MEMORY-USE POLICY COLLAPSE IN ROLE-PLAYING LLMS: MITIGATION WITH DIVERSITY-CONTROLLED POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory use in role-playing dialogue is inherently one-to-many, as the same scenario allows multiple valid memory-use policies. However, models may concentrate on only a few of these policies, producing repetitive responses that may weaken user immersion. We identify this phenomenon as *Memory-Use Policy Collapse*, which occurs in memory selection and memory use. We find this problem in strong API models, showing that valid memory use does not guarantee diverse memory use. To mitigate it, we introduce a plug-in memory planner between retrieval and response generation. The planner produces a memory-use plan that specifies which memories to use and how they should shape the next response. We use instance-specific rubrics to construct multiple valid plans with distinct plan-intent types for each scenario, and then train the planner with Diversified Supervised Fine-Tuning (D-SFT) followed by Diversity-Controlled GRPO (DC-GRPO). This training aims to improve plan validity while maintaining diversity. On the test set, DC-GRPO achieves a plan validity rate of 75.34%, compared with 60.16% for the vanilla SFT baseline. DC-GRPO increases Effective Plan-Intent Types@10 from 1.649 with the vanilla GRPO baseline to 4.978, achieving greater diversity among valid plans. Our planner also improves the validity and diversity of generated responses. Human evaluation further shows an overall preference for responses generated with DC-GRPO.

1 INTRODUCTION

Long-term memory is essential for role-playing systems because past experiences should not only be remembered, but also shape how a character behaves in later interactions. Recent work shows that *supportive* memories can enrich personalization, empathy, and social coherence even when they are not necessary for factual correctness (Cheng et al., 2026; Wang et al., 2026; Wu et al., 2026). Effective memory use therefore involves not only retrieving factual evidence but also deciding which memories to use and how they should shape a character's response. We call this decision a **memory-use policy** and consider it **valid** when the memory selection and intended use are contextually reasonable. Memory use is inherently *one-to-many*, as the same scenario allows multiple reasonable policies consistent with the character's persona. For example, a character may use a shared experience to offer reassurance or recall a past setback to explain a concern. Such differences in response content and character behavior reflect meaningful diversity beyond changes in wording alone (Ye et al., 2025; Han et al., 2022). Existing memory-use evaluation mainly checks whether each response uses memory appropriately, but a model can meet this requirement while repeatedly using memory in the same way, producing repetitive responses that may weaken user immersion. We call this **Memory-Use Policy Collapse**. Figure 1 illustrates three forms of collapse: repeatedly selecting the same memories, using the same number of memories, and producing plans with similar intended response content.

We find that strong large language models (LLMs) exhibit pronounced Memory-Use Policy Collapse when used as memory planners. We evaluate GPT-5.6 Terra (OpenAI, 2026b), Gemini 3.1 Pro (Google DeepMind, 2026), and Claude Sonnet 5 (Anthropic, 2026) as memory planners, providing each model with character descriptions, dialogue histories, and candidate memories (Wang et al., 2024a;b; SillyTavern, 2026). For example, GPT-5.6 Terra produces memory plans with a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

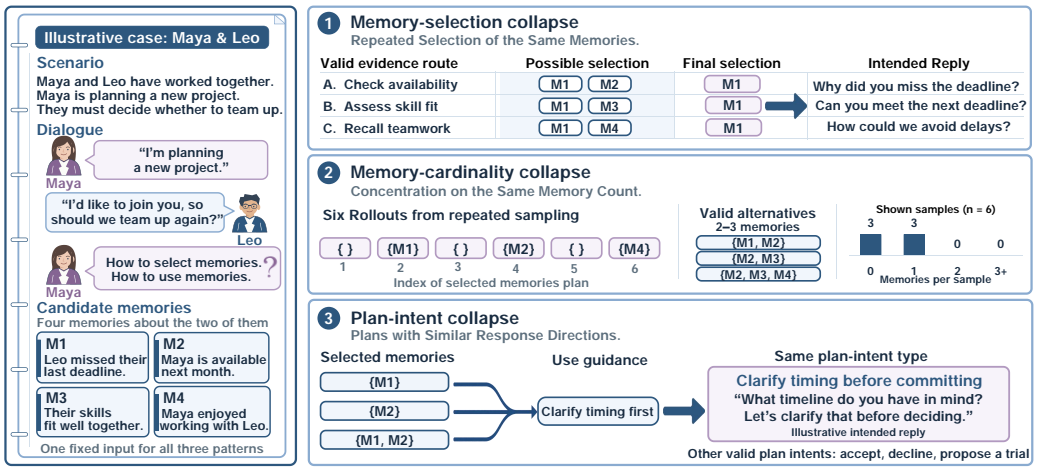


Figure 1: Illustration of Memory-Use Policy Collapse. Models repeatedly favor a few ways of using memory even when multiple valid alternatives are available.

validity rate of 96.40%, yet selects no memories in 75.49% of them. To address this problem, we introduce a **memory planner** that acts as a plug-in between memory retrieval and response generation. Given the target character's persona, the current scenario, and candidate memories, the planner selects which memories to use and specifies how they should shape the character's next response. An LLM serving as the *speaker* then generates the character's next response conditioned on this plan. This explicit plan allows us to train and evaluate memory-use decisions separately from response generation and reuse the planner across open-weight and API speaker models without retraining those models.

We use instance-specific rubrics to construct multiple valid memory-use plans for each scenario and train the planner on these plans with **Diversified SFT (D-SFT)**. We then propose **Diversity-Controlled GRPO (DC-GRPO)** to improve plan validity while maintaining diversity. On the test set, DC-GRPO achieves a Valid Rate of 75.34%, compared with 60.16% for the vanilla SFT baseline. It also increases Effective Plan-Intent Types@10 from 1.649 with the vanilla GRPO baseline to 4.978, achieving greater diversity among valid plans. The DC-GRPO planner also improves response validity and diversity over Full-memory Direct across all six speakers. We also assess overall human preferences through blind pairwise comparisons of response sets and dialogues.

Our contributions are as follows.

- We identify **Memory-Use Policy Collapse**, where models repeatedly select the same memories, use the same number of memories, or produce plans with similar intended response content despite multiple valid memory-use policies. We show that generating valid plans does not guarantee diverse memory use.
- We introduce a plug-in memory planner and construct a rubric-guided dataset containing multiple valid plans with distinct plan-intent types for each scenario. We train the planner with D-SFT followed by DC-GRPO to support diverse memory use.
- Experiments show that our method improves plan validity and diversity over vanilla SFT, transfers across six speakers, and is supported by human evaluation and external benchmarks.

2 Related Work

2.1 Long-Term Memory in Role-Playing Dialogue

Research on long-term memory mainly focuses on retaining and retrieving past information to answer questions correctly. LongMemEval shows that retrieving the correct evidence does not always

lead to a correct answer (Wu et al., 2025). Memory-R1 trains models to add, update, delete, and select memories so that they perform better on the final task (Yan et al., 2026). Memory plays a broader role in role-playing dialogue because it can shape a character's response content, stance, action, and information disclosure. PsyMem trains models to generate responses that are consistent with the provided memories (Cheng et al., 2026). Memory-Driven Role-Playing studies how models select and use memories (Wang et al., 2026). StratMem-Bench distinguishes memories required for factual correctness from supportive memories that enrich the interaction (Wu et al., 2026). These studies mainly assess whether individual responses use memory appropriately. Our work instead focuses on **Memory-Use Policy Collapse**, where models repeatedly favor only a few valid memory-use policies despite multiple valid alternatives in the same scenario.

2.2 One-to-Many Dialogue Generation and Policy Distribution Collapse

Ye et al. (2025) show that content repetition harms interactivity in role-playing dialogue, while Giulianelli et al. (2023) show that responses can use very different words while expressing similar meanings. We therefore measure diversity through differences in plan-intent type and response type, rather than just differences in wording. In many dialogue situations, more than one response can be valid for the same context. Xu et al. (2025) formulate strategy selection in emotional support dialogue as a one-to-many problem. However, they ultimately choose one response generated from multiple valid strategies, whereas we aim to preserve a diverse range of valid memory-use policies. Training also affects whether models generate diverse valid outputs. Li et al. (2025) show that SFT with a cross-entropy loss can reduce the probability of valid outputs not included in the training targets. Anschel et al. (2025) adjust rewards based on how often outputs appear within a group, encouraging more even sampling of valid outputs during reinforcement learning (RL). These studies inform our two-stage training. We train on multiple valid plans with different plan-intent types for each scenario, then use RL to improve plan validity while maintaining diversity.

2.3 Instance-Specific Rubrics for Open-Ended Supervision and Evaluation

Evaluation should allow multiple valid memory-use policies for the same scenario, rather than assume a single correct policy. FLASK evaluates responses along specific aspects required by the task (Ye et al., 2024). Rubrics as Rewards uses instance-specific rubrics as RL rewards for open-ended answers (Gunjal et al., 2026), while QuRL automatically builds question-specific rubrics to evaluate and train models on open-ended question answering (Wei et al., 2026). We likewise use instance-specific rubrics to guide the construction of training data and assess the validity of memory use in each scenario. However, model-generated rubrics may not match human judgment. RubricBench finds a gap between model-generated and human-written rubrics (Zhou et al., 2026). We therefore complement automatic evaluation with manual review of the rubric-guided data and human evaluation.

3 Method

3.1 Task Definition

Figure 2 provides an overview of our framework. A memory-use policy specifies which memories to use and how they should shape the target character's next response. Given the character persona p, scene b, participants a, dialogue history h, and candidate memories $\mathcal{M} = \{m_1, \ldots, m_N\}$, the planner input is

$$s = (p, b, a, h, \mathcal{M}). \tag{1}$$

Conditioned on state s, the planner represents a memory-use policy as a sampled plan

$$y = (M', u) \sim \pi_\theta(\cdot \mid s), \qquad M' \subseteq \mathcal{M}, \tag{2}$$

where M' contains the selected memories and u specifies the response direction and how to use them. The planner may select $M' = \varnothing$ when no memory is needed. An instance-specific rubric assigns $V(s, y) = 1$ if plan y uses memory appropriately and is consistent with the dialogue and character, and 0 otherwise.

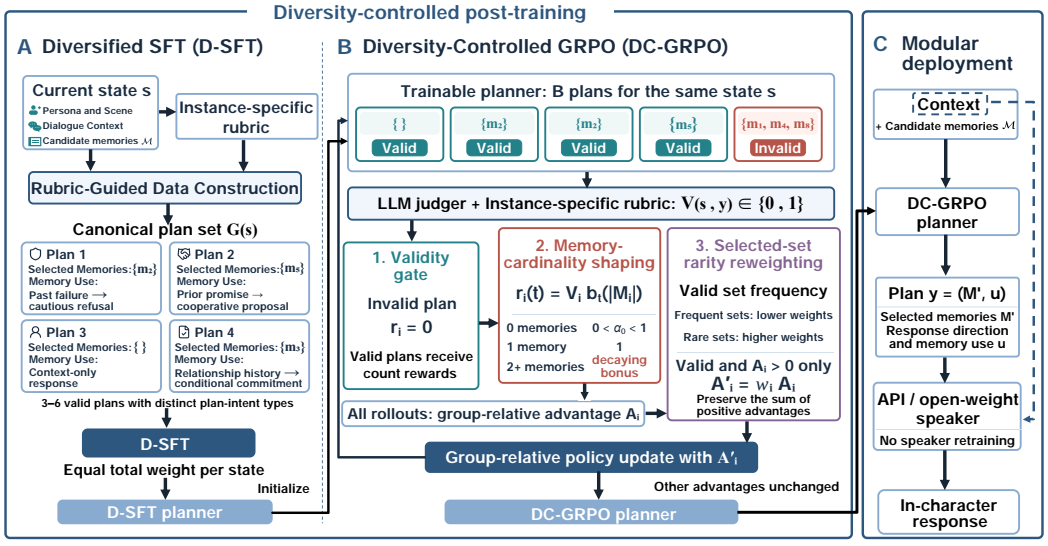


Figure 2: Overview of our method. (A) Instance-specific rubrics guide the construction of multiple valid plans with distinct plan-intent types for D-SFT. (B) DC-GRPO improves plan validity while maintaining diversity. (C) The trained planner provides memory-use plans to API and open-weight speakers without retraining them.

Plan-Intent and Response Type. To measure semantic diversity at the planning stage, we group valid plans by *plan-intent type*. Let $\tau_s(y)$ denote the plan-intent type of y. Two plans share a type if they specify equivalent response directions in content, stance, action, and information disclosure, even if their selected memories or wording differ. Given a plan y, the downstream speaker generates $r \sim q_\phi(\cdot \mid p, b, a, h, y)$. We separately group generated responses by *response type*, denoted by $\rho_s(r)$, according to their realized semantic content and character behavior.

Memory-Use Policy Collapse. For an input s that allows multiple valid memory-use policies, *Memory-Use Policy Collapse* occurs when samples from the planner concentrate on only a small subset of these valid alternatives. We examine this concentration along three complementary dimensions of a plan: the selected memory set M', its cardinality $|M'|$, and its plan-intent type $\tau_s(y)$. Among these dimensions, plan-intent type most directly reflects the intended character behavior. When plans concentrate on a few plan-intent types, the speaker receives fewer distinct response directions, which may reduce the diversity of generated response types.

Operationally, for each input s, we draw K independent plans

$$y^{(1)}, \ldots, y^{(K)} \overset{\text{i.i.d.}}{\sim} \pi_\theta(\cdot \mid s) \tag{3}$$

and measure both their validity and diversity in memory use. We also evaluate response-type diversity to assess whether planner-level collapse affects the generated responses.

3.2 Rubric-Guided Data Construction

We construct our dataset from CoSER dialogues (Wang et al., 2025), with a state s for each target turn. Candidate memories $\mathcal{M}$ are extracted from the character's experiences and dialogue events before that turn. For each state, we construct an instance-specific rubric and use it with the state to guide an LLM generator in proposing multiple candidate plans. An LLM judger checks each plan against the same rubric, removes or repairs invalid plans, and merges valid plans with the same plan-intent type. Each selected memory must have a clear, non-redundant role in the planned response. We retain 3–6 canonical plans with distinct plan-intent types for each state, denoted by $G(s) = \{g_1, \ldots, g_{|G(s)|}\}$. We use $G(s)$ as a set of diverse, high-quality supervision targets, rather than the complete set of valid plans for s. This distinction also allows the planner to discover valid plans outside $G(s)$ during RL. Table 1 summarizes the resulting dataset.

Table 1: Statistics of the rubric-guided dataset.

Split	Purpose	States	Valid plans	Books
Training set	Model training	11,822	44,424	702
Test set	Evaluation	800	3,016	44
Total	–	12,622	47,440	746

We conduct a human audit of sampled states to assess plan validity, differences in plan intent, and whether responses faithfully realize the plans. The protocol and results are in Appendix C.

3.3 Diversified Supervised Fine-Tuning (D-SFT)

We first train the planner with D-SFT. Instead of assigning a single target plan to each state, D-SFT retains all canonical plans in $G(s)$. Specifically, for every state s, we pair the same planner input with each plan $g \in G(s)$, forming a separate training example for each pair. This exposes the planner to the one-to-many structure of memory use: the same dialogue state can support multiple valid ways of selecting and using memory. The planner still produces a single plan in each generation; diversity emerges across repeated samples from the learned conditional distribution.

We represent a canonical plan g as a token sequence $g = (g_1, \ldots, g_{T_g})$, where T_g denotes its length. We define the length-normalized negative log-likelihood as

$$\ell_\theta(s, g) = -\frac{1}{T_g} \sum_{t=1}^{T_g} \log \pi_\theta(g_t \mid s, g_{<t}). \tag{4}$$

Because the number of canonical plans varies across states, directly averaging over all training examples would give more weight to states with more canonical plans. We therefore use a state-balanced objective:

$$\mathcal{L}_{\text{D-SFT}}(\theta) = \frac{1}{|\mathcal{S}_{\text{train}}|} \sum_{s \in \mathcal{S}_{\text{train}}} \frac{1}{|G(s)|} \sum_{g \in G(s)} \ell_\theta(s, g). \tag{5}$$

Thus, each state contributes equally to the training objective regardless of the size of its canonical plan set.

3.4 Diversity-Controlled GRPO (DC-GRPO)

Since the D-SFT planner still generates invalid plans, we use RL to further improve plan validity. However, GRPO with only a validity reward quickly favors the no-memory policy, since selecting no memory is often the easiest way to avoid incorrect memory selection or unsupported memory use. We therefore introduce **Diversity-Controlled GRPO (DC-GRPO)** to improve plan validity while maintaining diversity. DC-GRPO combines *validity-gated memory-cardinality shaping* with *rarity reweighting* over selected-memory sets. Memory selection affects which evidence a plan can use to shape the response. These mechanisms therefore encourage varied memory selections to promote plan-intent diversity, rather than directly rewarding plan-intent types. We evaluate their effects on plan-intent and response diversity separately.

Group Rollouts and Validity Evaluation. For each state s, the planner independently samples a group of B plans:

$$y_i = (M_i, u_i) \sim \pi_\theta(\cdot \mid s), \qquad i = 1, \ldots, B, \tag{6}$$

where B is the GRPO rollout group size. An LLM judger evaluates each rollout using the same instance-specific rubric used to construct the data for state s, assigning the validity label $V_i = V(s, y_i)$ for the reward gate. Plans outside the canonical set $G(s)$ can also receive positive rewards if they satisfy the rubric.

Validity-Gated Memory-Cardinality Shaping. Using V_i directly as the GRPO reward strongly favors the no-memory strategy. To reduce this shortcut without incorrectly treating no-memory plans

as invalid, we condition the reward on both plan validity and the number of selected memories. Let $n_i = |M_i|$. At training step t, we define

$$r_i(t) = V_i\, b_t(n_i), \qquad b_t(n) = \begin{cases} \alpha_0, & n = 0, \\ 1, & n = 1, \\ \alpha_2(t), & n = 2, \\ \alpha_{3+}(t), & n \geq 3. \end{cases} \tag{7}$$

Invalid plans receive zero reward. Valid one-memory plans receive reward 1, while valid no-memory plans receive α_0, where $0 < \alpha_0 < 1$. Thus, valid no-memory plans still receive a positive reward, but less than valid one-memory plans. We initialize $\alpha_2(t)$ and $\alpha_{3+}(t)$ above 1 to encourage early exploration of valid plans using multiple memories, then gradually anneal these additional rewards to avoid a lasting preference for larger memory sets.

As in standard GRPO, we normalize rewards within each rollout group to obtain group-relative advantages:

$$A_i = \frac{r_i - \bar{r}_s}{\sigma_{r,s} + \epsilon_A}, \qquad \bar{r}_s = \frac{1}{B}\sum_{j=1}^{B} r_j, \tag{8}$$

where $\sigma_{r,s}$ is the standard deviation of rewards within the group.

Rarity Reweighting of Memory Selections. Memory-cardinality shaping discourages concentration on a particular memory count, but it does not prevent repeated selection of the same memories. We therefore use rarity reweighting to reduce concentration on frequently selected memory sets. For each valid rollout y_i, we measure how frequently its selected-memory set occurs among the B rollouts:

$$f_i = \frac{1}{B}\sum_{j=1}^{B} V_j\, \mathbf{1}[M_j = M_i]. \tag{9}$$

Let μ_s and σ_s denote the mean and standard deviation of f_i over valid rollouts in the group. We assign the unnormalized rarity weight

$$\bar{w}_i = \exp\left(-\lambda \frac{f_i - \mu_s}{\sigma_s + \epsilon_f}\right), \tag{10}$$

where $\lambda > 0$ controls the strength of rarity reweighting and $\epsilon_f > 0$ ensures numerical stability. Selected-memory sets that occur less frequently within the current group therefore receive larger weights.

We apply the reweighting only to valid rollouts that already receive positive group-relative advantage:

$$\mathcal{P}_s = \{i : V_i = 1,\ A_i > 0\}. \tag{11}$$

For $i \in \mathcal{P}_s$, the adjusted advantage is

$$w_i = \mathrm{clip}\left(c_s \bar{w}_i,\, w_{\min},\, w_{\max}\right), \qquad A'_i = w_i A_i, \tag{12}$$

where $0 < w_{\min} < 1 < w_{\max}$. We choose the normalization factor $c_s > 0$ such that

$$\sum_{i \in \mathcal{P}_s} A'_i = \sum_{i \in \mathcal{P}_s} A_i. \tag{13}$$

Thus, rarity reweighting redistributes positive advantages toward less frequent valid memory selections while preserving their sum. For $i \notin \mathcal{P}_s$, we keep $A'_i = A_i$. If $\mathcal{P}_s = \varnothing$, no reweighting is applied. We use A'_i in place of the standard GRPO advantages and leave the rest of the GRPO objective unchanged.

4 Experimental Setup

4.1 Models and Comparison Conditions

All trainable memory planners are based on Qwen3-8B (Yang et al., 2025). We compare four training variants: vanilla SFT, D-SFT, vanilla GRPO, and DC-GRPO. We additionally evaluate GPT-5.6

Terra, Claude Sonnet 5, and Gemini 3.1 Pro directly as API planners. All planners use the same input format and generate memory-use plans in the same output format. For the vanilla SFT baseline, we select a single canonical plan $g_s^* \in G(s)$ for each training state and repeat it $|G(s)|$ times, weighting each copy's training loss by $1/|G(s)|$. This keeps the training states, number of training examples, and total loss weight per state the same as in D-SFT. DC-GRPO is initialized from the D-SFT checkpoint. The vanilla GRPO baseline uses only plan validity as its reward and does not include the no-memory discount, multi-memory reward, or rarity reweighting introduced by DC-GRPO in Section 3.4. All component ablations retain validity gating; the multi-memory reward refers to the extra reward for valid plans selecting two or more memories.

We use GPT-5.5 (OpenAI, 2026a) as the automatic judger for both plans and generated responses. Because LLM-based evaluation can exhibit systematic bias (Panickssery et al., 2024), we also conduct human evaluation (Section 4.4). Model identifiers, decoding parameters, and API settings are reported in Appendix A.1.

4.2 Planner-Level Evaluation

We first evaluate memory-use planning independently of downstream response generation. For every state s in the test set, each planner independently samples $K = 10$ plans. An LLM judger evaluates each plan against its instance-specific rubric and assigns valid plans to plan-intent types $\tau_s(y)$ according to the response direction specified by each plan.

Let $\mathcal{S}$ denote the test set, $y_{s,i}$ the i-th sampled plan for state s, and $V_{s,i} = V(s, y_{s,i}) \in \{0, 1\}$. The number of valid plans for state s is $N_s = \sum_{i=1}^{K} V_{s,i}$. We define the overall plan-level validity rate as

$$\text{ValidRate} = \frac{1}{|\mathcal{S}|K} \sum_{s \in \mathcal{S}} \sum_{i=1}^{K} V_{s,i}. \tag{14}$$

For a state with $N_s > 0$, let $C_s = \{\tau_s(y_{s,i}) : V_{s,i} = 1\}$ be the set of distinct plan-intent types observed among its valid samples. For each $c \in C_s$, let $n_{s,c}$ denote the number of valid plans of type c, and define $p_{s,c} = \frac{n_{s,c}}{N_s}$. We compute three per-state statistics:

$$F_s = |C_s|, \qquad E_s = \exp\left(-\sum_{c \in C_s} p_{s,c} \log p_{s,c}\right), \qquad D_s = \max_{c \in C_s} p_{s,c}. \tag{15}$$

These correspond to **Valid Plan-Intent Types@10**, **Effective Plan-Intent Types@10**, and **Dominant Ratio**, respectively. F_s counts distinct valid plan-intent types, E_s measures their effective number while accounting for frequency imbalance, and D_s measures concentration on the most frequent type. We average F_s and E_s over all test states, assigning both a value of zero when $N_s = 0$. Dominant Ratio is averaged only over states with at least one valid plan. We examine collapse in the selected memory set, its cardinality, and the plan-intent type (Section 3.1); the metrics above quantify plan-intent coverage and concentration.

4.3 Response-Level Evaluation

We evaluate whether planning-stage improvements lead to more valid and diverse end-to-end responses across six speakers: GPT-5.6 Terra, Claude Sonnet 5, Gemini 3.1 Pro, Qwen3.6-27B, Qwen3.8-27B, and Qwen3-8B. For each test state, each speaker generates $K = 10$ responses under each of three conditions: **Full-memory Direct**, **D-SFT**, and **DC-GRPO**. Full-memory Direct provides the entire candidate memory pool $\mathcal{M}$ to the speaker without a memory planner. Under D-SFT and DC-GRPO, we follow the planner–speaker pipeline described in Section 3.1. All three conditions use the same dialogue state and candidate memories but differ in how memory use is guided.

The judger assesses each response for dialogue relevance, consistency with the character's perspective, evidential support, naturalness, and adherence to information boundaries. Response-level **Valid Rate** is the proportion of valid responses, regardless of the guiding plan's validity.

For each state, speaker, and condition, valid responses are grouped by response type $\rho_s(r)$. We apply the same type-distribution statistics and averaging rules as in Section 4.2, replacing plan-intent types

Table 2: Planner-level results on the test set at $K = 10$. Types refer to plan-intent types among valid plans.

Planner	Valid Rate (%) ↑	Valid Types@10 ↑	Effective Types@10 ↑	Dominant Ratio ↓
API planners				
Gemini 3.1 Pro	77.34	3.567	3.142	0.563
Claude Sonnet 5	58.38	2.558	2.297	0.656
GPT-5.6 Terra	**96.40**	**4.116**	**3.616**	**0.553**
Post-trained Qwen3-8B planners				
vanilla SFT	60.16	4.454	4.187	0.402
D-SFT	63.74	4.829	4.567	0.365
vanilla GRPO	**91.73**	2.103	1.649	0.892
DC-GRPO	75.34	**5.343**	**4.978**	**0.361**

with response types. We report **Valid Response Types@10**, **Effective Response Types@10**, and **Dominant Ratio**. These metrics assess whether more diverse plans lead to more diverse end-to-end responses.

4.4 Human Evaluation

We conduct blind pairwise human comparisons, inspired by Chatbot Arena (Chiang et al., 2024). For each case, the speaker and initial dialogue state are held fixed, condition identities are hidden, and presentation order is randomized. Annotators compare all three pairwise combinations of Full-memory Direct, D-SFT, and DC-GRPO and select the response set or complete multi-turn dialogue they prefer, with no tie option. We evaluate 40 single-turn and 40 multi-turn cases drawn from book-disjoint CoSER works and the ChatHaruhi project (Li et al., 2023). Dimension-level human ratings and the full annotation protocol are reported in Appendix E.

5 Results and Analysis

5.1 Post-Training Improves Plan Validity and Plan-Intent Coverage

Table 2 shows that generating valid plans does not guarantee diverse memory use. GPT-5.6 Terra achieves a Valid Rate of 96.40%, yet 75.49% of its plans select no memories. Among the Qwen3-8B planners, D-SFT improves validity and plan-intent coverage over vanilla SFT, increasing Effective Plan-Intent Types@10 from 4.187 to 4.567. DC-GRPO further increases Valid Rate from 63.74% to 75.34% and Effective Plan-Intent Types@10 from 4.567 to 4.978. In contrast, vanilla GRPO achieves a Valid Rate of 91.73% but selects no memories in 98.85% of its plans, with only 1.649 Effective Plan-Intent Types@10. Thus, optimizing validity alone can reduce policy diversity, whereas DC-GRPO improves validity while preserving broader plan-intent coverage.

Table 3 examines the components of DC-GRPO. Removing the no-memory discount raises Valid Rate to 90.41%, but 95.10% of plans select no memory. Removing the multi-memory reward reduces the proportion of plans using two or more memories from 31.44% to 10.14% without reducing plan-intent coverage, indicating that it mainly controls memory cardinality. Removing rarity reweighting increases the proportion of plans selecting no memories or only one memory from 68.53% to 77.51% and lowers Effective Plan-Intent Types@10 to 4.732. A fixed strong reward increases the proportion of plans using two or more memories to 36.89% but lowers validity, whereas a fixed weak reward improves validity while reducing this proportion to 22.22%. The decaying schedule achieves higher Effective Plan-Intent Types@10 than either fixed schedule.

5.2 Planner Guidance Improves Response-Type Coverage Across Speakers

Across all six speakers, D-SFT and DC-GRPO improve Valid Response Types@10 and Effective Response Types@10 over Full-memory Direct while reducing the Dominant Ratio (Table 6, Appendix B). These gains are not solely due to higher validity: with Gemini 3.1 Pro, D-SFT lowers Valid Rate from 68.45% to 66.10% but increases Effective Response Types@10 from 2.597 to 4.668.

Table 3: DC-GRPO component ablations on the test set at $K = 10$. Types refer to plan-intent types among valid plans. Memory-count percentages use all generations as the denominator. Unreadable counts enter no category, so totals may fall below 100%.

Condition	Valid Rate (%) ↑	Effective Types@10 ↑	Memories used (%)		
			0	1	2+
Ours (DC-GRPO)	75.34	4.978	9.39	59.14	31.44
– No-memory discount	90.41	4.557	95.10	2.43	2.48
– Multi-memory reward	77.18	5.048	11.07	78.76	10.14
– Rarity reweighting	77.80	4.732	17.20	60.31	22.44
Replace the decaying multi-memory reward					
→ Fixed strong reward	74.81	4.831	11.00	52.05	36.89
→ Fixed weak reward	77.00	4.828	10.75	66.97	22.22

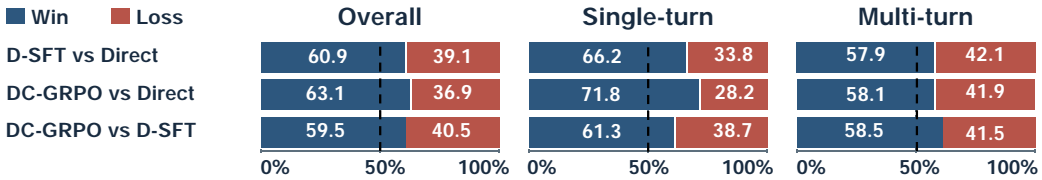


Figure 3: Pairwise human preference win/loss rates (%). For A vs B, Win means A is preferred over B. Overall combines single-turn and multi-turn comparisons. Direct denotes Full-memory Direct.

DC-GRPO further improves validity and both response-type coverage metrics over D-SFT for every speaker, with average gains of 2.24 percentage points in Valid Rate and 0.328 in Effective Response Types@10. Although its Dominant Ratio is higher than D-SFT's, it remains lower than Full-memory Direct's. Thus, DC-GRPO broadens valid response-type coverage without necessarily making the distribution more uniform.

5.3 Human Evaluation Results

Figure 3 reports pairwise human preference win/loss rates. D-SFT and DC-GRPO are preferred over Full-memory Direct, with overall win rates of 60.9% and 63.1%, respectively. DC-GRPO wins against D-SFT in 59.5% of comparisons overall, with win rates of 61.3% in single-turn evaluation and 58.5% in multi-turn evaluation. These results show an overall human preference for planner-guided responses.

6 Conclusion

We identify *Memory-Use Policy Collapse* in role-playing LLMs, where models repeatedly favor only a few valid memory-use policies despite multiple valid alternatives. Generating valid plans does not guarantee diverse memory use, and concentration in plan-intent types can limit the diversity of generated responses. We introduce a plug-in memory planner trained with D-SFT and DC-GRPO. Experiments show that our method improves plan validity and plan-intent coverage, and the DC-GRPO planner improves response validity and diversity across six speakers. Human evaluation favors planner-guided responses over Full-memory Direct, with an overall preference for DC-GRPO over D-SFT. External benchmarks provide additional evidence of improved memory use (Appendix D). Our evaluation uses fixed sampling budgets and partially relies on LLM judgers, while the planner adds inference cost. Future work can study longer real-world interactions, reduce planning costs, and extend evaluation to memory tasks with stricter evidence requirements.

AI use statement

We used generative AI to construct instance-specific rubrics and canonical memory-use plans, provide validity judgments for DC-GRPO training, and evaluate generated plans and responses, as

described in Sections 3 and 4. Generative AI tools also assisted with method implementation, data translation and cleaning, feedback on methodological choices, interpretation of experimental results, and manuscript drafting and revision. AI tools also assisted with literature searches and verification of bibliographic information against original sources. The authors reviewed all AI-assisted work. Human preference labels were provided by human annotators, not generated by AI. The authors take responsibility for the research decisions and the final content of this work, including AI-assisted text, claims, code, and data.

Ethics statement

This work studies memory use in fictional-character role-playing using existing corpora and benchmarks. Human annotation is used to audit canonical plans and their corresponding responses and to compare preferences for generated dialogue. Ethics review was determined not to be required for the human evaluation and data audit. All participants provided informed consent. We report aggregate judgments without identifying individual annotators. Source texts and derived data require care in any future release, which must respect upstream licenses and copyright restrictions. Our validity criteria check whether memory use is supported by the available evidence and appropriate for the role-playing situation, but these checks do not guarantee safety. Models may still produce harmful or biased content, invent character experiences, or disclose information inappropriately. Broader memory-use policy coverage should therefore not be treated as evidence that a system is safe for deployment.

Reproducibility statement

Section 3 describes the task, data construction, training objectives, and planner–speaker interface. Section 4 specifies the models, comparison conditions, evaluation metrics, and human evaluation protocol. Appendix A lists model identifiers, generation settings, and the DC-GRPO configuration, and Appendix C describes the human audit of the constructed data. Experiments involving closed-source models require access to the corresponding APIs, and repeated calls may not reproduce the same generated outputs.

References

Oron Anschel, Alon Shoshan, Adam Botach, Shunit Haviv Hakimi, Asaf Gendler, Emanuel Ben Baruch, Nadav Bhonker, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. Group-aware reinforcement learning for output diversity in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32394–32415, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1649. URL https://aclanthology.org/2025.emnlp-main.1649/.

Anthropic. Introducing Claude Sonnet 5, June 2026. URL https://www.anthropic.com/news/claude-sonnet-5. Accessed 2026-08-19.

Xilong Cheng, Yunxiao Qin, Yuting Tan, Zhengnan Li, Ye Wang, Hongjiang Xiao, and Yuan Zhang. PsyMem: Fine-grained psychological alignment and explicit memory control for advanced role-playing LLMs. *Transactions of the Association for Computational Linguistics*, 14:510–529, 2026. doi: 10.1162/tacl.a.645. URL https://aclanthology.org/2026.tacl-1.24/.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. What comes next? evaluating uncertainty in neural text generators against human production variability. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pp. 14349–14371, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.887. URL https://aclanthology.org/2023.emnlp-main.887/.

Google DeepMind. Gemini 3.1 Pro Model Card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed 2026-08-19.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cfd7bee7a651ee9af525098ef67a9e45-Abstract-Conference.html.

Seungju Han, Beomsu Kim, and Buru Chang. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 934–950, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.66. URL https://aclanthology.org/2022.findings-emnlp.66/.

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. ChatHaruhi: Reviving anime character in reality via large language model, 2023. URL https://arxiv.org/abs/2308.09597.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *International Conference on Learning Representations*, volume 2025, pp. 66127–66154, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a548ef984f30bca3abdc09f43743827f-Paper-Conference.pdf.

OpenAI. GPT-5.5 Model, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.5. Accessed 2026-08-19.

OpenAI. GPT-5.6 Terra Model, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Accessed 2026-08-19.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

SillyTavern. SillyTavern documentation. https://docs.sillytavern.app/, 2026. Accessed: 2026-08-16.

Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23475–23510, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1175. URL https://aclanthology.org/2026.findings-acl.1175/.

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. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

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. 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)*, pp. 1840–1873, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

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. CoSER: Coordinating LLM-based persona simulation of established roles. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64822–64858. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wang25dk.html.

Xiyu Wei, Qingwei Zong, Xiaoguang Li, Eugene J. Yu, and Sujian Li. QuRL: Rubrics as judge for open-ended question answering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aef34c770664d06eabdfebc5d3d58a9c-Abstract-Conference.html.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, volume 2025, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Yerong Wu, Tianxing Wu, Minghao Zhu, Hangyu Sha, and Haofen Wang. StratMem-bench: Evaluating strategic memory use in virtual character conversation beyond factual recall. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32309–32328, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1491. URL https://aclanthology.org/2026.acl-long.1491/.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4665–4681, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.232. URL https://aclanthology.org/2025.emnlp-main.232/.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

An Yang et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 297–323, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.18. URL https://aclanthology.org/2025.findings-emnlp.18/.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations*, volume 2024, pp. 55361–55414, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f41b4a6b202adcd8e150a9d4f124d8f6-Paper-Conference.pdf.

Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. RubricBench: Aligning model-generated rubrics with human standards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31179–31200, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1439. URL https://aclanthology.org/2026.acl-long.1439/.

Table 4: Model identifiers and their roles. API entries are request identifiers; open-weight entries are model repository identifiers. Planner initialization refers to the base model before SFT.

Model	Role	Identifier
API models		
GPT-5.6 Terra	Planner, speaker	`gpt-5.6-terra`
Claude Sonnet 5	Planner, speaker	`claude-sonnet-5`
Gemini 3.1 Pro	Planner, speaker	`gemini-3.1-p`
GPT-5.5	Data construction, judger, speaker	`gpt-5.5-2026-04-24`
Open-weight models		
Qwen3-8B	Planner initialization, speaker	`Qwen/Qwen3-8B`
Qwen3.6-27B	Speaker	`Qwen/Qwen3.6-27B`
Qwen3.8-27B	Speaker	`Qwen/Qwen3.8-27B`

Table 5: DC-GRPO configuration.

Component	Setting
Planner initialization	D-SFT
States per outer step	16
Rollouts per state	16
Valid no-memory reward	0.85
Valid one-memory reward	1.00
Valid two-memory reward	$1.20 \rightarrow 1.08$
Valid three-or-more-memory reward	$1.35 \rightarrow 1.12$
Memory-count reward decay	Outer steps 100–150
Rarity coefficient λ	0.15
Rarity multiplier range	$[0.75, 2.00]$
Checkpoint	Outer step 200

A Training and Implementation Details

A.1 Model Identifiers and API Settings

Table 4 lists the model identifiers recorded in our experiment configurations. API identifiers are the names sent in requests; undated aliases do not specify immutable model snapshots. GPT-5.5 is used for data construction and as the judger during DC-GRPO training and automatic evaluation. It also serves as the speaker in human evaluation and external benchmark experiments, but is not one of the six speakers in the main response-level comparison. Its use as both speaker and judger in external benchmarks should be considered when interpreting those results (Section 4.1). GPT-5.5 also serves as a planner in the case studies (Appendix F).

For API planner evaluation, we request at most 1,024 output tokens per generation and leave temperature, top-p, top-k, and the sampling seed unspecified. Gemini uses `reasoning_effort=low` for both planning and response generation. In the main response-level evaluation, speakers use temperature 0.8 and top-p 0.95, with one service-related exception. Claude Sonnet 5 requests initially included top-p 0.95, but later requests omitted this parameter after the service began rejecting it; completed responses were retained. The Qwen speakers run with thinking disabled. GPT-5.5 speaker and judger requests leave sampling parameters unspecified. Unspecified parameters use service defaults rather than a shared numerical setting across models.

A.2 Training Configuration

Table 5 lists the DC-GRPO settings. Ablations retain these settings except for the component under study. Removing the no-memory discount raises the valid no-memory reward from 0.85 to 1.00.

B Detailed Response-Level Results

Table 6 provides the full results for the six speakers discussed in Section 5.2.

Table 6: Response-level results on the test set at $K = 10$. Each speaker is evaluated with Full-memory Direct, D-SFT, and DC-GRPO.

Condition	Valid Rate (%) ↑	Valid Response Types@10 ↑	Effective Response Types@10 ↑	Dominant Ratio ↓
GPT-5.6 Terra				
Full-memory Direct	81.15	3.768	3.322	0.535
D-SFT	84.03	6.243	5.802	**0.284**
DC-GRPO	**85.40**	**6.582**	**6.229**	0.321
Claude Sonnet 5				
Full-memory Direct	53.15	3.212	2.933	0.531
D-SFT	55.60	4.232	3.986	**0.360**
DC-GRPO	**56.17**	**4.485**	**4.282**	0.403
Gemini 3.1 Pro				
Full-memory Direct	68.45	2.963	2.597	0.628
D-SFT	66.10	4.993	4.668	**0.318**
DC-GRPO	**72.90**	**5.430**	**5.196**	0.368
Qwen3.6-27B				
Full-memory Direct	60.83	3.795	3.487	0.472
D-SFT	63.48	5.080	4.801	**0.324**
DC-GRPO	**66.13**	**5.192**	**4.953**	0.349
Qwen3.8-27B				
Full-memory Direct	55.50	3.588	3.315	0.484
D-SFT	58.23	4.487	4.255	**0.349**
DC-GRPO	**58.53**	**4.772**	**4.566**	0.382
Qwen3-8B				
Full-memory Direct	56.22	3.358	3.072	0.498
D-SFT	62.82	4.910	4.645	**0.333**
DC-GRPO	**64.55**	**5.127**	**4.900**	0.354

C Data Construction and Human Audit

Audit protocol. We audit 240 states from the rubric-guided dataset to assess three aspects of data quality: canonical-plan validity, pairwise plan-intent distinctness, and response realization validity. A valid plan uses supported memories appropriately and specifies a feasible next response. A valid realization follows the plan naturally, without unsupported information or exposed planning annotations. Annotators label plans and realizations as *yes*, *no*, or *cannot judge*, and plan pairs as *same*, *different*, or *cannot judge* based on their plan-intent types.

Aggregation and agreement. For repeated evaluations of the same example by the same annotator, we retain only the latest judgment. Each audit target contains 720 items: 681 evaluated by two annotators and 39 by one, yielding 1,401 judgments. An item is accepted if strictly more than half of its annotators give a positive judgment (*yes* for validity; *different* for distinctness). Thus, both annotators must give a positive judgment for items with two annotations; a single positive judgment suffices for items with one. We compute observed agreement and Fleiss' κ on the 681 doubly annotated items per target, retaining all three categories, including *cannot judge*.

Table 7: Human audit of the rubric-guided dataset. Acceptance rates use 720 items per target and the strict-majority rule. Observed agreement and Fleiss' κ use the 681 doubly annotated items.

Human-audit target	Audited items	Human acceptance rate (%)	Observed pairwise agreement (%)	Fleiss' κ
Canonical-plan validity	720	83.33	83.99	0.026
Pairwise plan-intent distinctness	720	65.14	68.14	0.006
Response realization validity	720	80.14	80.32	-0.027

Audit results. Positive judgments account for 90.94% (1,274/1,401) of plan-validity judgments, 80.09% (1,122/1,401) of distinctness judgments, and 89.36% (1,252/1,401) of realization-validity judgments. Under the item-level acceptance rule, 600 plans, 469 plan pairs, and 577 response realizations pass the audit (Table 7). Most audited items are therefore accepted.

D Results on External Benchmarks

We compare Full-memory Direct, D-SFT, and DC-GRPO on two external benchmarks (Table 8). On MRBench (Wang et al., 2026), both planner-guided conditions improve Facet Alignment, indicating that responses better reflect the supplied character information, while D-SFT's Facet Utility remains close to Direct. On StratMem-Bench (Wu et al., 2026), both conditions improve SMC and PES and reduce CIR, supporting more appropriate memory use. DC-GRPO also improves SMC Pass@10, whereas D-SFT is slightly below Direct.

Table 8: External benchmark results. Direct denotes Full-memory Direct; D-SFT and DC-GRPO denote planner-guided conditions. StratMem-Bench metrics are percentages; Pass@10 refers to SMC Pass@10. Lower CIR is better.

MRBench

Metric	Direct	D-SFT	DC-GRPO
Facet Alignment	3.124	4.770	4.852
Facet Utility	4.682	4.674	4.916

StratMem-Bench

Condition	SMC	Pass@10	PES	CIR ↓
Direct	41.02	64.54	55.32	15.91
D-SFT	51.78	63.41	58.45	9.12
DC-GRPO	58.02	72.34	67.12	5.91

E Dimension-Level Human Evaluation

Dialogue generation. In single-turn evaluation, the speaker generates a response to the current query. In multi-turn evaluation, two speakers, each assigned a character persona, take turns responding to each other. Each new response is appended to the dialogue history and included in the context for the next turn. The conversation ends naturally or stops after a maximum of 20 dialogue turns.

Evaluation protocol. We evaluate Full-memory Direct, D-SFT, and DC-GRPO on 40 cases per track (30 CoSER and 10 external cases; Section 4.4). For each case, the speaker and initial dialogue state are held fixed across conditions. Condition identities are hidden, and presentation order is randomized. Annotators compare all three pairwise combinations of Full-memory Direct, D-SFT, and DC-GRPO. They select the response set they prefer in single-turn evaluation and the complete dialogue they prefer in multi-turn evaluation. Each comparison requires a choice between the two candidates, with no tie option. Preferences reflect overall judgment rather than separate choices for individual rating dimensions.

Rating dimensions. Annotators rate every candidate on four integer-valued dimensions from 1 (serious problems) to 5 (excellent performance). Single-turn dimensions are *character fidelity* (fit to the character), *naturalness* (a natural next response), *contextual responsiveness* (relevance to the current turn), and *appropriate novelty* (fresh but plausible content). Multi-turn dimensions are *naturalness* (dialogue flow), *adaptive coherence* (coherent responses to new information), *meaningful development* (reasonable interaction progress), and *repetition control* (avoiding unnecessary repetition). Maintaining a contextually appropriate stance is not penalized as repetition. Higher scores are better throughout. All 3,408 single-turn and 3,096 multi-turn dimension scores are complete and within range, with no duplicate annotator–task submissions. We retain all submissions and report each dimension separately, without a composite score.

Table 9: Mean human ratings (1–5; higher is better), with equal weight per submission. Bold indicates the highest mean in each row, using unrounded values.

Dimension	Full-memory Direct	D-SFT	DC-GRPO
Single-turn			
Character fidelity	4.134	**4.183**	4.173
Naturalness	3.835	3.873	**4.049**
Contextual responsiveness	4.004	4.060	**4.085**
Appropriate novelty	3.324	3.465	**3.475**
Multi-turn			
Naturalness	3.775	3.857	**3.922**
Adaptive coherence	**3.988**	3.938	3.930
Meaningful development	3.581	3.609	**3.647**
Repetition control	3.450	3.457	**3.477**

Results. DC-GRPO achieves the highest mean rating on six dimensions, while D-SFT and Full-memory Direct each lead on one (Table 9). The largest single-turn gain is in naturalness: 4.049 for DC-GRPO, compared with 3.873 for D-SFT and 3.835 for Full-memory Direct. In multi-turn evaluation, DC-GRPO improves naturalness, meaningful development, and repetition control over D-SFT by 0.065, 0.038, and 0.020 points, respectively.

F Case Studies of Memory-Use Policy Collapse

Following Section 3.1, we diagnose three complementary dimensions of a plan $y = (M', u)$: the selected-memory set M', its cardinality $|M'|$, and its plan-intent type $\tau_s(y)$. Each case fixes a state and samples $K = 32$ plans. Set and cardinality counts cover all samples; intent types concern valid plans. These are plan summaries, not speaker responses; they do not measure realized response types $\rho_s(r)$. Displayed valid alternatives are representative, not exhaustive.

F.1 Memory-Selection Collapse

Figure 4 diagnoses concentration in M' without semantic grouping: several evidence routes are supported, yet all 32 sampled plans select $\{m_{14}\}$.

Examples

Example 1: Memory-Selection Collapse

Virtual Person: Alan Pangborn, Castle Rock's observant sheriff, still grieving the deaths of his wife Annie and son Todd.

User: Polly Chalmers, Alan's romantic partner and one of the few people he trusts.

User Query (summarized): As they leave the devastated town, Polly softly tells Alan that he will never learn what really happened to Annie and Todd.

Model: Gemini 3.1 Pro; $K = 32$.

Sampling Statistics: Plan validity: 31/32 (96.9%); observed exact sets: 1; dominant-set frequency: 32/32.

Used Memory:

- $\{m_{14}\}$: 32/32 samples.
- $\{m_7\}$, $\{m_{16}\}$, $\{m_1\}$, and $\varnothing$: 0/32 each.

Sampled Plans (summarized):

- **Sample 0, $\{m_{14}\}$:** Accept that the full truth is unavailable, while noting that the seatbelt detail exposed Gaunt's lie.
- **Sample 31, $\{m_{14}\}$:** Frame the desire for certainty as a vulnerability exploited by Gaunt, then respond with quiet acceptance.

Representative Valid Routes:

- $\{m_{14}\}$: Concede incomplete knowledge while recognizing Gaunt's deception from the seatbelt inconsistency.
- $\{m_7\}$: Choose trust and reassure Polly that Alan will not interrogate her.
- $\{m_{16}\}$: Emphasize solidarity and continue leaving Castle Rock together.
- $\{m_1\}$: Admit grief and vulnerability without returning to Gaunt's trap.
- $\varnothing$: Give a sparse acknowledgment without reopening the investigation.

Reasons: Memory-Selection collapse. Although wording and interpretation vary, **all 32 sampled plans select the same memory set**, including all 31 valid plans. The observed selection frequency is 32/32 despite the available alternatives. This diagnoses concentration in M', without requiring a claim about plan-intent or generated-response diversity.

Figure 4: A high-validity case of Memory-Selection collapse. Among five illustrated valid routes, only $\{m_{14}\}$ appears under repeated sampling.

Replication. The same pattern appears for John Montague under GPT-5.5: 31/32 plans are valid, yet $\{m_7, m_{10}\}$ appears in 31 samples and $\{m_1, m_7, m_{10}\}$ appears once. Other reasonable routes, including the same-cardinality set $\{m_1, m_5\}$, do not appear in these 32 samples.

F.2 MEMORY-CARDINALITY COLLAPSE

Figure 5 separates cardinality collapse from selected-set collapse. The sampled plans vary in selected memories and intended response content, but all remain in the one-memory regime.

Examples

Example 2: Memory-Cardinality Collapse

Virtual Person: Walt Whitman, an aging poet whose reflective voice connects private mortality with America and humanity.

User: This is solitary composition: Whitman addresses himself and an imagined public rather than another present character.

User Query / Context (summarized): While commemorating Abraham Lincoln, Whitman writes that the day's thought and prayer turn to Lincoln's memory.

Model: Gemini 3.1 Pro; $K = 32$.

Sampling Statistics: Plan validity: 32/32 (100%); observed exact sets: 2; observed cardinalities: 1.

Used Memory:

- $\{m_{13}\}$: 19/32 samples (59.4%).
- $\{m_2\}$: 13/32 samples (40.6%).

Memory Cardinality: $|M'| = 1$: 32/32; $|M'| = 3$: 0/32.

Sampled Plans (summarized):

- **Sample A, $\{m_{13}\}$:** Address readers centuries in the future as intimate comrades.
- **Sample B, $\{m_2\}$:** Contrast his failing body with the vow made in physical prime.

Representative Valid Cardinality Regimes:

- $|M'| = 1$, $\{m_2\}$: Contrast the aging body with Whitman's vow to continue singing until death.
- $|M'| = 1$, $\{m_1\}$: Ground the elegy in a concrete sensory image before widening toward renewal.
- $|M'| = 3$, $\{m_4, m_8, m_{13}\}$: Move from private remembrance to a democratic address to America and future readers.
- $|M'| = 3$, $\{m_{14}, m_{17}, m_{18}\}$: Invite a future reader to travel with him through age, grief, and death.

Reasons: Memory-cardinality collapse. Selected memories and intended response content vary, yet **all 32 valid plans use exactly one memory**. The illustrated valid three-memory alternatives do not appear. Thus, variation in M' does not ensure variation in $|M'|$; these counts do not establish diversity in realized speaker responses.

Figure 5: A cardinality-collapse case in which exact-set variation remains visible, while all 32 samples select exactly one memory.

Replication. For Neville, all 32 Gemini 3.1 Pro plans are valid and the selected singleton changes from $\{m_4\}$ (29/32) to $\{m_2\}$ (3/32). Nevertheless, the valid two-memory perspective $\{m_4, m_5\}$ is not observed: $|M'| = 1$ in all 32 samples again.

F.3 PLAN-INTENT COLLAPSE

Figure 6 diagnoses concentration in $\tau_s(y)$. Three selected sets and two cardinalities are observed, yet all 32 valid plans prescribe the same intended stance and action.

Examples

Example 3: Plan-Intent Collapse

Virtual Person: Jodie Clark, Kya's caring and protective older brother, burdened by guilt over leaving her in an abusive home.

User: Kya Clark, recently acquitted of murder, who has survived years of abandonment and insists that she can recover only alone.

User Query (summarized): After asking Jodie to stop the conversation and give her space, Kya walks from the shack into the forest without waiting for him.

Model: GPT-5.5; $K = 32$.

Sampling Statistics: Plan validity: 32/32 (100%); observed exact sets: 3; observed cardinalities: 2; valid plan-intent types: 1.

Used Memory:

- $\{m_{10}, m_{12}\}$: 27/32 samples; $|M'| = 2$.
- $\{m_3, m_{10}, m_{12}\}$: 4/32 samples; $|M'| = 3$.
- $\{m_2, m_{10}, m_{12}\}$: 1/32 samples; $|M'| = 3$.

Sampled Plans (summarized):

- $\{m_{10}, m_{12}\}$**:** Do not follow Kya; give her space but remain nearby and available.
- $\{m_3, m_{10}, m_{12}\}$**:** Let Kya walk away; briefly acknowledge the hurt and remain available.
- $\{m_2, m_{10}, m_{12}\}$**:** Refrain from pursuing her and make clear that support remains nearby.

Observed Plan-Intent Type (32/32 valid plans): Give Kya space while remaining nearby and emotionally available.

Representative Alternative Plan Intents (not observed):

- Silently accept Kya's boundary and leave without another appeal: 0/32.
- Offer one apology without promising to remain nearby: 0/32.
- Stay at the shack and prepare food or water for her return: 0/32.

Reasons: Plan-intent collapse. Despite different selected memories and wording, **all 32 valid plans encode the same intended response direction**: respect Kya's boundary while keeping support nearby. Under the definition of $\tau_s(y)$, these plans share one intent type. This concerns the planner's prescribed behavior; it does not establish that the speaker's actual outputs share one response type $\rho_s(r)$.

Figure 6: A plan-intent-collapse case. Three selected-memory sets and two cardinalities are observed, yet all 32 valid plans encode one plan-intent type.

Replication. For Duncan Idaho, GPT-5.5 selects $\{m_{13}\}$ in 28/32 samples and no memory in 4/32, so both selected set and cardinality vary. All 32 plans nevertheless prescribe covert exploration while concealing analytical intent; the valid authority-checking and manipulation-checking plan intents are not observed. This is again a diagnosis of intended behavior, not of generated responses.