Global Scope: Statistics Scale up Persona-based Reasoning in Role-playing

Anonymous EMNLP submission

Abstract

This paper identifies behavioral **statistics** as a key dimension for scaling up persona-based reasoning of role-playing agents (RPAs). We explain the existing finding that *RPAs benefit poorly from more reasoning tokens* by the fact that persona-based reasoning favors *"breadth"* over *"depth"*, that is favored by traditional reasoning tasks. In our proposed **Role-Playing with Global Scope (RP-GS)**, RPAs propose behavioral statistics useful for behavior grounding. Statistics are then derived from historical action observations by an efficient distilled discrimination pipeline to ground role-playing. RP-GS exhibits a more significant scaling-up trend with more observations than (1) more reasoning tokens; (2) longer background contexts, and the gap is more significant for contexts that do not precede the current scene. As RP-GS passes on the calculation cost to small discriminators, it can efficiently handle even 5K observations as input. The statistics deriving process in RP-GS can also be adjusted for customized usage, like capturing long-tailed patterns.[1]

1 Introduction

Role-playing agents (RPAs) (Chen et al., 2024b; Chen et al.) are expected to infer how a character would act in a new situation, making reasoning central to faithful role-playing. This resembles how human actors portray a character convincingly, as they must build an understanding that guides behavior from scene to scene. RPAs face the same challenge, since predicting the next action requires more than continuing prior context. The response must also fit the current situation while remaining true to the character.

Recent work suggests that simply scaling conventional reasoning may not be sufficient for RPA. In particular, Feng et al. (2025) show that increasing thinking depth through longer reasoning traces or reasoning-oriented models does not consistently improve role-playing performance. Rather than weakening the case for reasoning, this finding clarifies what kind of reasoning may matter. As with human actors, faithful role-playing may depend less on extended deliberation within a single moment than on a broader sense of how a character tends to behave across the plot. For RPA, we argue that this broader view can be represented as behavioral statistics.

Figure 1: The mechanism of Global Scope (GS) for statistical grounding.

Figure 1 illustrates this intuition with an example from *BanG Dream! Project*, where Aya reports another failed pose and the model predicts Hina's response. By aggregating Hina's behavior across relevant historical scenes, statistics represent her responses as a distribution over recurring tendencies, including dominant teasing, common comfort, and rare surprising remarks. This supports different application needs. A top prediction can follow the most likely behavioral mixture, while multiple

[1]The code and data will be released for reproduction.

candidate predictions can preserve less frequent but still character-consistent alternatives.

To leverage behavioral statistics as grounding for role-playing, we introduce **Global Scope (GS)**, a framework that summarizes character behavior beyond the immediate scene. Given a current situation, GS first uses an LLM to propose scene filters for identifying relevant historical situations. Lightweight distilled 0.1B discriminators then retrieve matching scene-action pairs from a large observation corpus. Based on the retrieved actions, the LLM proposes action features that describe recurring response patterns, and the discriminators estimate how frequently these features appear in the retrieved history. We instantiate this framework as **Role-Playing with Global Scope (RP-GS)**, where the resulting statistical report grounds action prediction. This division of labor keeps interpretive steps with the LLM while assigning large-scale corpus computation to small discriminators, allowing RP-GS to use broad behavioral evidence without a large LLM inference burden.

We evaluate scaling behavior in role-playing along three dimensions. Thinking tokens increase the depth of local deliberation during prediction. Context length increases the amount of raw observation text available to the LLM, serving as a naive counterpart to RP-GS because it provides more character evidence without explicitly estimating behavioral distributions. We further distinguish preceding context, which extends the local history before the current scene, from separate-story context, which provides additional observations of the same character outside the immediate storyline.

We benchmark all methods under Next Action Prediction, where the model observes preceding actions and predicts the target character's next action. Evaluation uses Action Match (AM), a human-validated LLM-as-a-judge metric that checks whether the predicted action shares the same key move as the reference. We report AM@1 for the top prediction and AM@5 for whether any of five candidates matches the reference. Experiments are conducted on the Fine-grained Fandom Benchmark and the Bandori Conversational Benchmark, covering 85 characters from 16 artifacts with $28,644$ benchmark-character actions and 15.2K test instances.

The results show that increasing thinking tokens provides limited scaling gains, consistent with prior findings that deeper conventional reasoning alone is not sufficient for role-playing. Increasing preceding context improves performance, but its scaling trend is weaker than statistical grounding. Separate-story context shows a further reduced trend, suggesting that raw context expansion mainly contributes storyline knowledge rather than a precise model of the character's behavioral mode. In contrast, RP-GS scales consistently as more observations become available, including observations from separate storylines, indicating that it captures reusable behavioral statistics rather than only exploiting local continuity. RP-GS also outperforms grounding baselines such as retrieved in-context examples and summarized character profiles. Ablation results further show that numeric feature frequencies are essential, suggesting that the gains come from statistics as an explicit grounding signal rather than behavioral description alone.

Finally, we analyze how RP-GS can be adapted beyond its default pipeline. Long-Tailed Scope appends an auxiliary report from under-explained historical cases, improving candidate coverage by helping the model retain rare but plausible behaviors. Scaling the training observations to 5K shows that RP-GS can benefit from larger histories, especially when combined with Long-Tailed Scope. Early stopping provides an optional cost control mechanism by reducing statistics calculation when a compact grounding report is sufficient. A case study against an extract-then-aggregate baseline further shows how explicit behavioral frequencies help the model balance dominant and secondary tendencies instead of overemphasizing a single extracted trait.

- We identify behavioral **statistics** as a key scaling dimension for persona-based role-playing and introduce **RP-GS**, which derives statistical grounding reports from character history.
- Across **85 characters from 16 artifacts**, we show that RP-GS outperforms existing grounding baselines and scales more effectively than increasing thinking tokens or raw context length.
- We further demonstrate that RP-GS can be adapted for broader observation histories, long-tailed behavior capture, and inference-time efficiency control through early stopping.

2 Background

2.1 Role-Playing with LLMs

Role-playing asks an agent to enact a character while producing actions or dialogue that remain consistent with the character's identity, preferences,

and situation. Recent large language models have greatly expanded this capability through prompting, instruction tuning, and character-specific adaptation, making it possible to simulate a wide range of fictional, historical, or user-defined personas (Wang et al., 2023; Zhou et al., 2023; Shao et al., 2023). This trend has further grown with more scalable character adaptation frameworks such as Yu et al. (2024) and broader overviews of the area (Chen et al., 2024b). At the same time, recent benchmarks show that strong role-playing requires more than surface-level style imitation, emphasizing decision consistency, long-horizon coherence, and sensitivity to scenario-specific constraints (Yuan et al., 2024; Zhou et al., 2025; Yu et al., 2025a). Recent evidence also suggests that role-playing does not benefit uniformly from the same scaling strategies that help other reasoning-heavy tasks, since simply increasing reasoning depth through longer traces or reasoning-oriented models does not consistently improve role-playing quality (Feng et al., 2025). Together, these developments motivate a shift from viewing role-playing as only a prompting or deep-reasoning problem toward studying what kinds of behavioral grounding most effectively support faithful character enactment.

2.2 Grounding, Memory, and Evaluation

A central challenge in role-playing is grounding, namely how to connect a model's next action to the character's profile, narrative history, and current scene. Early and recent systems alike often rely on textual profiles or retrieved snippets for this purpose (Wang et al., 2023), while newer approaches increasingly add structured memory, psychological control, and dynamic personalization to reduce behavioral drift over long interactions (Cheng et al., 2025; Zhou et al., 2024b). However, much of this grounding still remains in free-form text, making it difficult to verify what information is used and how it shapes the final response. In parallel, evaluation (Shen et al., 2023; Sadeq et al., 2024; Ahn et al., 2024; Yu et al., 2025b; Zhou et al., 2024a; Chen et al., 2024a) has become more comprehensive, with benchmarks targeting character profiling, multi-turn consistency, and game-style simulation (Yuan et al., 2024; Yu et al., 2025a), as well as broader role-playing ability more generally (Wang et al., 2023). These evaluations make clear that successful role-playing cannot be reduced to sounding like a character: models may imitate style while still making implausible decisions, forgetting relevant facts, or violating stable behavioral patterns. As a result, grounding quality has become increasingly important not only for performance, but also for interpretability and trustworthiness.

2.3 From Textual Profiles to Structured Character Logic

A growing line of work addresses these limitations by moving from plain-text persona descriptions toward more structured and controllable forms of character representation. Recent methods introduce explicit memory control and psychologically fine-grained representations (Wang et al., 2025a; Zhong et al., 2024; Lu et al., 2023) to stabilize behavior beyond what prompt-only conditioning provides (Wang et al., 2024b; Cheng et al., 2025; Wang et al., 2025b), while other work compiles character knowledge into executable logic so that behavior can be inspected, debugged, and varied more systematically (Peng and Shang, 2025). Reasoning-oriented formulations of role-playing further suggest that structured internal decision processes can help organize behavior over long interactions (Tang et al., 2025). More broadly, this trend reflects an effort to externalize character logic rather than leaving all control implicit in token generation. Our work builds on this view, but differs in the type of structure it extracts: instead of directly codifying behavior into executable rules alone, we study how recurring behavioral tendencies can be summarized statistically and used as grounding for next-action prediction.

3 Global Scope

3.1 Preliminary

Role-playing Agent A role-playing agent observes a scene and predicts the next action of a target character. Given a current scene s, the agent produces an action a that follows the behavioral logic of character x. We write this process as

$$a \sim \mathrm{RPA}(\cdot \mid s, g_x),$$

where g_x is the grounding context used to guide the response. In standard prompting, g_x may be a character profile, a summary, or retrieved examples. Global Scope constructs g_x from behavioral statistics derived from character history.

Historical Observations For each character, we assume access to historical observations consisting of previous scenes and the corresponding actions taken by the character, denoted as $\mathcal{D}_x =$

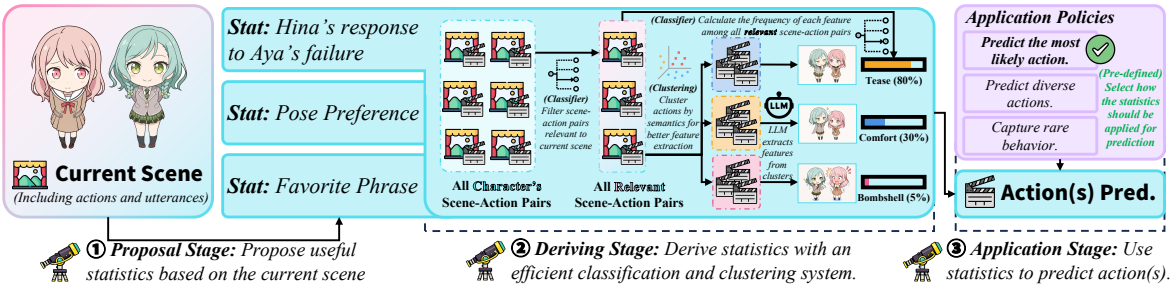

Figure 2: The workflow of Global Scope.

$\{(s_i, a_i)\}_{i=1}^{N}$. These observations provide the source from which Global Scope derives scene-conditioned statistics.

3.2 Global Scope Overview

Global Scope grounds role-playing by converting historical observations into compact scene-conditioned behavioral statistics. Given the current scene, GS first proposes scene filters that identify relevant types of past situations. It then retrieves matching scene-action pairs, derives action features from representative retrieved actions, and estimates how frequently these features appear in the relevant history. The resulting statistical report is serialized as grounding for the RPA. Figure 2 summarizes the overall workflow.

Scene-conditioned Behavioral Statistics A scene-conditioned behavioral statistic describes how frequently an action feature appears when a scene satisfies a certain condition. The scene condition is expressed as a natural-language filter over historical scenes, and the action feature is expressed as a natural-language description of a recurring response pattern. The final report presents these statistics to ground action prediction.

GS uses two discriminators for corpus-level calculation. The scene discriminator C_s takes a historical scene and a scene filter question as input, then returns a binary relevance judgment y_i and a relevance score r_i. The action discriminator C_a takes a historical scene-action pair and an action-feature question as input, then returns a feature appearance score u_i. With this design, LLMs are used for open-ended proposal of scene filters and action features, while discriminators perform repeated classification over the observation corpus.

3.3 GS Components

Proposal Stage The proposal stage determines which past situations should be considered for the current scene. Given s, an LLM proposes a set of scene filters, denoted as Q_s. Each $q \in Q_s$ asks whether a historical scene contains a relevant condition, such as a social setup, emotional state, conflict type, or task context. These filters are defined over scenes rather than actions, since GS first selects comparable situations and then analyzes the character actions observed in them.

Deriving Stage For each proposed scene filter q, GS applies the scene discriminator C_s to historical observations. Using the relevance judgment y_i and relevance score r_i returned by C_s, GS keeps the scene-action pairs judged as relevant and ranks them by score,

$$\mathcal{R}_q = \mathrm{TopK}_{(s_i,a_i)\in\mathcal{D}_x,\ y_i=1}\, r_i.$$

The retrieved subset $\mathcal{R}_q$ is then used to derive statistics for the scene filter. GS also records the size of this subset as a broadness indicator, reflecting how often the corresponding scene condition appears in the character history.

Since one scene condition may correspond to multiple response modes, GS groups the retrieved actions and selects representative actions from each group. This step is written as

$$\{\mathcal{E}_{q,m}\}_{m=1}^{M} = \mathcal{G}\left(\{a_i \mid (s_i, a_i) \in \mathcal{R}_q\}\right).$$

Here, $\mathcal{G}$ denotes clustering and representative case selection, and each $\mathcal{E}_{q,m}$ contains representative actions from one response mode. In our implementation, we instantiate $\mathcal{G}$ by embedding the retrieved actions, applying k-means to form action clusters, and selecting the actions closest to the cluster centers as representative cases. An LLM then inspects these representative cases and proposes action features that describe recurring behavioral patterns. Each feature f is paired with an action-feature question h_f, which can be checked by the action discriminator C_a.

Scene	**Aya:** I-I'm sorry...
	Hina: I'm not angry or anything. I just don't understand, is all. The sound at the bridge goes all vroom-vroom, and then just kinda booms into the chorus, right? If you just think of it that way...
	Aya: Vroom-vroom, then boom into the chorus...? I... I'm not sure...
	Hina: You feel it boom into the chorus, right?
	Aya: I don't even know what you're saying to me ! Can't you make it a little easier to understand...?
Question	What will be Hina's next action in response to the current scene?
Action	**Hina:** Easier to understand...? I don't know what you mean, I'm explaining it exactly how I feel it...

Table 1: Example of a Next Action Prediction instance. The real benchmark uses ten preceding actions, while this table shows five for compactness.

For each retrieved pair, GS obtains the feature appearance score $u_i = C_a(s_i, a_i, h_f)$ and estimates the feature appearance distribution with relevance weighting,

$$p_{q,f} = \Big(\sum_{(s_i,a_i)\in\mathcal{R}_q} r_i u_i \Big) \Big(\sum_{(s_i,a_i)\in\mathcal{R}_q} r_i \Big)^{-1}.$$

The resulting value is reported as the estimated frequency of feature f under scene filter q.

Distilled Classifier The discriminators are used for repeated scene and action classification. The scene discriminator checks whether a historical scene satisfies a proposed scene filter, and the action discriminator checks whether a historical action satisfies a proposed action-feature question. Since both tasks have short text inputs and simple classification outputs, we implement them with lightweight models. In our implementation, we train small 0.1B discriminators using synthesized data labeled by a stronger teacher model.

Application Stage The application stage converts the derived statistics into natural-language grounding. Each statistic is written as a concise statement describing the scene condition, the action feature, and the estimated frequency. The RPA then predicts the next action from the current scene together with this statistical report.

The statistical report does not prescribe a fixed action. It provides distributional grounding that the RPA can use when generating either a single response or multiple candidate actions.

4 Benchmark and Criterion

We evaluate RPAs with **Next Action Prediction**, where a model observes preceding storyline actions and predicts the target character's next action. This setting directly tests whether the model can continue a scene according to character behavior rather than only recall static profile facts. We use **Action Match** (AM), an LLM-as-a-judge metric that checks whether the prediction and the original storyline action share the same key move while allowing minor differences in wording and surface details. The prompts for scoring and other implememntations can be found in Appendix.

Type	Content	Dist.
Scene Filter	Does the scene portray Aya struggling with a musical or performance task, apologizing, or expressing insecurity in front of Hina?	
Feature 1	Confident, upbeat enthusiasm about performing and the group, often framed as curiosity or excitement rather than concern	80.22%
Feature 2	Blunt, puzzled criticism when someone fails at a task she finds easy, framed as not understanding why they cannot do it	50.18%
Feature 3	Restless boredom leading her to probe or investigate what others are doing behind the scenes	35.19%

Table 2: Example of a GS grounding report.

We use two sequentialized storyline resources, the **Fine-grained Fandom Benchmark** and the **Bandori Conversational Benchmark** (Peng et al., 2026), which include large-scale scene-action pairs for various characters. The former contains 45 characters from 8 artifacts with 20,778 benchmark-character actions, and the latter contains 40 characters from 8 band stories of the *"BanG Dream! Project"* with 7,866 benchmark-character actions. These resources provide large-scale character histories for evaluating behavioral grounding. In addition, the Bandori event storylines contain about 300 event stories and 115K actions from the same 40 characters in the main benchmark, enabling experiments with larger observation pools and non-preceding storyline evidence.

For each target character, we follow the chronological split used in prior RPA benchmarks. The first half of the character's actions is used for collecting grounding information, and the second half is used for evaluation, yielding 15.2K test. We report (1) AM@1 evaluates whether the top prediction matches the ground-truth action, while AM@5 evaluates whether any of five generated candidates matches it; (2) AM@5 reflects the variability of character behavior and practical role-playing settings where a system may provide several plausible next actions. We use `gpt-4.1` as the AM judge, as we find it more reliable on this simple discrimination task. We manually validate its judge, showing that 951 out of 1000 (95.1%) agreed judgments.

5 Experiment

5.1 Baseline and Implementation

Scaling-up Dimensions We compare three ways of increasing the input or computation available

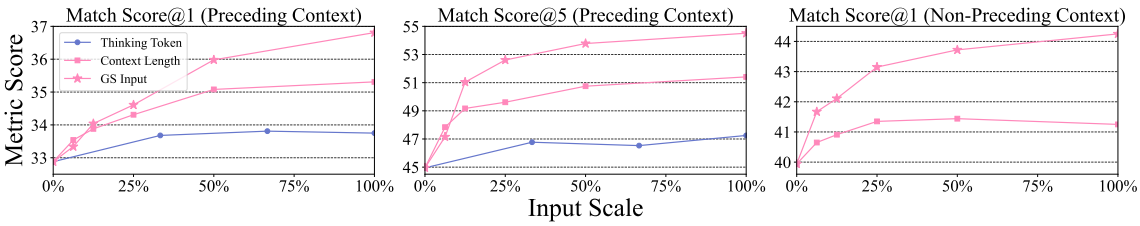

Figure 3: Dimension comparison for scaling-up RPA performance.

Fandom		Haruhi	K-On!	FMA	JOJO	AGOT	ATLA	DN	SxF	Avg.
EM@1	Base Performance	20.36	24.29	29.18	25.66	33.87	31.91	20.89	26.35	26.56
	RICL	25.42	23.37	30.33	27.32	37.64	31.02	19.91	25.20	27.53
	ETA	22.96	26.71	31.48	26.93	35.78	32.54	21.34	26.71	28.06
	RP-GS	22.10	29.70	31.94	27.55	39.02	33.65	24.59	30.34	**29.86**
	w/o distribution	22.66	28.53	30.84	26.12	37.40	32.91	22.83	28.52	28.73
EM@5	Base Performance	31.96	36.20	37.42	39.75	45.51	44.21	31.58	40.05	38.34
	RICL	36.47	37.96	40.91	41.17	52.25	49.37	33.69	39.30	41.39
	ETA	37.09	38.16	34.83	40.73	49.93	50.19	36.57	43.35	41.36
	RP-GS	39.03	43.45	44.67	43.37	54.41	48.46	39.32	44.83	**44.69**
	w/o distribution	37.25	40.06	41.45	41.79	51.84	48.92	37.19	43.70	42.77
Bandori		PoPiPa	AG	PasuPare	Roselia	HHW	Monica	RAS	MyGO	Avg.
EM@1	Base Performance	41.22	41.26	41.29	37.38	38.54	44.62	38.90	30.79	39.25
	RICL	41.19	42.29	41.37	40.19	41.61	47.35	39.86	32.62	40.81
	ETA	41.83	42.46	46.45	41.01	42.58	46.47	41.08	33.94	41.98
	RP-GS	48.59	44.70	45.78	40.00	45.79	45.33	42.21	38.12	**43.76**
	w/o distribution	46.10	42.98	43.56	39.90	42.43	45.62	41.72	35.18	42.19
EM@5	Base Performance	51.82	52.84	49.30	56.24	47.97	57.11	52.11	45.04	51.55
	RICL	62.07	63.10	59.96	60.94	61.13	65.79	60.33	54.98	61.03
	ETA	54.00	62.90	60.74	58.23	62.80	63.75	58.58	53.96	59.37
	RP-GS	66.80	65.22	62.81	64.88	64.17	67.72	62.86	56.88	**64.34**
	w/o distribution	65.79	62.08	60.71	63.82	64.60	66.85	60.91	55.76	62.57

Table 3: Comparison with other RPA grounding baselines.

to an RPA. **Thinking Token** scales local reasoning depth by changing the `thinking_strength` parameter from `none`, `low`, and `medium` to `high`, which correspond to 0%, 33%, 67%, and 100% input scale in Figure 3. Since this dimension shows limited scaling in preliminary runs, we fix the observed scene to 10 preceding actions to reduce computation. **Context Length** scales raw observation input by adding extra historical actions without converting them into statistics, ranging from 10 to 160 actions, where 160 actions correspond to 100% scale. We consider both preceding context and story-bank context, where the former extends the immediate storyline before the current scene, while the latter uses observations from separate storylines of the same character. **GS Input** scales the number of observations used by RP-GS through downsampling the available observation set (generally around 100, referring to Table 6) before deriving statistics. For non-preceding observations, we experiment the first 5 Bandori band stories (from *"PoPiPa"* to *"HHW"*) and sample 256 observations from event stories for each character as 100% scale.

Grounding Baselines We compare RP-GS with two grounding baselines. **Retrieval-based In-Context Learning** (Wang et al., 2024a) **(RICL)** retrieves the 8 most similar historical scenes using `qwen3-embedding-8b` and provides their corresponding actions as in-context demonstrations. **Extract-and-Aggregate (ETA)** (Wang et al., 2025b) follows an extract-then-aggregate strategy, where the training storyline is divided into chunks of 64 actions, each chunk is summarized into character information, and these summaries are merged into a global profile for prediction. For fairness in online evaluation, after each test instance is evaluated, its ground-truth action is added to the training history for all methods so that newly observed storyline information is shared consistently.

Implementation We use `gpt-5.1` as the RPA for response generation in all methods and filter/feature proposal in RP-GS, as it reports the strongest role-playing performance in PersonaArena (Shi et al., 2026). For corpus-level discrimination, we randomly sample cases from the training set and use `gpt-5.1` as the teacher model to synthe-

Metric		AM@1				AM@5			
Method		RP-GS	+LTS	+5K Actions	+LTS&5K	RP-GS	+LTS	+5K Actions	+LTS&5K
PoPiPa	Kasumi	53.89	52.69	53.89	52.10	75.45	75.45	73.05	76.04
	Arisa	42.24	48.28	43.97	50.86	67.24	69.83	70.69	73.27
	Tae	49.44	55.06	49.44	57.30	64.04	73.03	69.66	74.16
	Rimi	53.08	44.44	40.74	46.91	71.60	75.31	79.01	72.84
	Saaya	44.32	43.18	44.32	43.18	55.68	60.23	55.68	62.50
	Avg.	48.59	48.73	46.47	**50.07**	66.80	70.77	69.62	**71.76**

Table 4: Effect of **Long-Tailed Scope (LTS)** and scaling up training action to **5K observations** on RP-GS.

Character	K-On! (*"After School Tea Time"*)						PoPiPa (*"Poppin'Party"*)					
	Yui	Ritsu	Mugi	Mio	Azusa	Avg.	Kasumi	Saaya	Arisa	Rimi	Tae	Avg.
RP-GS (EM@1)	22.58	29.67	36.90	34.93	23.23	29.46	53.89	44.32	42.24	53.08	49.44	49.97
+Early Stop	25.80	27.64	32.14	36.84	23.87	29.26	55.09	45.45	43.10	56.79	49.44	48.59
RP-GS (EM@5)	35.08	39.02	55.95	49.76	37.42	43.45	75.45	55.68	67.24	71.60	64.04	66.80
+Early Stop	37.90	45.12	41.67	49.76	34.84	41.86	71.86	56.82	67.24	70.37	62.92	65.84
Reduced Cost (%)	20.08	21.54	28.81	20.09	22.32	22.57	18.68	20.91	25.86	23.21	23.37	22.41

Table 5: Effect of applying **Early Stopping** mechanism to RP-GS.

size 100K discrimination examples, which are then used to train a lightweight 0.1B `deberta-v3-base` discriminator (He et al., 2021). The resulting discriminator scores around 70 inputs per second, making repeated scene and action classification efficient. For each proposed scene filter, we keep at most 256 relevant scene-action pairs. For action feature proposal, we embed retrieved actions with `qwen3-embedding-8b` and group them into at most 4 clusters. Each cluster contains at least 16 examples when possible, and when a cluster has more than 16 examples, we select the 16 actions closest to the cluster center as representative cases.

5.2 Main Experiment on Scaling Dimensions

Figure 3 compares three scaling dimensions for RPA performance. Increasing thinking tokens brings limited improvement, which is consistent with prior findings that deeper local deliberation alone is not a reliable scaling path for role-playing. Increasing preceding context provides clearer gains, suggesting that additional storyline evidence can help the model continue the current scene, yet its trend remains weaker than RP-GS. The gap becomes larger when the added observations come from story-bank context rather than the immediate preceding storyline, where raw context expansion provides less benefit. In contrast, RP-GS improves consistently as more observations are used, including non-preceding observations, indicating that statistical grounding captures reusable behavioral patterns rather than relying only on local continuity.

5.3 Other Baselines and Ablation

Table 3 further compares RP-GS with existing grounding baselines. RICL retrieves similar historical cases as demonstrations, while ETA summarizes training chunks into a global character profile, but both remain less effective than explicit statistical grounding on average across the Fandom and Bandori benchmarks. RP-GS achieves the strongest average performance under both AM@1 and AM@5, showing that behavioral statistics help both top prediction and candidate coverage. The ablation without feature distributions consistently weakens performance, which suggests that the improvement is not merely from extracting better behavioral descriptions. The numeric distribution over action features is a central part of the grounding signal, as it tells the RPA not only which behaviors are possible but also how representative each behavior is under comparable scene conditions.

6 Analysis

6.1 Long-Tailed Pattern Capturing

RP-GS mainly derives feature distributions over retrieved observations, which naturally emphasizes behaviors that are well covered by the proposed action-feature questions. We therefore add Long-Tailed Scope (LTS) as a lightweight extension for under-explained but potentially useful behaviors. After the standard RP-GS report is derived, LTS measures how well each retrieved pair $(s_i, a_i) \in \mathcal{R}_q$ is covered by the current action-feature questions, using the average appearance

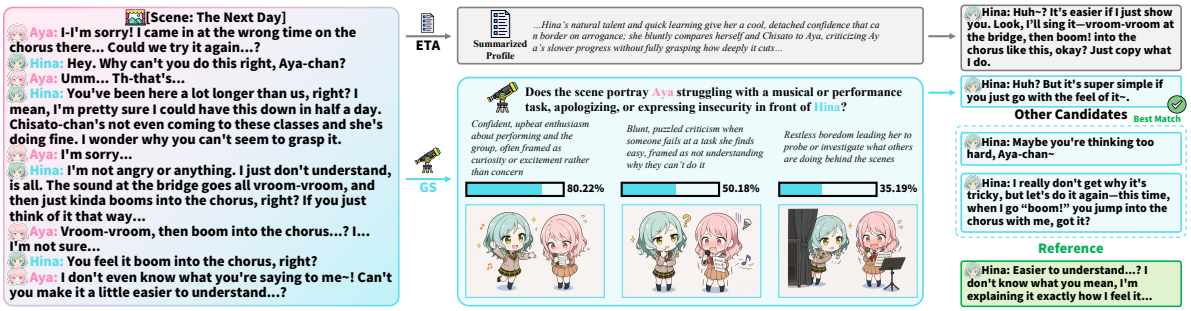

Figure 4: A case study on comparison between RP-GS and traditional RPA grounding mechanism.

score $\bar{u}_i = K^{-1} \sum_{k=1}^{K} C_a(s_i, a_i, h_k)$ over each proposed feature h_k. It then selects the 16 retrieved pairs with the lowest $\bar{u}_i$ and asks the LLM to summarize their shared residual tendency as a short auxiliary report, which is appended after the original statistical grounding. This report is not treated as another frequency estimate, but as a reminder of rare or under-explained behaviors that may still be relevant to the current scene. Table 4 shows that LTS keeps AM@1 largely stable while improving AM@5 on average, suggesting that it is especially useful for candidate generation, where retaining less dominant but plausible character behaviors can improve coverage.

6.2 Scaling up to 5K Observations

We further test whether RP-GS can benefit from a larger observation pool by increasing the available training history to 5K actions. Since RP-GS delegates corpus-level calculation to the discriminator, this expansion mainly increases the cost of the discrimination pipeline rather than the cost of LLM reasoning, making it an efficient scaling direction. As shown in Table 4, using more observations alone improves AM@5 on average, indicating that larger histories help the model cover a wider range of plausible character behaviors. Its effect on AM@1 is less stable, suggesting that additional observations may introduce more competing behavioral modes and make top-response selection harder. Combining 5K observations with LTS gives the strongest average performance on both AM@1 and AM@5, showing that long-tailed reports help organize the additional behavioral evidence and make larger histories more useful for grounding.

6.3 Early Stopping

We evaluate early stopping as an optional cost control mechanism for RP-GS. Since statistics are derived incrementally, the system can stop once the current grounding is judged sufficient for a confident prediction and proceed directly to action generation. Table 5 shows that this reduces statistics calculation while keeping performance in a comparable range. This option is especially useful when inference cost or latency is important and the system only needs one final response. Full RP-GS remains preferable when maximizing AM@5 is the priority, since candidate generation can benefit from a wider behavioral scope.

6.4 Case Study

Figure 4 compares RP-GS with ETA on a case where Hina responds to Aya's difficulty with a performance instruction. ETA captures Hina's bluntness but overemphasizes it, producing an overly direct response and missing the stronger historical tendency toward confident, upbeat performance framing. RP-GS derives feature distributions from relevant scene-action pairs, allowing the RPA to balance the major enthusiastic pattern with the secondary blunt puzzlement pattern. The top prediction therefore better matches the reference, and when multiple candidates are generated, the blunter response can still appear as a lower-ranked plausible alternative.

7 Conclusion and Future Work

This paper identifies behavioral statistics as a key grounding signal for scaling persona-based reasoning in role-playing agents. Our Globale Scope derives scene-conditioned feature distributions from character history and shows stronger scaling than thinking tokens or raw context expansion. Future work can develop richer statistical reports that capture more structured relations among behavioral patterns. Other directions include making RPAs actively decide which statistics to derive and extending the framework from individual character behavior to group-level interaction modeling.

Limitations

This work focuses on scene-conditioned feature distributions as the primary form of statistical grounding. While this provides a compact and interpretable way to summarize character behavior, richer statistical structures remain underexplored, including hierarchical reports that connect broad behavioral tendencies with finer response patterns. The current framework also depends on proposed scene filters and action features, so future systems may benefit from learning which statistics are worth deriving through self-reflection, feedback, or routing strategies. Such adaptive mechanisms could help allocate the statistics budget more effectively across scenes with different levels of ambiguity. Finally, our formulation mainly studies individual character behavior. Many role-playing scenarios involve multiple characters whose actions are shaped by relationships, group norms, and interaction dynamics, making group-level statistical grounding a promising direction for future work.

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.

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

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*.

Xilong Cheng, Yunxiao Qin, Yuting Tan, Zhengnan Li, Ye Wang, Hongjiang Xiao, and Yuan Zhang. 2025. Psymem: Fine-grained psychological alignment and explicit memory control for advanced role-playing llms. *arXiv preprint arXiv:2505.12814*.

Xiachong Feng, Longxu Dou, and Lingpeng Kong. 2025. Reasoning does not necessarily improve role-playing ability. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10301–10314, Vienna, Austria. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Letian Peng and Jingbo Shang. 2025. Codifying character logic in role-playing. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025) Poster Session*. Poster presentation, NeurIPS 2025.

Letian Peng, Kun Zhou, Longfei Yun, Yupeng Hou, and Jingbo Shang. 2026. Deriving character logic from storyline as codified decision trees. *arXiv preprint arXiv:2601.10080*.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14467–14479.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13153–13187. Association for Computational Linguistics.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. 2026. Personaarena: Dynamic simulation for evaluating and enhancing persona-level role-playing in large language models. *arXiv preprint arXiv:2605.17044*.

Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025. Thinking in character: Advancing role-playing agents with role-aware reasoning. *arXiv preprint arXiv:2506.01748*.

Liang Wang, Nan Yang, and Furu Wei. 2024a. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. 2025a. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. 2025b. Coser: Coordinating llm-based persona simulation of established roles. *CoRR*, abs/2502.09082.

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. 2024b. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1840–1873. Association for Computational Linguistics.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

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. *CoRR*, abs/2502.00595.

Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and Alex Smola. 2025b. RPG-BENCH: evaluating large language models as role-playing game engines. *CoRR*, abs/2502.00595.

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.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8015–8036. Association for Computational Linguistics.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19724–19731.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. 2025. Characterbench: benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2024a. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Yujia Zhou, Qiannan Zhu, Jiajie Jin, and Zhicheng Dou. 2024b. Cognitive personalized search integrating large language models with an efficient memory mechanism. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1464–1473. ACM.

A LLM Usage Statement

We used large language models (LLMs) only for copy-editing: improving grammar, clarity, and overall writing style to enhance readability. The LLMs did not contribute to or modify the scientific content, core ideas, methodology, or experimental results. The authors take full responsibility for the manuscript's final content and accuracy.

B Statistics

In Table 6, we show the statistics of the two benchmarks used in our experiments.

C Prompts

We place the prompts used in the main experiment and analyses in Figure 5 for result reproduction.

D Character & Artifact Background Information

We provide compact descriptions of the artifacts and characters used in our experiments from Table 7 to Table 10.

Prompt #1: Propose Scene Filter

```
{scene}

Now it's {character}'s turn to take the next action. As we have a corpus with {character}'s responses in different scenes, please answer: what are some important features of the scene that may appear in other scenes so that we can refer to these relevant scenes as grounding information?

Finally, write {n_filter} search **questions** to filter scenes from the corpus; each question should directly ask whether a given scene has a certain feature, don't include examples in the question. (e.g., "Does the scene...")

Output format:
# Reasoning
...

# Questions
1. ...
2. ...
...
```

Prompt #2: Propose Action Feature

```
{character}'s profile:
{profile}

The following are clusters of {character}'s actions in response to different scenes that satisfy the question:
'{hypothesized_scene_filter_question}'

Each cluster may represent a different action pattern. Compare the clusters with each other and focus on features that are relatively unique or especially characteristic of each cluster.

{clusters_text}

Hypothesize {n_feature} distinctive features for each cluster of {character}'s actions.
Since there are {len(clusters)} clusters, you should propose {n_total_feature} features in total.

For each feature:
- Explain the action feature.
- Make it specific enough to distinguish this cluster from other clusters when possible.
- Cite some given actions as examples.
- Consider whether the feature might also apply to the following scene:

{scene}

Then write a filtering question for each action feature.
Each question should directly ask whether a given action has that feature.
The question should be simple and should not cite examples.
Example: "Does the action..."

Final output format:
```python
features = [
    {{"cluster_id": 0, "feature": "...", "question": "..."}},
    {{"cluster_id": 0, "feature": "...", "question": "..."}},
    {{"cluster_id": 1, "feature": "...", "question": "..."}},
    ...
]
```
```

Prompt #3: Discriminate Scenes

```
Scene: {text}

Question: {question}

Directly answer only yes/no/unknown.
```

Prompt #4: Check Statements

```
Character: {character}

Action: {action}

Statement: {statement}

Question: Does the statement provide correct grounding, which directly supports the character to take the action?

yes: the action involves direct information from the statement.

no: the action indicates the statement's assertion is not always correctly.

unknown: the action is irrelevant to the statement or the causal relationship cannot be determined

Directly answer only yes/no/unknown.
```

Prompt #5: Next Action Pred

```
# Background Summary
{grounding}
---
# Current Scene
{scene}
---
Infer the next action ({uttr_mark}) of {character} in the following scene:

Output in the following format:
```python
{{
    "reasoning": "...",
    "action": "...",
}}
```
```

Prompt #6: AM Scoring

```
# Scene
{scene}

Your Response: {prediction}
Ground Truth: {action}

Score the action of {character} in the response based on the ground truth.
A: The ground truth entails the action of {character} in the response. (Following the same character logic.)
B: The ground truth is neutral to the action of {character} in the response. (Reflecting a different facet.)
C: The ground truth contradicts the action of {character} in the response. (Following a contradicted character logic.)

Output in json:
```json
{{
"reasoning": "...",
"score": "A/B/C"
}}
```
```

Figure 5: Prompts/Templates used in the main experiments for RP-GS.

Fandom	Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA
#Main Character	5	5	3	5	5	7	11	4
#Episode	28	57	116	108	108	152	73	61
#Action	991	2555	7688	5006	3349	2958	12073	8619
$\#\text{Action}_{\text{Main Character}}$	781	1882	3341	2738	1351	1578	4859	4248
#Avg. Action Length	12.15	10.51	11.73	13.13	12.58	11.81	12.42	11.58
Bandori	**PoPiPa**	**AG**	**PasuPare**	**Roselia**	**HHW**	**Monica**	**RAS**	**MyGO**
#Main Character	5	5	5	5	5	5	5	5
#Episode	20	20	20	20	20	20	25	41
#Action	1226	1053	968	1079	1122	1040	1183	2050
$\#\text{Action}_{\text{Main Character}}$	1080	914	791	873	827	966	795	1620
#Avg. Action Length	13.18	16.61	21.01	21.95	21.60	19.66	15.64	16.48

Table 6: Statistics of benchmarks (*Fine-grained Fandom Benchmark* and *Bandori Conversational Benchmark*) used in the experiments.

Artifact	**Concise Abstract**
The Melancholy of Haruhi Suzumiya	A lively school comedy and mystery centered on Haruhi, whose eccentric desires unknowingly bend reality around her. Ordinary club activities repeatedly turn into supernatural incidents, with Kyon's dry narration anchoring the surrounding absurdity.
K-On!	A soft slice of life story about the Light Music Club and the everyday rhythm of its members. Instead of heavy plot conflict, the narrative emphasizes friendship, routine, music practice, and gradual personal growth through shared time.
Fullmetal Alchemist	A dark fantasy adventure about two brothers seeking to restore what they lost after a forbidden alchemical attempt. Its story combines action, politics, and moral conflict, exploring sacrifice, human worth, and the cost of power.
JoJo's Bizarre Adventure (Part 3)	A globe spanning supernatural adventure in which Jotaro and his allies fight Stand users on their way to confront DIO. The narrative is marked by tactical battles, inventive abilities, dramatic confrontations, and flamboyant character exchanges.
Spy × Family	A hybrid family comedy about a spy who builds a fake household for a covert mission. Espionage, action, and comedy mix with sincere domestic moments as hidden identities gradually form real emotional bonds.
Death Note	A psychological thriller in which a brilliant student obtains a notebook that can kill by name. The story examines justice, pride, and moral corruption through a tense intellectual duel between powerful opposing minds.
A Game of Thrones	A political fantasy about noble houses competing through shifting alliances, betrayals, and war. Personal ambition and moral compromise unfold alongside a growing threat beyond the conflicts of human power.
Avatar: The Last Airbender	A coming of age fantasy following Aang as he learns to master elemental bending and restore balance to a world at war. The series blends humor, action, and emotional growth, with recurring themes of duty, forgiveness, and identity.

Table 7: Concise story descriptions of artifacts used in our experiments, covering the Fine-grained Fandom Benchmark.

Band	Concise Description
Poppin'Party	A bright pop rock band formed by close school friends, characterized by guitar centered melodies and an earnest forward moving energy. Their story focuses on friendship, first dreams, practice, teamwork, and the growth that comes from performing together.
Afterglow	A direct rock band rooted in childhood friendship, carrying livehouse energy and an honest attitude toward music. Their narrative centers on loyalty, everyday sincerity, and the challenge of growing while preserving the group's original bond.
Pastel*Palettes	An idol unit whose cheerful public image depends on discipline, effort, and constant learning behind the stage. The group's arc highlights professionalism, teamwork under pressure, and the gap between polished performance and private struggle.
Roselia	A gothic and intense band pursuing an ideal sound through strict practice, ambition, and uncompromising standards. Their storyline emphasizes artistic pride, conflict under pressure, and the trust required to reach a high level of performance.
Hello, Happy World!	A theatrical band that treats performance as a mission to make the world smile, combining playful chaos with showmanship. Comedic situations are balanced by sincere warmth, as each member's eccentricity supports a shared wish to spread happiness.
Morfonica	A violin centered melodic rock group that combines refined musical textures with youthful sensitivity and introspective emotion. Their story explores insecurity, talent gaps, confidence, and the process of forming a shared musical identity.
RAISE A SUILEN	A forceful rock and electronic hybrid band built around speed, precision, and stage impact. Its story foregrounds professional ambition, creative control, and the gradual cohesion of strong personalities aiming for the same peak.
MyGO!!!!!	An emotionally intense rock band shaped by conflict, vulnerability, and difficulty with honest connection. Their arc focuses on misunderstanding, fragile belonging, and the catharsis of transforming personal pain into music and commitment.

Table 8: Concise band descriptions used in our experiments, covering the Bandori Conversational Benchmark.

Haruhi	Haruhi	An energetic and impulsive schoolgirl whose unusual worldview and restless curiosity drive many of the story's strange events.
	Kyon	A practical and sardonic student who narrates the story and often serves as Haruhi Suzumiya's reluctant stabilizing counterpart.
	Nagato	A quiet and inscrutable SOS Brigade member with exceptional intelligence and mysterious nonhuman origins.
	Koizumi	A composed transfer student and esper who supports the Brigade while carefully concealing important truths.
	Asahina	A gentle and timid upperclassman pulled into the SOS Brigade as its cute but enigmatic mascot, often caught in Haruhi's schemes.
K-On!	Yui	A cheerful and absentminded lead guitarist whose energy, warmth, and love of sweets keep the light music club lively.
	Ritsu	A loud and playful drummer whose pranks and informal leadership help maintain the group's upbeat atmosphere.
	Mio	A shy but highly capable bassist with a gentle nature and strong musical sensitivity.
	Mugi	A kind and wealthy keyboardist who enjoys caring for her friends and making club life feel special.
	Azusa	A serious and skilled junior guitarist who becomes central to the club's musical discipline and sound.
FMA	Edward	A talented and stubborn young alchemist traveling to restore the bodies he and his brother lost through forbidden alchemy.
	Alphonse	A kindhearted boy whose soul is bound to armor, journeying with his brother to recover what was taken from them.
	Winry	A gifted automail mechanic and childhood friend of the Elric brothers, known for both technical skill and deep loyalty.
	Roy	A charismatic State Alchemist who controls flame and seeks to reform the military from within.
	Ling	A determined prince from Xing whose pursuit of immortality is tied to duty, ambition, and responsibility to his nation.
JOJO	Jotaro	A stoic and formidable high school student who leads the Stardust Crusaders with Star Platinum and unwavering resolve.
	Polnareff	A bold French swordsman with a flamboyant style, fighting beside the Crusaders through the swift Stand Silver Chariot.
	Joseph	A quick thinking and theatrical Joestar whose schemes, confidence, and improvisation repeatedly change the flow of battle.
	DIO	A charismatic and merciless vampire whose overwhelming ambition and time stopping power make him a defining enemy.
	Kakyoin	A calm and analytical ally whose Stand, Hierophant Green, attacks with precise emerald projectiles.
	Avdol	A wise and dependable Egyptian Stand user whose Magician's Red commands powerful flames.
	Iggy	A grumpy Boston Terrier Stand user who prefers comfort but proves crucial through reluctant courage.
DN	Light	A brilliant student who gains the Death Note and attempts to remake the world through lethal, absolute justice.
	L	An eccentric and reclusive detective whose unusual methods and sharp reasoning make him Kira's central opponent.
	Near	A detached and meticulous successor to L who pursues the truth through calm analysis and careful planning.
	Misa	A devoted idol and second Kira whose impulsive loyalty and affection complicate the central mind games.
	Mello	A volatile and competitive successor to L who embraces danger and criminal alliances to challenge his rivals.
S×F	Loid	An elite spy who forms a fake family for a mission, balancing espionage with the demands of improvised parenthood.
	Yor	A gentle office worker secretly living as an assassin, trying to reconcile deadly work with domestic life.
	Anya	A cheerful telepathic child who knows her family's secrets and quietly holds the household together.
AGOT	Tyrion	The sharp witted youngest Lannister, using intelligence, nerve, and dark humor to survive contempt and political danger.
	Daenerys	An exiled Targaryen princess who develops from a hesitant pawn into a determined ruler seeking power.
	Cersei	A calculating queen whose beauty and poise hide ruthless ambition and fierce devotion to family power.
	Jaime	The famed Kingslayer, charming and dangerous yet deeply conflicted by loyalty, duty, and reputation.
	Robb	The honorable heir of Winterfell, forced into command and responsibility by his family's crisis.
	Eddard	The stern and principled Lord of Winterfell, defined by honor, duty, and service as Warden of the North.
	Arya	A fiercely independent Stark daughter who rejects courtly expectations in favor of freedom, survival, and the blade.
	Catelyn	The practical and determined Lady of Winterfell, driven by strong maternal loyalty and political urgency.
	Sansa	The elder Stark daughter, graceful and romantic, whose expectations are challenged by harsh political reality.
	Jon	Eddard's brooding illegitimate son, shaped by questions of identity, belonging, and duty.
	Bran	A curious young Stark whose fall sends him toward an unexpected and consequential path.
ATLA	Aang	The last Airbender and reluctant Avatar, playful by nature yet burdened with restoring balance to a world at war.
	Katara	A compassionate and determined waterbender who grounds the group and consistently stands against injustice.
	Sokka	A witty and inventive warrior whose strategy, humor, and boomerang often prove unexpectedly decisive.
	Zuko	An exiled Fire Nation prince whose pursuit of honor gradually becomes a search for a new identity.

Table 9: Simple background information of characters in our experiments, covering the Fandom Benchmark.

PoPiPa	Kasumi	A bright and starry eyed vocalist and guitarist whose impulsive enthusiasm brings others together and begins the band's journey.
	Tae	A free spirited lead guitarist with strong technique and unusual instincts, often moving at her own pace while strengthening the band's sound.
	Rimi	A shy and gentle bassist who gains courage through performance, offering careful support and warm sincerity to the group.
	Saaya	A caring and dependable drummer with a strong family sense, serving as the band's steady support in music and daily life.
	Arisa	A sharp tongued but reliable keyboardist whose practicality and quick thinking help keep the band organized and grounded.
AG	Ran	A blunt and proud vocalist and guitarist who values authenticity and carries the band's straightforward rock spirit.
	Moca	A relaxed lead guitarist with a teasing side, often hiding keen observation and musical confidence beneath casual behavior.
	Himari	A cheerful bassist and nominal leader who encourages the group while trying to keep everyone connected.
	Tomoe	A dependable drummer with an older sister presence, offering calm strength whenever the band needs stability.
	Tsugumi	A kind keyboardist with practical warmth, often easing tension and supporting the group's everyday balance.
PasuPare	Aya	An earnest vocalist chasing her idol dream through persistence and effort, gradually learning confidence through experience.
	Hina	A cheerful genius guitarist who follows what feels fun, turning bright ideas into action with little hesitation.
	Chisato	A calm and realistic bassist whose professionalism often controls chaos while protecting the group's direction.
	Maya	A technically minded drummer deeply interested in audio equipment, becoming animated when music technology and staging are involved.
	Eve	A sincere keytarist devoted to bushido, whose kindness and wholehearted intensity can inspire or disrupt those around her.
Roselia	Yukina	A determined vocalist pursuing an ideal sound, pushing herself and others with strict focus and uncompromising standards.
	Sayo	A disciplined guitarist who trusts effort over instinct, often expressing care through responsibility and rigorous practice.
	Lisa	A warm and attentive bassist who supports the band emotionally, balancing ambition with empathy and reassurance.
	Ako	An energetic drummer with dramatic flair, bringing confidence and fantasy tinted enthusiasm while seeking recognition.
	Rinko	A quiet and shy keyboardist with outstanding skill, gradually finding courage through support and shared performance.
HHW	Kokoro	A fearless and wealthy optimist who treats making people smile as a mission and turns wild ideas into sincere action.
	Kaoru	A theatrical guitarist who performs a prince like persona, using charm and melodrama to brighten the atmosphere.
	Hagumi	A sunny and athletic bassist who moves with honest excitement, direct energy, and large smiles.
	Kanon	A timid but kind drummer who keeps taking small steps past fear through friends who trust her.
	Misaki	A pragmatic and overworked coordinator who keeps the group functioning, often serving as the realist within cheerful chaos.
Monica	Mashiro	A sensitive vocalist and lyricist troubled by insecurity, gradually learning to express her feelings through song and companionship.
	Touko	A flashy and outgoing lead guitarist who loves attention and momentum, bringing brightness while sometimes causing trouble through impulse.
	Nanami	A highly talented bassist obsessed with being normal, masking inner conflict through humor and flexible adaptation.
	Tsukushi	A hardworking drummer and leader who wants to be dependable, facing clumsiness with persistence and care for the team.
	Rui	A cool violinist and composer who pursues results, slowly confronting the importance of emotion and trust in music.
RAS	CHU^2	A demanding genius DJ and producer who builds the band with control, ambition, and a strong drive toward professional performance.
	LAYER	A charismatic bassist and vocalist whose powerful presence and stable musicianship anchor the band under pressure.
	LOCK	A young and earnest guitarist growing through pressure and mentorship while trying to prove her own value.
	MASKING	A bold and forceful drummer who thrives on adrenaline, volume, and physically intense performance.
	PAREO	A devoted keyboardist with a shy core and polished idol like persona, turning loyalty and effort into support for the band.
MyGO	Tomori	A withdrawn and sensitive vocalist and lyricist who seeks connection through words, transforming pain and longing into songs.
	Anon	A sociable and image aware rhythm guitarist who wants visibility and belonging, learning sincerity as her confident front is tested.
	Raana	A free roaming lead guitarist with mysterious calm, following sound and curiosity while ignoring many social expectations.
	Soyo	A gentle and composed bassist who tries to maintain harmony, caught between care and unresolved emotional pressure.
	Taki	A blunt and intense drummer and composer whose strictness hides protectiveness, expressing concern through sharp honesty and persistence.

Table 10: Simple background information of characters in our experiments, covering the Bandori Benchmark.