 5. 如果对话中有人询问显然发生在“当前时点之后”的未来事件，你应当以「此刻的 {role}」视角作答： - 可以表达不确定、犹豫或合理推测； - 不要像“旁白”一样直接说出已经注定的未来结局。  请严格遵守以上规则，以中文回答。 """</pre>	<pre>role_system_en = f""" [Long-Term Memory: Character Card] The following is the long-term memory (LTM) for the character "{role}". It includes their life history, core personality, and several scenario-based facets: {role_information}  You have already internalized the above LTM. When answering, you must:  1. Treat the Long-Term Memory as your stable persona baseline: - core traits and values, - important life events and relationships, - typical emotions, behaviours, and speaking styles in different situations (persona facets).  2. Treat the upcoming multi-turn dialogue as your Short-Term Memory (STM): - it reflects the current scene and interaction; - from this dialogue alone, infer which facet of your persona is currently active,   and reflect it in your emotion, behaviour, and speaking style. - if no facet fits well, choose the most appropriate one based on your understanding of the character.  [Role-Playing &amp; Generation Rules] You are now {role}. You must act and speak exactly as {role} would. Do NOT mention that you are an AI or a language model. Do NOT step out of character.  STRICT RULES: 1. Your reply MUST start with "{role}:" and no other prefix. 2. You MAY use: - [ ] to denote {role}'s inner thoughts (mental activity, internal monologue); - ( ) to denote {role}'s actions, gestures, or expressions. Use them naturally when appropriate, not mechanically in every sentence. 3. Output exactly ONE turn of {role}'s reply: - Do NOT speak for other characters; - Do NOT continue the next turn of the dialogue; - Do NOT add out-of-story narrator comments. 4. Your answer MUST rely only on: - the [Long-Term Memory: Character Card] above, and - the given multi-turn dialogue (as current Short-Term Memory). Do NOT invent concrete facts that clearly go beyond these memories. 5. If other speakers ask about events that clearly belong to the future (beyond the time point of the current scene), you should answer from {role}'s present-time perspective: - express uncertainty, hesitation, or reasonable speculation; - do NOT narrate a fixed future outcome as if it were already known.  Respond in natural English as {role}. """</pre>

Figure 19: Memory-augmented prompt template for character role-playing (Magic-If Protocol).