Deriving Character Logic from Storyline as Codified Decision Trees

Anonymous ACL submission

Abstract

Role-playing (RP) agents rely on behavioral profiles to act consistently across diverse narrative contexts, yet existing profiles are largely unstructured, non-executable, and weakly validated, leading to brittle agent behavior. We propose **Codified Decision Trees (CDT)**, a data-driven framework that induces an executable and interpretable decision structure from large-scale narrative data. CDT represents behavioral profiles as a tree of conditional rules, where internal nodes correspond to validated scene conditions and leaves encode grounded behavioral statements, enabling deterministic retrieval of context-appropriate rules at execution time. The tree is learned by iteratively inducing candidate scene–action rules, validating them against data, and refining them through hierarchical specialization, yielding profiles that support transparent inspection and principled updates. Across multiple benchmarks, CDT substantially outperforms human-written profiles and prior profile induction methods on 85 characters across 16 artifacts, indicating that codified and validated behavioral representations lead to more reliable agent grounding.

1 Introduction

LLM-based Role-playing (RP) (Chen et al., 2024; Chen et al.), (i.e., building established characters into LLMs) focuses on *"How LLMs can interact in a preferred way"* (Yang, 2024), and exhibits broaden applications spanning from emotional support (Ye et al., 2025), creative writing (Gurung and Lapata, 2025), and gaming engines (Yu et al., 2025). Fundamentally, an RP system takes a scene as the input and outputs an action, aiming to imitate the target character's behavior pattern. Character profile drives it by grounding the LLM with character-specific information. **Codified profile** (Peng and Shang, 2025) converts symbolic rules (e.g., *werewolf transforms into wolf under the full moon*) to executable functions (*state = "wolf" if is_full_moon(scene) else "human"*). Such a structured design provides explicit, executable constraints, leading to better consistency and interpretability than static textual profiles.

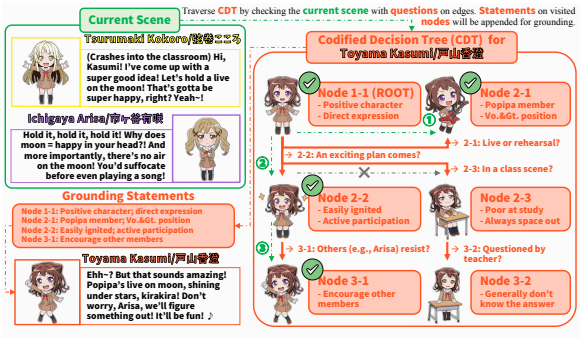

Figure 1: Grounding RP by Codified Decision Tree.

Since manually specified profiles are often unavailable or prohibitively expensive to construct, an alternative is to build character profiles from storylines (profiling) automatically. Existing profiling methods (Wang et al., 2025b) predominantly focus on constructing textual profiles by recurrently or aggregatively merging sub-profiles summarized from blocks of storylines, which often conflates behaviors exhibited under different situational contexts. To address this limitation, we introduce a framework that directly induces codified profiles from storylines, enabling systematic, situation-specific grounding for scene-level action inference. Specifically, inspired by the structure of decision trees (DTs) and human profiling practices, we propose a novel **Codified Decision Tree (CDT)** method. The core idea is to let the LLM hypothesize candidate scene→action triggers from similar (scene, action) pairs in the storyline, and then validate these triggers against the complete set of observed pairs.

Concretely, we construct CDT by inducing a recursively defined tree of behavioral rules from observed (scene, action) pairs. We first cluster semantically similar pairs using text embeddings to sur-

face candidate scene→action regularities. Within each cluster, an LLM proposes codified triggers of the form *"if A then B"* that map interpretable situational predicates A to actions or action modifiers B. These hypothesized triggers are then systematically validated against the entire dataset and used to grow the tree recursively: highly predictive triggers are promoted to internal nodes that partition the data, unsubstantiated triggers are discarded, and partially predictive triggers are refined into child subtrees over filtered subsets of pairs. This construction process yields a hierarchy in which each node stores a compact set of codified grounding statements, and edges correspond to discriminative questions about whether a scene satisfies the associated conditions.

At inference time, a novel scene is routed through the CDT by answering these questions, going through edges whose question return *"True"*. Statements on nodes along the visited path form its situation-specific grounding information, which is then provided to the RP policy for action generation. The resulting structure supports transparent inspection and editing of behavioral rules, principled incorporation of additional data via local subtree updates, and deterministic retrieval of context-appropriate codified profiles for diverse scenes. An example for CDT traversal is presented in Figure 1, where our protagonist *"Toyama Kasumi"* (from *"BanG Dream! Project"*) is facing a scene inviting her to hold a live on the moon. Kasumi's CDT works by following only the edges with questions satisfied by the scene, accumulating statements (easily ignited and vocalist position) from visited nodes and deliberately omitting ones from unvisited alternatives (study-related branches), which demonstrates the situation-aware nature of CDT.

To test the performance of CDT, we first adapt the existing RP benchmark (Peng and Shang, 2025) to a finer-grained one through all action extraction and action sequence modeling. Then, we enrich it by adding new artifacts, curating new styles, and collecting large-scale event story conversations. In total, we have 45 characters, $20,778$ scene-action pairs in the fine-grained Fandom benchmark, 40 characters, $7,866$ pairs in the Bandori benchmark, and an extra $77,182$ pairs of event story conversations from *"BanG Dream!"*. Based on these benchmarks, we compare CDT against alternative ways of leveraging the training data, including model fine-tuning, retrieval-based in-context learning, and prior textual profiling methods, and find that CDT yields substantially stronger action-prediction performance. Notably, CDT even surpasses human-written profiles, highlighting the effectiveness of data-driven profiling when coupled with codification for situation-specific grounding.

We further conduct an in-depth analysis of CDT by examining both its profiling and traversal configurations, including hyperparameter choices, efficiency, clustering mechanisms, and top-k statement selection strategies. In addition, we build a conversion pipeline that translates CDT into a reader-friendly wiki-style textual profile, and the resulting profiles still outperform human-written ones, underscoring the intrinsic quality of CDT-induced knowledge. We also study the data-driven scaling-up of CDT, showing that more data results in stronger CDT profiling. Finally, we present relation modeling as a case study of goal-driven CDT, where the tree is constructed to emphasize a specific behavioral aspect of a character, demonstrating that such targeted profiling can further improve RP precision. Our contributions are three-fold:

- We propose Codified Decision Trees (CDT), a novel data-driven algorithm that induces executable, situation-specific character profiles from storylines for grounded RP inference;
- We introduce enriched benchmarks and conduct comprehensive evaluations, concluding that CDT consistently outperforms other profiling approaches, and even human-written profiles;
- We provide an in-depth analysis of CDT, including configuration setup, training and inference strategies, scaling up, and converting CDT to in-depth textual profiles.

2 Background

Role-playing. Role-playing (RP) tasks aim to sustain coherent and persona-consistent behaviors across evolving narrative or simulated environments (Riedl and Bulitko, 2012; Shao et al., 2023a; Chen et al.). Early RP systems relied on handcrafted character sheets and rules that describe goals, traits, and responses, but these were often limited in scope and lacked systematic validation. Recent advances in large language models (LLMs) have enabled richer, adaptive personas (Yan et al., 2023; Moore Wang et al., 2024; Peng and Shang, 2024), where characters can recall, reason, and act through long-horizon interactions. However, maintaining behavioral stability and transparency remains challenging, because profiles represented as plain text are hard to verify or execute, and also

prone to cause inconsistency. Structured representations such as reasoning graphs, hierarchical memories, and codified constraints (Cheng et al., 2025; Peng and Shang, 2025; Tang et al., 2025) have been proposed to stabilize persona behavior, but these often require extensive manual definition. Our work builds on this direction by inducing interpretable, executable behavioral structures directly from narrative data, enabling grounded, deterministic role-play without handcrafted profiles.

Grounding System. Grounding connects agents' internal reasoning to external narrative or environment states, ensuring that actions and dialogue remain contextually valid (Liang et al., 2022; Wu et al., 2023). In RP and simulation contexts, grounding allows to condition agent responses on evolving world facts, social relations, and temporal cues rather than free-form text history. Prior methods achieve this via world models (Zhang et al., 2024; Liu et al., 2024), graph-based memory (Li et al., 2024), or structured symbolic stores that track entities and events (Sun et al., 2024). These methods improve consistency but often rely on manually crafted schemas or domain-specific rules. In contrast, we automatically construct codified decision trees from large-scale narrative corpora. The induced structure explicitly encodes grounding conditions and decision logic, allowing systematic traversal and verification during execution.

Rule Mining. Rule mining seeks to discover conditional regularities, typically expressed as if–then statements that govern observable behaviors in data (Wang et al., 2024b). Traditional approaches identify statistical or symbolic dependencies between events, while recent LLM-based methods infer and validate such rules through natural language understanding and generation (Gan et al., 2024; Yoneda et al., 2024; Wang et al., 2025a). These rules have been used for planning (Yao et al., 2023; Gao et al., 2023), reasoning (Li et al., 2022), and social simulation (Sun et al., 2024), providing interpretable scaffolds for decision control. However, most existing work treats rule extraction as a flat, unstructured process and offer limited mechanisms for recursive refinement or contextual specialization. Our CDT differs by inducing hierarchical, executable rule structures: candidate triggers are mined, validated, and recursively expanded, forming a transparent tree that encodes both coverage and precision for role-specific behaviors.

3 Codified Decision Tree

3.1 Preliminary

Role-playing system It can be formatted as a function $a = \text{RP}(s|g_x)$ that takes scenes s as inputs and outputs actions a conditioning on a character x's information g_x. As most modern RP systems are driven by LLMs, we replace $\text{RP}(\cdot)$ by $\text{LLM}_{\text{RP}}(\cdot)$ in this paper. The process to incorporate g_x into the action generation is called grounding. A grounding system can be static (e.g., appending x's profile to the context) or dynamic (e.g., retrieving a similar scene-action pair).

Profile Codification Among dynamic systems, codification (Peng and Shang, 2025) is a recently proposed RP grounding paradigm, which converts the static profiles into executable functions $f : s \rightarrow g_x$. For example, a statement in x's profile:

> *"x is always brave against all kinds of challenges."*

will be codified as

```
if check(s, "challenge exists?") return "brave"
```

where `check(`s`,` q`)` is a discriminator calling that returns `True/False/None` (`None` represents "Unknown") based on the scene s and question q. The discrimination can be executed by the RP LLM itself or a fine-tuned model. Such a codification process improves the RP LLM's consistency with the profile logic by enforcing reasoning paths rather than leaving it for LLMs to explain. A limitation of the existing codification framework is the reliance on human-written profiles for conversion. This paper explores a different data-driven codification, which takes a set of scene-action pairs $\mathcal{D} = \{(s_i, a_i)\}_{i=1:|\mathcal{D}|}$ as input to directly build the codified function f without given profiles.

3.2 Data Structure

Similar to traditional decision trees, CDTs take a textual scene s as the input and enable moving between nodes v based on discrimination, which is implemented as the mentioned `check(`s`,` q`)` function. The difference is that CDTs do not output a single label like traditional decision trees. As a grounding system, CDTs will instead be traversed, and all grounding information inside visited nodes will be incorporated for grounding. As shown in Figure 1, a node in CDT includes two key elements: 1) a set H of statements h; 2) a set of child nodes with questions for traversal checking.

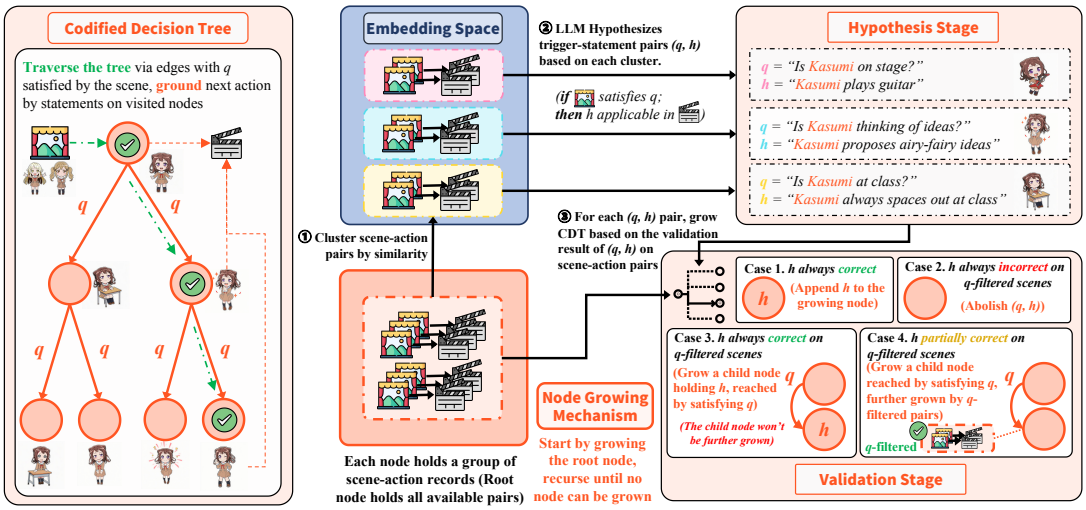

Figure 2: The workflow of codified decision tree (CDT).

CDT Traversal (Shown in Figure 1, with another running case available in Figure 7) starts from the root node with an input scene s, and we first append all statements inside the root node to the grounding set g, then for each child node, we check whether `check(`s, q`)` is `True` for the assigned question. If the check is passed, we visit the corresponding child node and recursively execute the flow: 1) appending statements; 2) checking to visit child nodes operations until all reachable nodes are visited. Follow the proposed structure, each node in CDT can be understood as the triggered behavior statements under conditions filtered by multiple questions q.

3.3 Recursive Hypothesis-Validation

Based on CDT's definition above, we design the methodology to grow a CDT (Figure 2). For initialization, we have a root node with no statements or children, but the whole training set of scene-action pairs $\mathcal{D}_{train}$. Then we apply a rule mining system (elaborated in § 3.4) to hypothesize the causality between scenes and actions (hypothesizing case available in Figure 6). Each hypothesis is in the format (q, h) where q is the filtering question and h is the behavioral statement.

For each (s, a) inside $\mathcal{D}_{train}$, we first check whether h is a global (q-free) behavior statement according to the h-a natural language inference (NLI) relation. By NLI discrimination on all data, we get the ratio of entailed r_e, neutral r_n, and contradicted r_c. If the accuracy $\frac{r_e}{r_e+r_c}$ reaches an accept threshold θ_{acc}, h will be directly added to the statement set H as a globally applicable behavior.

For (q, h) failed to be established as a global statement, we then filter $\mathcal{D}$ into a subset $\mathcal{D}'$ of (s, a) pairs that s passes the `check(`s, q`)`. Based on the accuracy, an accept threshold θ_{acc}, and a reject threshold θ_{rej}, we select the next operation step.

- If accuracy $> \theta_{\text{acc}}$, we add a leaf node with only h in H, which can be accessed from the parent by `check(`s, q`)` and won't be further grown, representing a critical hit.
- If accuracy $< \theta_{\text{rej}}$, (q, h) is abolished with no further operation, representing a failed hypothesis.
- If accuracy falls between θ_{rej} and θ_{acc}, we view it as a logic that requires further exploration for more complex structures. Thus, we will build a new empty node connected to its parent by `check(`s, q`)`, the new node will take the filtered dataset $\mathcal{D}'$ for recursive growth. To avoid recursion lasting too long, we set a filtering threshold θ_f so that the recursion is only triggered when $\frac{|\mathcal{D}'|}{|\mathcal{D}|} < \theta_f$, which gradually reduces $|\mathcal{D}|$ during recursion. The recursion can also be stopped by a too small $|\mathcal{D}|$ or too deep v in the CDT.

By such a recursive hypothesis-validation mechanism, we can model the logic in different complexities to ensure broad and deep coverage.

3.4 Rule Mining

In the recursive hypothesis-validation mechanism (§ 3.3), the rule mining system is a crucial component to make high-potential hypotheses for (q, h). As we target a more precise and efficient mining of (s, a) pairs that represent the causality, we apply a clustering algorithm (K-Means) based on the textual embeddings of s and a. For a, we apply a semantic textual embedding (Gao et al., 2021). For s, as we care more about the potential triggered character behavior, we follow the idea of instruction-following embedding (Peng et al., 2024) to use the

Scene	Haruhi is unworried, pointing out the North High uniform is not distinctive enough to identify at a glance
	Haruhi reveals her concept for the film is for the timid Asahina to go through great struggles and suffering to make the happy ending more satisfying
	Koizumi asks if there will be any actors in the film other than Asahina, Nagato, and himself
Question	What'll be Haruhi's next action in response to the current scene?
Action	Haruhi is inspired to recruit people to play Nagato's minions

Table 1: Examples of training/test cases. (10 preceding actions as the scene in real benchmarks)

last token prediction hidden state from a generative LM as the embedding following the prompt below:

"{scene} Thus, {character} decides to"

where the hidden states predicted from *"to"* will contain the distribution of all kinds of verbs triggered by s to guide the clustering. In contrast, a normal semantic embedding will be distracted by surface similarity, especially by clustering scenes from the same episode together, ignoring their difference in impact on the target character. Based on clustering the concatenated embedding between scenes and actions, we prompt LLM to summarize potential rules from each cluster.

Diversification We let parent nodes pass their question paths (the sequence of questions to reach the node) and established statements (the validated statements when reaching the node) to children. Such information is incorporated into the hypothesis prompt to instruct the LLM to propose something else to avoid redundant checking.

4 Benchmark

While early RP works rely on synthesized datasets (Shao et al., 2023b; Wang et al., 2024c), the RP community has also begun to use humans' annotations (e.g., profiles and synopses) (Ran et al., 2025; Wang et al., 2025b) on well-known artifacts for benchmarking. Fandom Benchmark (Peng and Shang, 2025) is such an RP benchmark, which utilizes Fandom[1]'s rich human annotations for both profiles and synopses. Fandom Benchmark extracts **only key actions** of characters in scenes to evaluate whether RP systems can take a given scene to make an action entailed by the ground-truth original actions in NLI. We extend such a benchmarking strategy to **all actions** in the gathered artifacts.

Fine-grained Fandom Benchmark We use the same source of story synopses crawled from Fandom as the original benchmark. Instead of extracting only key actions in the original benchmark, we utilize an LLM to break narrations into sequences of actions annotated with acting characters (*"environment"* when no active character). The preceding 10 actions of each action are taken as the input scene. For each character, we perform a chronological split of scene–action pairs, training on the first half of the storyline and evaluating on the second half, which both prevents data contamination and mirrors real-world prediction of future behavior from past evidence. In addition to the original 6 artifacts (*"Haruhi"*, *"K-On!"*, *"Fullmetal Alchemist"*, *"JOJO"*, *"A Game of Thrones"*, *"Avatar: The Last Airbender"*), we incorporate *"Death Note"* and *"Spy × Family"* to broaden the benchmarking scope. We experiment on the main characters that appear throughout the storyline for long-horizon testing.

Bandori Conversational Benchmark We curate the Bandori conversational benchmark for assessing dialogue-level role-playing behavior. We collect conversations from the first band story of all eight bands (from *"Poppin'Party"* to *"MyGO!!!!!"*)[2], where each utterance is treated as an action. Artifact and character background information is provided in Appendix H.

Criterion and Statistics For both benchmarks, predicted next actions are compared against reference actions using reference-prediction NLI relation (*"entailed"* for 100, *"neutral"* for 50, *"contradicted"* for 0), following prior practice (Peng and Shang, 2025). Overall, the fine-grained Fandom benchmark contains 45 characters and 20,778 scene-action pairs, while the Bandori benchmark comprises 40 characters and 7,866 pairs. We further collect *"BanG Dream!"* event-story conversations, yielding extra 77,182 pairs for scaling-up analysis. Statistics are provided in Appendix A. Our main content focuses on the NLI score for next action prediction, while multi-dimensional scoring, out-of-domain scenario, and human evaluation can be found in Appendix D, which validates NLI to be cross-metric and human consistent.

5 Experiment

5.1 Evaluation and Baselines

We include comprehensive baseline methods to cover all types of existing methods that utilize known plots to ground RP in new scenes.

[1] https://www.fandom.com/

[2] e.g., bandori.fandom.com/wiki//Poppin'Party/Band_Story/

Fandom		Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA	Avg.
Data-driven	Vanilla	55.08	49.92	56.10	62.49	55.66	54.66	57.05	53.56	55.57
	Fine-tuning	51.49	51.01	49.14	50.79	44.07	34.92	41.20	42.84	45.68
	RICL	56.83	55.74	56.86	62.80	56.33	49.77	56.46	52.25	56.01
	ETA	60.54	53.83	58.00	63.29	57.12	51.00	55.28	56.23	56.91
	CDT (Ours)	61.16	57.93	60.35	66.34	58.57	57.40	63.79	61.05	60.82
	CDT-Lite (Ours)	62.17	57.24	59.79	67.00	59.04	57.26	64.27	61.32	**61.01**
Human	Human Profile	55.87	55.86	59.14	64.75	58.54	55.11	59.35	57.98	58.33
	Codified Human Profile	57.94	55.93	59.38	65.56	57.01	56.56	62.07	59.97	59.30
Bandori		PoPiPa	AG	PasuPare	Roselia	HHW	Monica	RAS	MyGO	Avg.
Data-driven	Vanilla	66.39	66.76	68.29	66.83	65.13	64.06	67.20	59.37	65.50
	Fine-tuning	69.52	62.76	64.63	61.83	62.39	62.35	62.64	56.72	62.86
	RICL	73.56	67.56	73.06	67.24	73.63	65.04	69.44	61.10	68.86
	ETA	75.29	72.49	78.00	70.91	78.92	66.82	72.68	62.89	72.25
	CDT (Ours)	84.25	79.92	78.93	71.93	80.03	77.33	80.08	69.17	77.71
	CDT-Lite (Ours)	88.38	80.49	82.47	72.81	79.66	78.67	79.51	70.33	**79.04**
Human	Human Profile	73.73	72.43	77.11	70.08	73.14	68.08	71.74	63.91	71.28
	Codified Human Profile	73.02	74.00	78.65	71.23	72.47	69.14	71.41	65.02	71.87

Table 2: RP performance comparison (NLI score) on fine-grained Fandom and Bandori benchmarks.

- **Vanilla** directly prompts the RP model with the input scene, without any additional information.
- **Fine-tuning** adapts the RP model by supervised training on the character-specific scene–action pairs, enabling implicit memorization of behavioral patterns from past storylines.
- **Retrieval-based In-Context Learning (RICL)** (Wang et al., 2024a) retrieves a set of scene–action examples from the training data that are most similar to the input scene as in-context examples to guide action generation.
- **Extract-Then-Aggregate (ETA)** (Wang et al., 2025b) first extracts textual sub-profiles from blocks of storylines and then aggregates them into a single character profile, which is appended to the prompt as grounding information.
- **Human Profile** grounds RP using human-written character descriptions, containing natural language canonical traits and behaviors. Our experiment applies the ones written in Fandom wiki.
- **Codified Human Profile** (Peng and Shang, 2025) converts human-specified symbolic rules into executable grounding functions that condition actions on scene-specific predicates.

5.2 Implementation Details

Hyperparameters Detailed hyperparameter setups are placed in Appendix B. Hyperparameter impact analysis can be found in Appendix C.

Clustering We embed actions by using `qwen3-embedding-8b` (Zhang et al., 2025). Scenes are embedded using instruction-following representations produced by the generative model `qwen3-8b` (Yang et al., 2025).

Hypothesis-Validation We use `gpt-4.1` to hypothesize candidate scene→action triggers in natural-language "if–then" form. Trigger validation is conducted with `gpt-4.1-mini` via NLI-style judgments over all scene–action pairs. We also introduce a lightweight variant CDT-Lite, by replacing `gpt-4.1-mini` with a 0.1B encoder model (`deberta-v3-base` (He et al., 2021)) distilled from 1% of `gpt-4.1-mini` discrimination.

RP Model `llama-3.1-8b-instruct` is used as the RP model to generate character actions and responses in main experiments. It is also used to answer discriminative questions during CDT traversal. When using CDT-Lite, question discrimination is instead performed by `deberta-v3-base`.

5.3 Main Results

As shown in Table 2, across all 8 Fandom artifacts, CDT(-Lite) achieve the best scores, consistently surpassing Vanilla prompting, fine-tuning, RICL, and ETA. Fine-tuning often underperforms even the Vanilla baseline, which can be attributed to the strict chronological split of training and test sets, as fitting tokens in the first half storyline leads to repeating similar actions rather than learning global behavior pattern. RICL and ETA provide noticeable gains over Vanilla in several settings, confirming the value of grounding on past storylines, but they remain clearly behind CDT(-Lite), which benefit from structured, situation-specific codification rather than purely textual aggregation.

CDT also outperforms the "Human Profile" and "Codified Human Profile" baselines, which use human-written profiles as commonly accepted ground truth. While these human profiles are

	Fandom	Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA	Avg.
CDT-Lite ($d_{\max} = 4$)		62.17	57.24	59.79	67.00	59.04	57.26	64.27	61.32	61.01
Ablation	w/ Abolished Statements	61.18	55.09	58.85	66.21	58.01	56.24	61.29	59.33	59.52
Ablation	$d_{\max} = 1$	60.99	55.75	60.88	66.29	58.03	54.56	61.03	60.43	59.65
Ablation	$d_{\max} = 2$	61.03	56.61	60.83	68.59	58.82	55.61	62.13	60.80	60.55
Ablation	$d_{\max} = 3$	61.53	56.50	59.68	67.91	57.87	57.15	63.30	61.58	60.69
Ablation	w/o Clustering	59.76	55.14	58.29	66.39	59.52	55.32	60.25	60.88	59.44
Ablation	w/o Inst. Embed.	58.87	57.19	58.15	65.51	59.23	56.65	61.21	61.06	59.73
Ablation	w/o Diversification	60.75	57.04	59.96	66.25	57.84	57.00	62.89	59.83	60.20
Variant	TopK (Depth Rank)	59.51	56.24	58.25	64.82	57.26	54.26	61.00	60.59	58.99
Variant	TopK (Accuracy Rank)	61.40	56.46	58.90	66.87	58.95	56.14	64.49	61.04	60.53
Variant	TopK (Usability Rank)	62.21	56.11	59.27	67.78	58.72	57.97	63.95	61.85	60.98
Variant	Verbalized CDT	62.17	57.04	57.93	65.32	58.54	55.01	61.60	60.98	59.82
Variant	Wikified CDT	59.21	54.93	56.77	67.76	57.11	54.28	59.15	57.95	58.40

Table 3: Ablation study and variant experiments. (Bandori part placed in Table 7)

strong and generally superior to unguided or purely data-hungry approaches, CDT(-Lite) obtain higher scores for every Fandom artifact, demonstrating that data-driven profiling under the codification framework can exceed manually constructed descriptions in both coverage and situational fidelity.

On the Bandori benchmark, we observe an even larger margin: CDT(-Lite) achieve the top performance for all 8 bands, with CDT-Lite slightly ahead of CDT in most cases. Traditional grounding methods (fine-tuning, RICL, ETA) improve over Vanilla but still lag behind our codified approach, and even the best human-written or codified human profiles are outperformed by CDT(-Lite). These results indicate that 1) codified, situation-aware profiling is particularly beneficial for conversational RP, and 2) the lightweight CDT-Lite architecture retains most of the benefits of CDT, often improving further, while enabling cheaper validation, making data-driven codified profiling practical at scale.

6 Analyses

Because of the page limitation, we place hyperparameter analysis in Appendix C, efficiency analysis in Appendix E, and running cases in Appendix F.

6.1 Ablation Study

With abolished statements This ablation verifies the importance of the statement validation stage by simply appending the abolished statements back to the CDT. Comparing with CDT-Lite in Table 3 shows that blindly accepting LLM hypotheses is consistently worse, highlighting the importance of explicit validation in improving reliability.

Node Depth We next ablate the maximum traversal depth $d_{\max}$, which controls how far a scene is allowed to descend in the tree. The case $d_{\max} = 1$ degenerates CDT into a flat, codified profile with only root-level rules. Ablation result shows that RP performance generally improves when deepening the CDT, but it also observes a gradual saturation.

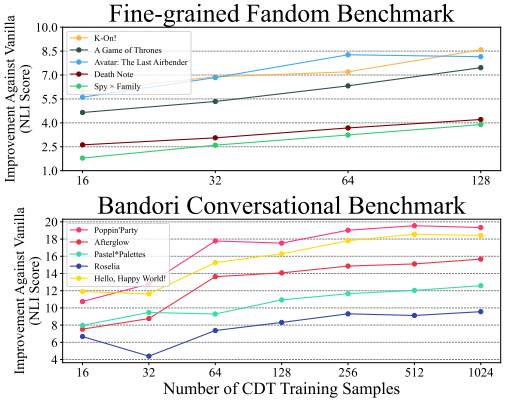

Figure 3: Performance scales up with training data.

Clustering Mechanism To study the role of clustering, we consider two variants: (1) **w/o Clustering** disables clustering altogether and hypothesize triggers on the entire set of scene–action pairs at each node, forcing the LLM to explain a more heterogeneous mixture of behaviors; (2) **w/o Inst. Embed.** retains clustering but replace instruction-following action embeddings with simpler representations, reducing the semantic alignment between scenes and actions. Both lead to consistent drops compared to CDT-Lite (Table 3), indicating that (i) grouping semantically similar instances before rule induction and (ii) using instruction-following embeddings to represent actions are both crucial for discovering clean, reusable triggers.

Diversification As shown in Table 3, turning off diversification also results in a drop of CDT performance, which validates its contribution in probing more potential candidate behaviors.

6.2 Variants

Top-K Policy During inference, CDT may activate multiple codified statements along a traver-

sal path, requiring a selection policy. We evaluate three **Top-K** ranking strategies: **Depth Rank**, which favors deeper (more specific) nodes; **Accuracy Rank**, which orders statements by validated accuracy $\frac{r_e}{r_e+r_c}$; and **Usability Rank**, which prioritizes empirically applicable rules $\frac{r_e}{r_e+r_n+r_c}$. As shown in Table 3, Usability Rank generally yields the strongest results, suggesting that broadly applicable rules provide better grounding for RP.

CDT-to-Wiki We also explore using CDTs as textual profiles. **Verbalized CDT** linearizes the tree into explicit if-then rules and appends them to the prompt, preserving conditional structure, while **Wikified CDT** further rewrites these rules into a wiki-style narrative (pipeline described in Appendix F), removing most control-flow cues. Both variants eliminate runtime traversal and can be used by standard RP models. As shown in Table 3, they remain strong baselines and outperform human-written profiles, despite sacrificing fine-grained situational control and performance.

6.3 Scaling-up by Data

We study how CDT scales with additional supervision by varying the number of training scene–action pairs. Figure 3 reports CDT performance under different training sizes. For the Fandom benchmark, we select characters with at least 128 actions and randomly sample $\{16, 32, 64, 128\}$ pairs per character. For the Bandori benchmark, using the richer event-story data, we further scale training up to 1024 pairs for five bands.

Across both benchmarks, CDT exhibits a clear scaling trend: performance steadily improves as more training data are used. On Fandom, CDT trained with only 64 pairs already surpasses human-written profile baselines, showing that limited behavioral evidence can outperform manually authored descriptions. On Bandori, gains continue beyond 128 and up to 1024 pairs, indicating that RP performance benefits from scaling up the number of interaction data.

6.4 Relation Modeling

Beyond general profiling, CDT also supports *goal-driven* specialization toward specific behavioral aspects by **instructing the trigger-hypothesis prompt**. We demonstrate this capability with **relation modeling**, which aims to capture characteristic interaction patterns between pairs of characters. For each target character, we filter the training data to retain only actions that occur immediately after a designated related character speaks or acts, and train a relation-specific *goal-driven* CDT that is instructed to focus on their special interaction patterns. At inference time, scenes are routed through this relation-specific CDT whenever the latest context action is taken by the related character; otherwise, the general CDT is used. As shown in Table 4, this +GD variant consistently outperforms the base CDT on the target subset (cases when interacting with the related character) across all nine character pairs and thus results in gains on the full test set, demonstrating that CDT can model fine-grained relational behaviors without sacrificing generality.

Character	Relation with	Target Subset		Full Test	
		CDT	+GD	CDT	+GD
Suzumiya Haruhi	Kyon	56.45	**67.74**	67.91	**70.14**
Akiyama Mio	Tainaka Ritsu	44.93	**50.00**	52.63	**54.30**
Yor Forger	Loid Forger	64.13	**70.65**	60.05	**60.76**
Alphonse Elric	Edward Elric	75.00	**81.82**	63.43	**64.82**
Toyama Kasumi	Ichigaya Arisa	85.95	**88.02**	85.78	**86.44**
Mitake Ran	Aoba Moca	81.25	**85.00**	85.12	**86.68**
Hikawa Hina	Hikawa Sayo	85.05	**88.92**	86.84	**87.69**
Hikawa Sayo	Hikawa Hina	76.52	**86.87**	76.15	**80.39**
Chihaya Anon	Takamatsu Tomori	69.33	**72.00**	82.18	**83.06**

Table 4: Performance of Goal-driven (GD) CDT.

6.5 Case Study

Figure 8 contrasts a human-written character profile with a Wikified CDT for *"Haruhi Suzumiya"*, highlighting differences in granularity and behavioral coverage. While the human profile offers a compact, intuitive summary of key traits, the Wikified CDT expands these traits into systematically grounded patterns that capture a wider range of recurring behaviors and situational triggers. This comparison illustrates how CDT provides more comprehensive behavioral grounding than manually authored profiles.

7 Conclusion and Future Work

We present *Codified Decision Trees* (CDT), a data-driven framework that induces executable, situation-specific character profiles from storylines. We show that CDT consistently outperforms previous methods, even using human-written profiles. As future work, we plan to extend CDT to multi-character joint profiling and interaction modeling, continual and online updates from interactive RP logs, and multi-modal settings where scenes include visual or game-state signals, moving toward more broadly controllable, behavior-grounded agents beyond the RP domain.

Limitations

While CDT demonstrates strong performance and interpretability, it currently relies only on storyline-derived scene-action pairs and does not incorporate pre-designed persona information (e.g., canonical traits or author intent). We adopt this setting to isolate the effect of data-driven profiling and to avoid introducing external human priors that may be inconsistent or unavailable across domains; incorporating such information as soft priors for trigger induction is a natural extension. CDT is also constructed offline and remains static, which simplifies validation and benchmarking but does not capture evolving narratives; enabling incremental or continual CDT updates as new storylines unfold is an important direction for future work. Finally, although we focus on role-playing, the CDT framework is general and could be extended to other situation-aware behavior modeling tasks, such as task-oriented agents or embodied decision-making.

References

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

Xiangyu Robin Gan, Yuxin Ray Song, Nick Walker, and Maya Cakmak. 2024. Can large language models help developers with robotic finite state machine modification? *arXiv preprint arXiv:2412.05625*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Alexander Gurung and Mirella Lapata. 2025. Learning to reason for long-form story generation. *CoRR*, abs/2503.22828.

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*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, et al. 2024. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12758–12786.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*.

Zeyang Liu, Xinrui Yang, Shiguang Sun, Long Qian, Lipeng Wan, Xingyu Chen, and Xuguang Lan. 2024. Grounded answers for multi-agent decision-making problem through generative world model. *Advances in Neural Information Processing Systems*, 37:46622–46652.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024. Rolellm: benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Letian Peng and Jingbo Shang. 2024. Quantifying and optimizing global faithfulness in persona-driven role-playing. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. 2024. Answer is all you need: Instruction-following text embedding via answering the question. 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 459–477. Association for Computational Linguistics.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. BOOKWORLD: from novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15898–15912. Association for Computational Linguistics.

Mark Riedl and Vadim Bulitko. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 2160–2165.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023a. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023b. 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.

Libo Sun, Siyuan Wang, Xuanjing Huang, and Zhongyu Wei. 2024. Identity-driven hierarchical role-playing agents. *arXiv preprint arXiv:2407.19412*.

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.

Xiaochen Wang, Junqing He, Liang Chen, Gholamreza Haffari, Yiru Wang, Zhe Yang, Xiangdi Meng, Kunhao Pan, and Zhifang Sui. 2025a. Sg-fsm: A self-guiding zero-shot prompting paradigm for multi-hop question answering based on finite state machine. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6025–6037.

Xiaochen Wang, Junqing He, Yiru Wang, Xiangdi Meng, Kunhao Pan, Zhifang Sui, et al. 2024b. Fsm: A finite state machine based zero-shot prompting paradigm for multi-hop question answering. *arXiv preprint arXiv:2407.02964*.

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. 2024c. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1840–1873. Association for Computational Linguistics.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102.

Ming Yan, Ruihao Li, Hao Zhang, Hao Wang, Zhilan Yang, and Ji Yan. 2023. Larp: Language-agent role play for open-world games. *arXiv preprint arXiv:2312.17653*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Diyi Yang. 2024. Human-ai interaction in the age of large language models. In *Proceedings of the AAAI 2024 Spring Symposium Series, Stanford, CA, USA, March 25-27, 2024*, pages 66–67. AAAI Press.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023.

React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4646–4669. Association for Computational Linguistics.

Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R Walter. 2024. Statler: State-maintaining language models for embodied reasoning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15083–15091. IEEE.

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

Lin Zhang, Zihan Liu, Yuchen Zhou, Tong Wu, and Jintao Sun. 2024. Grounding large language models in real-world environments using imperfect world models. *IJACSA) International Journal of Advanced Computer Science and Applications*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

A Statistics Information

Tables 5 and 6 summarize the statistics of the benchmarks and the resulting CDTs used in our experiments. The fine-grained Fandom benchmark spans eight artifacts with diverse narrative lengths and cast sizes, ranging from relatively compact series (e.g., *Haruhi*) to long-running storylines with large ensembles (e.g., *AGOT* and *ATLA*). The Bandori conversational benchmark covers all eight bands with balanced main-character counts, while the event-story extension substantially increases data scale, yielding over 77K actions in total. Across datasets, average action lengths remain stable, indicating consistent annotation granularity.

Table 6 reports structural statistics of the induced CDTs. Despite large variation in data volume, the resulting trees remain shallow and compact, with moderate numbers of nodes and statements per character, reflecting the effectiveness of validation and pruning. Average statement lengths are consistent across artifacts and benchmarks, suggesting that CDT induces comparable, interpretable rule granularity in different narrative domains. These statistics further support that CDT scales to long storylines without uncontrolled growth in profile complexity.

B Setup Details

In our experiments, we set $\theta_{\text{accept}} = 0.75$, $\theta_{\text{reject}} = 0.50$, $\theta_{\text{f}} = 0.75$, max depth $= 4$, $|\mathcal{D}|_{min}$ for node growing $= 16$, 3 hypotheses per cluster, minimal 16 cases per cluster, and a maximum of 8 clusters per node for all main experiments. For distillation in CDT-Lite, we train the `deberta-v3-base` discriminator with AdamW (Loshchilov and Hutter, 2019) initialized by 10^{-5} learning rate, 32 batch size for 1 epoch.

For baselines, **Fine-tuning** adopts LoRA (Hu et al., 2022) fine-tuning with batch size 16, AdamW initialized by 2×10^{-4}, which enables smooth loss drop. We report the result after the first epoch since the performance continues dropping while fitting on the first half storyline. **RICL** retrieves the scene with the 8 most similar scene based on `qwen3-embedding-8b`. **ETA** segment scene-action pairs by 16 samples in each block, with prompts available in Appendix 13. **Human profile and codified human profile**'s setups follow the previous codification work (Peng and Shang, 2025) using the same distilled discriminator for CDT-Lite.

C More Variant Studies

The full ablation and variant study results are placed in Table 7, which reaches a similar conclusion about component contribution, top statement selection strategy, and verbalization effect.

RP Model Variants Table 8 reports results when varying the RP backbone while keeping CDT-Lite unchanged. We evaluate smaller RP models (1B and 3B) and compare CDT-Lite against Vanilla prompting and human-written Human Profiles. Across both the Fandom and Bandori benchmarks, CDT-Lite consistently delivers large gains over Vanilla and Human Profile for all model sizes, demonstrating that codified, situation-aware grounding is effective even with limited-capacity RP models. While absolute performance improves with stronger RP backbones, the relative advantage of CDT-Lite remains stable, indicating that its benefits are largely orthogonal to model scale. Notably, CDT-Lite with smaller RP models can approach or even surpass the performance of larger models grounded with human-written profiles, highlighting its practical value in resource-constrained settings.

Codifier Variants Table 11 compares different codifiers used for trigger hypothesis and validation when constructing CDT-Lite. Using `gpt-4.1` as the codifier generally yields the strongest performance, while `gpt-4.1-mini` or `qwen3-coder` achieves slightly lower but still competitive results on both benchmarks. The relatively small gap between these settings suggests that CDT-Lite is robust to codifier strength and can be instantiated with lighter or open-source models at reduced cost. These results further support the scalability of CDT, showing that high-quality codified profiles can be induced without relying exclusively on the strongest available LLMs.

θ_{accept}		0.70	0.75	0.80	0.85
θ_{reject}	0.40	69.28	69.74	69.35	68.82
	0.50	69.13	70.16	70.53	69.54
	0.60	66.05	67.82	69.33	68.89
#Hypothesis		1	2	3	4
		68.53	69.46	70.16	70.33
#Max Cluster		4	8	12	16
		68.92	70.16	70.22	70.43

Table 10: CDT performance with different hyperparameter setups.

Hyperparameter Setup Table 10 reports the sensitivity of CDT performance to different hyper-

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
#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
#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
Bandori (Events)	**PoPiPa**	**AG**	**PasuPare**	**Roselia**	**HHW**	**Monica**	**RAS**	**MyGO**
#Main Character	5	5	5	5	5	5	5	5
#Episode				1498				
#Action				77182				
#Action$_{\text{Main Character}}$	12553	11365	12058	11821	10287	4758	2863	745
#Avg. Action Length	19.63	19.87	21.61	21.07	20.82	20.73	19.53	15.90

Table 5: Statistics of benchmarks (*Fine-grained Fandom Benchmark* and *Bandori Conversational Benchmark*) used in the experiments.

Fandom	Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA
#Node	3.80	32.80	162.33	45.80	73.40	32.14	79.73	309.50
#Statement	16.80	90.80	259.67	135.20	88.20	50.14	159.55	497.25
#Avg. Statement Length	18.03	19.75	20.46	19.63	19.46	18.49	20.56	20.20
Bandori	**PoPiPa**	**AG**	**PasuPare**	**Roselia**	**HHW**	**Monica**	**RAS**	**MyGO**
#Node	10.40	28.40	41.40	21.80	5.20	52.00	59.00	9.80
#Statement	61.00	107.40	99.40	68.20	27.40	112.00	114.60	82.40
#Avg. Statement Length	18.35	19.23	19.74	20.62	18.23	18.37	19.05	18.66

Table 6: Statistics of CDTs used in the experiments.

parameter configurations, evaluated on a sampled 10% subset of the test set. We vary the trigger acceptance threshold θ_{accept}, rejection threshold θ_{reject}, the number of hypotheses proposed per cluster, and the maximum number of clusters expanded at each node. Overall, performance is relatively stable across a wide range of settings, indicating that CDT is not overly sensitive to precise hyperparameter tuning. We apply $\theta_{\text{accept}} = 0.75$, $\theta_{\text{reject}} = 0.50$, 3 hypotheses per cluster, and a maximum of 8 clusters per node for all main experiments, which achieves a strong balance between performance and computational cost.

	Bandori	PoPiPa	AG	PasuPare	Roselia	HHW	Monica	RAS	MyGO	Avg.
CDT-Lite		88.38	80.49	82.47	72.81	79.66	78.67	79.51	70.33	79.04
Ablation	w/ Abolished Statements	86.90	76.35	78.69	73.52	78.52	77.44	79.54	67.90	77.36
	$d_{\text{max}} = 1$	87.15	77.23	82.19	74.07	77.88	77.96	76.97	69.77	77.90
	$d_{\text{max}} = 2$	87.40	79.79	82.85	72.71	79.49	78.44	79.40	70.04	78.77
	$d_{\text{max}} = 3$	88.24	79.84	81.76	72.90	79.63	78.97	79.06	70.75	78.89
	w/o Clustering	86.99	79.38	80.67	73.22	80.52	78.21	77.43	69.70	78.27
	w/o Inst. Embed.	86.97	77.66	78.78	75.91	78.16	79.16	78.93	67.00	77.82
	w/o Diversification	88.74	80.24	80.97	73.38	79.13	78.26	78.10	69.97	78.60
Variant	TopK (Depth Rank)	87.58	76.78	79.57	71.77	78.93	75.77	82.24	68.86	77.69
	TopK (Accuracy Rank)	87.73	81.26	81.62	72.65	79.82	76.19	80.55	69.85	78.71
	TopK (Usability Rank)	88.17	81.28	83.76	75.19	78.92	78.84	77.72	70.90	79.35
	Verbalized CDT	88.57	79.92	81.32	75.18	78.39	78.01	78.10	70.12	78.70
	Wikified CDT	86.56	75.92	82.56	73.24	81.21	77.14	78.20	64.35	77.40

Table 7: Bandori part of ablation study and variant experiments.

Fandom	Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA
Vanilla	43.60	44.20	47.16	51.47	43.71	41.45	42.56	43.18
Human Profile	42.70	49.31	53.98	54.26	47.97	49.42	49.14	45.13
CDT-Lite (Ours)	54.30	53.88	58.61	60.07	49.82	52.49	50.92	51.92
Vanilla	50.47	49.73	51.67	56.97	55.74	49.57	49.46	52.77
Human Profile	50.58	52.87	55.31	60.45	53.12	50.96	52.38	53.12
CDT-Lite (Ours)	54.75	53.75	56.00	62.57	53.16	49.74	53.14	54.45
Human Profile (8B)	55.87	55.86	59.14	64.75	58.54	55.11	59.35	57.98

Bandori	PoPiPa	AG	PasuPare	Roselia	HHW	Monica	RAS	MyGO
Vanilla	58.38	56.44	60.14	56.57	58.50	56.06	59.97	53.60
Human Profile	65.34	62.90	66.75	61.03	65.76	55.56	64.73	54.75
CDT-Lite (Ours)	78.58	80.05	77.47	68.14	78.90	73.42	73.68	61.16
Vanilla	63.91	62.27	66.49	60.90	64.63	63.06	62.95	59.62
Human Profile	71.96	68.13	74.00	67.64	70.64	65.16	69.27	59.39
CDT-Lite (Ours)	79.13	77.72	77.29	67.46	79.26	74.18	74.30	66.51
Human Profile (8B)	73.73	72.43	77.11	70.08	73.14	68.08	71.74	63.91

Table 8: RP performance on variant RP models (1B and 3B), compared with key baselines.

Fandom	Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA
`gpt-4.1-mini`	60.45	58.66	59.07	67.27	59.19	55.35	62.06	60.95
`gpt-4.1`	62.17	57.24	59.79	67.00	59.04	57.26	64.27	61.32
`qwen3-coder`	61.37	57.05	58.83	66.39	56.60	56.88	61.07	60.14

Bandori	PoPiPa	AG	PasuPare	Roselia	HHW	Monica	RAS	MyGO
`gpt-4.1-mini`	85.62	80.24	79.26	73.66	79.63	77.80	77.49	73.10
`gpt-4.1`	88.38	80.49	82.47	72.81	79.66	78.67	79.51	70.33
`qwen3-coder`	82.95	81.13	79.83	74.06	80.46	77.37	78.57	71.09

Table 9: RP performance on variant codifiers, compared with key baselines.

D Evaluation Generality Validation

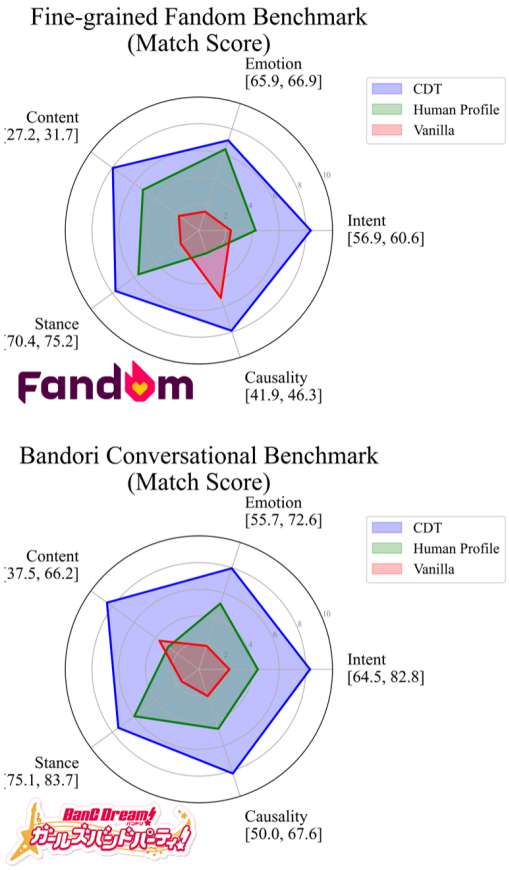

Figure 4: The comparison between grounding methods with matching scores. (CDT represents CDT-Lite)

Cross-metric and Manual Validation Figure 4 further shows that the NLI-based matching score is consistent with prior LLM-based multi-dimensional judgment protocols, validating its use as a reliable automatic evaluator. Across both benchmarks, CDT exhibits clear and systematic gains on all evaluated dimensions, *intent*, *emotion*, *content*, *stance*, and *causality*, with particularly strong improvements on intent and causality, reflecting CDT's ability to condition behavior on situation-specific triggers rather than generic traits. Improvements in content and stance indicate that CDT grounding helps the RP model select contextually appropriate details and attitudes, while emotion gains suggest better alignment with character affect under varying scenes. We also verify that the trends observed in NLI scores closely match those of manual inspections: based on 200 samples for each class, the NLI discrimination achieves 90.50%, 92.00%, and 88.50% consistency with human judgment on *"entailed"*, *"neutral"*, and *"contradicted"* while matching score has 89.00% and 90.50% consistency on *"matched"* and *"mismatched"*. This confirms that both NLI accuracy and the aggregated matching score serve as dependable proxies for RP quality.

Open-domain Test We further include two types of open-domain evaluations to assess whether the induced CDTs generalize beyond the storylines on which they are trained. On the Fandom benchmark, we propose 200 starting scenes that initiate novel interactive settings (e.g., battles, casual conversations) between characters from the same artifact, and roll out 10 turns of interaction for each scene. CDT is compared against the Human Profile baseline, with responses evaluated by both an LLM judge (`gpt-4.1`) and human annotators. As shown in Figure 5, CDT is preferred over Human Profile in 38.0% versus 26.0% of cases by the LLM judge, and 39.0% versus 27.5% by human evaluators, indicating that CDT generalizes better to unseen interactive scenarios.

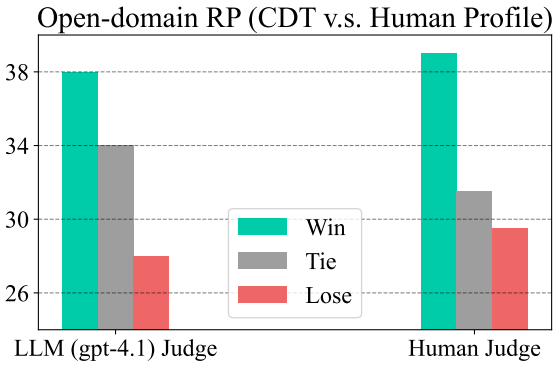

Figure 5: Comparison on open-ended RP by Human and LLM judgment. (CDT represents CDT-Lite)

On the Bandori benchmark, we sample 100 scene–action pairs per band member from the wild event stories (drawn from a pool of 77.2K actions) and evaluate CDTs trained only on the first half of the main band stories. Table 11 shows that CDT consistently outperforms Vanilla prompting and Human Profiles across all eight bands, demonstrating robust out-of-domain generalization from curated storylines to heterogeneous, real-world conversational data. Together, these results confirm that codified, situation-aware profiles learned by CDT transfer effectively to novel scenes and interaction patterns beyond the original training distributions.

Bandori	PoPiPa	AG	PasuPare	Roselia
Vanilla	71.80	70.20	68.60	70.60
Human Profile	74.60	69.90	73.70	73.50
CDT	85.40	81.40	85.10	74.50
Bandori	HHW	Monica	RAS	MyGO
Vanilla	72.70	69.60	72.20	63.10
Human Profile	79.00	70.00	74.90	69.80
CDT	80.60	81.10	81.10	72.00

Table 11: Out-of-domain (OOD) RP performance evaluation with CDT trained on main band stories tested on wild event stories.

	Metric	CDT-Lite	Boosted CDT-Lite
Fandom	#Gen. Call	58.47	23.28
	#Disc. Call	40.62K	9.04K
	#Node	80.13	5.18
Bandori	#Gen. Call	37.35	18.57
	#Disc. Call	12.06K	3.08K
	#Node	28.50	5.82

Table 13: Training and Grounding efficiency of (Boosted) CDT-Lite per character.

E Cost and Efficiency

Boosted CDT To further improve efficiency, we introduce **Boosted CDT**, which retains only the most important hypothesized triggers (eight in our implementation) before validation. This reduces redundancy and focuses validation on the core behavioral logic, at the cost of omitting some fine-grained details. As shown in Table 12, Boosted CDT achieves performance close to CDT-Lite on both Fandom and Bandori, while reducing the number of nodes in the tree by an order of magnitude.

Training and Grounding Efficiency Table 13 summarizes the computational efficiency of CDT-Lite and Boosted CDT. A key observation is that CDT construction is *validation-heavy*: the number of discrimination calls far exceeds the number of generation calls, since each hypothesized trigger must be tested against all scene–action pairs. This motivates the use of distilled discriminators (e.g., `deberta-v3-base`) to replace expensive LLM validators. With distilled discrimination, CDT becomes practical to train at scale, as validation, which is the dominant cost, can be performed cheaply without significant performance degradation.

The number of nodes in each CDT (Table 12, bottom rows) also approximates the maximal traversal steps needed during inference. Because traversal is performed via distilled discriminators rather than full LLM calls, its runtime cost is negligible compared to generating the final RP response. Consequently, CDT grounding remains lightweight at inference time, and Boosted CDT offers additional speedups in trade-off for lower accuracy by further shrinking the tree.

Fandom		Haruhi	K-On!	S×F	DN	FMA	JOJO	AGOT	ATLA
NLI	Human Profile	55.87	55.86	59.14	64.75	58.54	55.11	59.35	57.98
	CDT-Lite	62.17	57.24	59.79	67.00	59.04	57.26	64.27	61.32
	Boosted CDT-Lite	61.70	57.19	60.04	66.29	58.53	58.15	61.42	60.00
#Node	CDT-Lite	3.80	32.80	162.33	45.80	73.40	32.14	79.73	309.50
	Boosted CDT-Lite	1.60	6.60	8.33	2.40	5.60	4.71	4.91	10.00
Bandori		PoPiPa	AG	PasuPare	Roselia	HHW	Monica	RAS	MyGO
NLI	Human Profile	73.73	72.43	77.11	70.08	73.14	68.08	71.74	63.91
	CDT-Lite	88.38	80.49	82.47	72.81	79.66	78.67	79.51	70.33
	Boosted CDT-Lite	86.96	76.19	79.81	74.88	78.34	75.44	78.29	69.74
#Node	CDT-Lite	10.40	28.40	41.40	21.80	5.20	52.00	59.00	9.80
	Boosted CDT-Lite	5.80	6.60	9.80	4.60	3.20	4.40	10.40	1.80

Table 12: Further boosting the efficiency of CDT and resulted performance.

F Profiling and Inference Cases

We visualize running cases of CDT behavior hypothesis construction in Figure 6, and the traversal (grounding) stage in Figure 7, using *"Haruhi Suzumiya"* as the instance.

Figure 6 shows how CDT generalizes from clustered scene–action evidence into reusable `IF-THEN` hypotheses. Each cluster groups semantically similar scenes with an observed action, and the LLM summarizes them into trigger conditions (`IF`) and behavioral statements (`THEN`). In this example, the hypotheses characterize Haruhi as someone who eagerly pursues unusual or unexplained phenomena, initiates group activities or investigations, and frequently assigns roles or tasks to other SOS Brigade members. The extracted rules are stored as nodes in the CDT for later retrieval.

At inference time, Figure 7 illustrates traversal over the hypothesis tree given a new scene. The system checks candidate `IF` conditions, selects those supported by the scene, and collects the corresponding `THEN` statements along the traversal path. These statements are merged into a compact guidance set that constrains the response generation. In the running case, the merged guidance supports predicting Haruhi will actively steer the filming (e.g., instructing Kyon to modify a scene to match her vision), grounded by an entailed reference action from the evidence (e.g., Haruhi using prop guns to drive pigeons toward Asahina).

Wikification pipeline The CDT is first linearized into a verbalized CDT. Then, the prompts in Figure 13 are used to propose chapters and fill in chapter contents to build the wikified CDT.

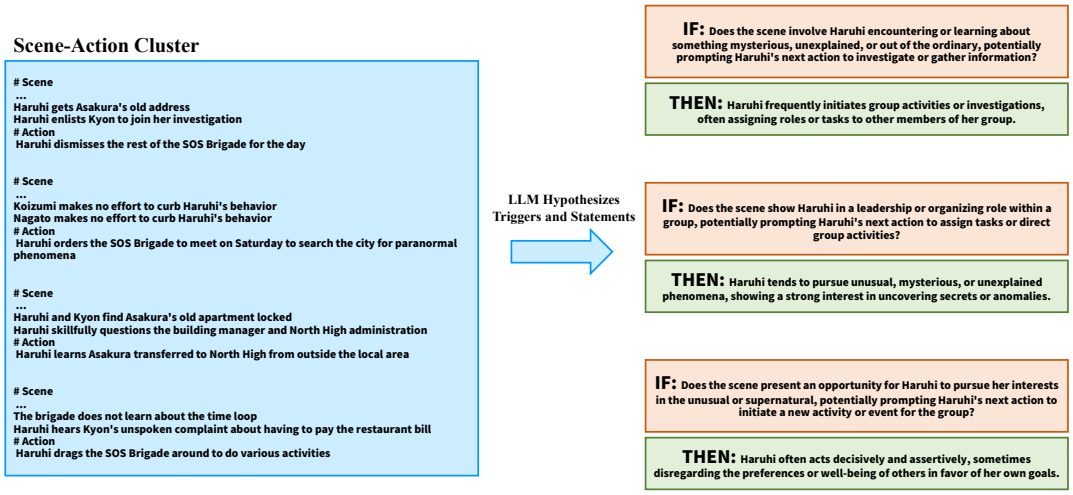

Figure 6: A running case for behavior hypothesis in the CDT building.

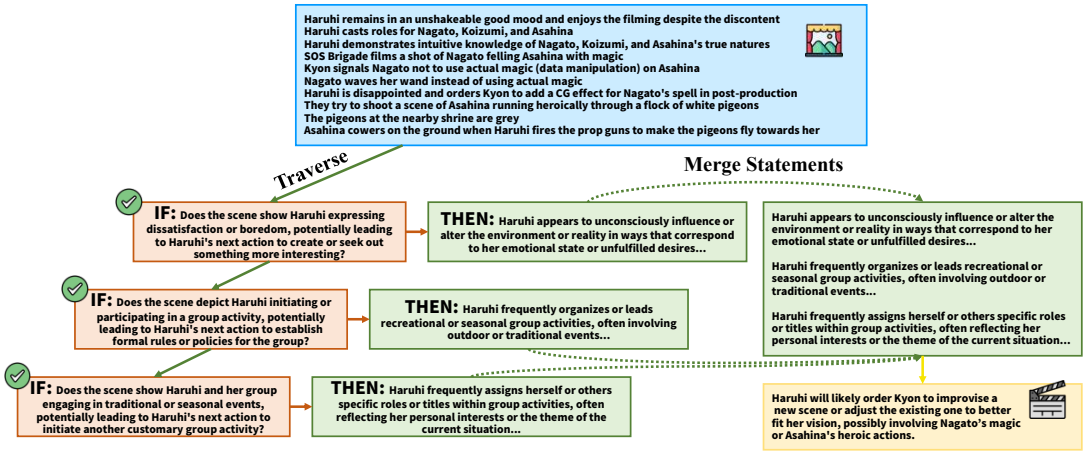

Figure 7: A running case for tree traversal in the CDT grounding.

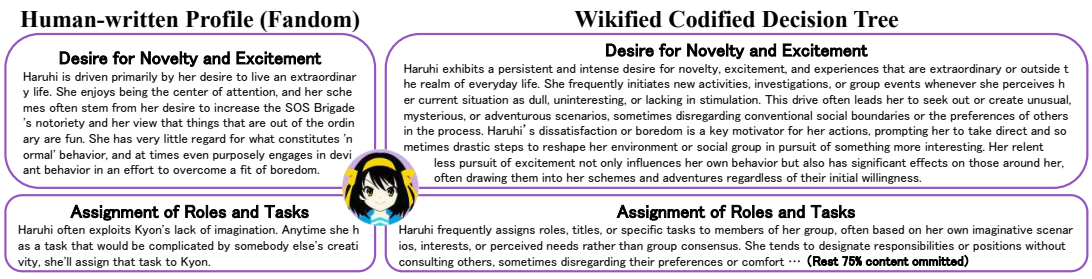

Figure 8: Granularity and coverage comparison between human profile and CDT. (Full wikified CDT in Figure 9, 10, and 11)

- Initiative and Leadership -

Haruhi consistently demonstrates strong initiative and leadership qualities within her group. She frequently takes charge in ambiguous si tuations, often organizing and directing group activities or investigations, especially when she perceives her environment as dull or lacki ng excitement. Haruhi is proactive in seeking out or creating opportunities for adventure, mystery, or the supernatural, and she rarely he sitates to act on her ideas or impulses. She tends to assign roles and responsibilities to others based on her own goals or i maginative scen arios, sometimes disregarding their preferences or consent. Haruhi often establishes formal rules, policies, or titles for her group, reflecti ng her vision and sense of purpose. Her leadership style is assertive and sometimes coercive, as she may use manipulation or threats to m otivate others or secure resources for her plans. Overall, Haruhi's initiative and leadership are characterized by decisiveness, creativity, a nd a willingness to challenge social norms to achieve her objectives.

- Desire for Novelty and Excitement -

Haruhi exhibits a persistent and intense desire for novelty, excitement, and experiences that are extraordinary or outside the realm of ev eryday life. She frequently initiates new activities, investigations, or group events whenever she perceives her current situation as dull, u ninteresting, or lacking in stimulation. This drive often leads her to seek out or create unusual, mysterious, or adventurous scenarios, so metimes disregarding conventional social boundaries or the preferences of others in the process. Haruhi's dissatisfaction or boredom is a key motivator for her actions, prompting her to take direct and sometimes drastic steps to reshape her environment or social group in pu rsuit of something more interesting. Her relentless pursuit of excitement not only influences her own behavior but also has significant eff ects on those around her, often drawing them into her schemes and adventures regardless of their initial willingness.

- Tendency to Organize and Direct Others -

Haruhi consistently demonstrates strong organizational and leadership tendencies within her social group. She frequently initiates group activities, investigations, or events, often taking charge without consulting others or considering their preferences. Haruhi assigns roles, tasks, and responsibilities to group members based on her own goals or imaginative scenarios, sometimes disregarding group consensus or individual comfort. She is known to appropriate or repurpose resources, spaces, and even people to suit her objectives, regardless of t heir original purpose or ownership. Haruhi often establishes formal rules, policies, or titles for her group and activities, reflecting her per sonal vision and sense of purpose. Her assertive approach can include manipulation, coercion, or threats to motivate others or secure nec essary resources. Overall, Haruhi's drive for novelty and excitement is closely tied to her tendency to organize and direct those around he r, often resulting in significant changes to group structure and dynamics.

- Disregard for Social Norms and Others' Preferences -

Haruhi frequently disregards established social norms, rules, and the preferences or comfort of others in pursuit of her own interests and goals. She often acts unilaterally, making decisions for her group without seeking consensus or considering the wishes of her peers. Haru hi is known to assign roles, tasks, or responsibilities to others based on her own imaginative scenarios, sometimes without their consent or understanding. She may appropriate or repurpose resources, spaces, or even people to suit her objectives, regardless of their original purpose or ownership. Additionally, Haruhi is willing to use manipulation, coercion, or threats to motivate others or secure what she nee ds for her plans. Her disregard for conventional boundaries and willingness to challenge or ignore social expectations often leads to unpre dictable and sometimes significant consequences for those around her.

- Proactive Response to Boredom or Dissatisfaction -

When Haruhi experiences boredom or dissatisfaction with her current circumstances, she exhibits a strong tendency to take imm ediate a nd proactive action to seek out or create more interesting, novel, or extraordinary experiences. This often involves initiating new activiti es, investigations, or group events, frequently without consulting others or considering their preferences. Haruhi's desire for excitement and aversion to monotony can lead her to disregard social norms, established rules, or the comfort of those around her. In so me cases, he r emotional state and unfulfilled desires unconsciously manifest as reality-altering phenomena, resulting in significant and sometimes su pernatural changes to her environment. These actions may include the sudden appearance or disappearance of people or objects, the cre ation of alternate spaces, or the repetition of time loops. Haruhi's proactive response to boredom not only drives the narrative forward b ut also has far-reaching consequences for her surroundings and the people involved with her.

- Reality-Altering Abilities (Unconscious) -

Haruhi possesses the extraordinary, unconscious ability to alter, reshape, or influence reality in accordance with her desires, expectation s, or emotional state. These reality-altering phenomena often manifest without her explicit awareness or intent, typically in response to f eelings of boredom, dissatisfaction, or unmet expectations. Haruhi's powers can result in large-scale or supernatural changes, such as the creation of alternate spaces, time loops, or the sudden appearance and disappearance of people, objects, or phenomena. Her em otional fl uctuations and desires can unconsciously trigger these events, sometimes leading to significant and unpredictable consequences for her environment and those around her. Haruhi's reality-altering abilities are closely tied to her interactions with her close group of peers and are often catalyzed by her pursuit of novelty, excitement, or the resolution of mysteries. Despite the magnitude of these effects, Haruhi r emains unaware of her own influence, and her actions are typically driven by her search for extraordinary experiences rather than any co nscious intent to manipulate reality.

- Influence of Emotional State on Environment -

Haruhi's emotional state, particularly feelings of boredom, dissatisfaction, or agitation, has a profound and often unconscious influence o n her environment. When Haruhi perceives her surroundings as uninteresting or lacking excitement, she may unconsciously alter or resha pe reality to better align with her desires or emotional needs. This can result in the sudden appearance or disappearance of people, object s, or phenomena, as well as large-scale or supernatural changes such as time loops or the creation of alternate spaces. These reality-alteri ng effects typically occur without Haruhi's explicit awareness of her own role in causing them. Her actions, driven by her em otional state, can have significant and unpredictable consequences for her surroundings, often affecting not only her immediate group but also the bro ader environment or timeline. This influence is cyclical, as Haruhi's boredom or agitation frequently precedes major changes, and her pur suit of novelty or excitement can lead to repeated or escalating alterations in reality.

- Initiative and Leadership -

Haruhi consistently demonstrates strong initiative and leadership qualities within her group. She frequently takes charge in ambiguous si tuations, often organizing and directing group activities or investigations, especially when she perceives her environment as dull or lacki ng excitement. Haruhi is proactive in seeking out or creating opportunities for adventure, mystery, or the supernatural, and she rarely he sitates to act on her ideas or impulses. She tends to assign roles and responsibilities to others based on her own goals or i maginative scen arios, sometimes disregarding their preferences or consent. Haruhi often establishes formal rules, policies, or titles for her group, reflecti ng her vision and sense of purpose. Her leadership style is assertive and sometimes coercive, as she may use manipulation or threats to m otivate others or secure resources for her plans. Overall, Haruhi's initiative and leadership are characterized by decisiveness, creativity, a nd a willingness to challenge social norms to achieve her objectives.

- Desire for Novelty and Excitement -

Haruhi exhibits a persistent and intense desire for novelty, excitement, and experiences that are extraordinary or outside the realm of ev eryday life. She frequently initiates new activities, investigations, or group events whenever she perceives her current situation as dull, u ninteresting, or lacking in stimulation. This drive often leads her to seek out or create unusual, mysterious, or adventurous scenarios, so metimes disregarding conventional social boundaries or the preferences of others in the process. Haruhi's dissatisfaction or boredom is a key motivator for her actions, prompting her to take direct and sometimes drastic steps to reshape her environment or social group in pu rsuit of something more interesting. Her relentless pursuit of excitement not only influences her own behavior but also has significant eff ects on those around her, often drawing them into her schemes and adventures regardless of their initial willingness.

Figure 9: Full wikified CDT for *"Haruhi Suzumiya"* (1/3)

- Pattern of Group Activity Initiation -

Haruhi consistently demonstrates a strong tendency to initiate group activities, often acting as the driving force behind her club or social circle's events. She frequently proposes or organizes new activities, especially when she perceives her current environment as dull or lacking excitement. Haruhi's approach is proactive and sometimes impulsive, with little regard for the preferences or comfort of others; she often involves her peers in her plans without their prior agreement or understanding. Her desire for novelty and adventure leads her to seek out or create situations that are unusual, mysterious, or extraordinary, and she is quick to take charge in ambiguous situations. Haruhi also tends to assign roles or tasks to others according to her own imaginative scenarios, and she may establish formal rules or policies for the group to suit her objectives. This pattern of group activity initiation is closely tied to her emotional state—periods of boredom or dissatisfaction often precede her most significant and reality-altering actions.

- Assignment of Roles and Tasks -

Haruhi frequently assigns roles, titles, or specific tasks to members of her group, often based on her own imaginative scenarios, interests, or perceived needs rather than group consensus. She tends to designate responsibilities or positions without consulting others, sometimes disregarding their preferences or comfort. Haruhi often reassigns or changes these roles as her interests shift or as new activities arise. This assignment of roles can include formal titles, thematic designations, or specific duties within group activities, and is sometimes accompanied by manipulation, coercion, or threats to ensure compliance. Her approach to assigning roles reflects her strong leadership tendencies and desire to organize group dynamics according to her vision, often appropriating people, spaces, or resources to suit her objectives.

- Manipulation and Coercion -

Haruhi frequently employs manipulation, coercion, or even threats to achieve her objectives or to motivate others to participate in her plans. She often disregards the preferences, comfort, or consent of those around her, assigning roles or responsibilities based on her own goals rather than group consensus. Haruhi is willing to use blackmail or other forms of pressure to acquire resources or secure cooperation, especially when pursuing activities she finds interesting or necessary for her group. This tendency is evident both in her day-to-day interactions and in situations involving the organization of group events, investigations, or the appropriation of spaces and resources. Her assertive and sometimes forceful approach to leadership often results in others being compelled to follow her direction, regardless of their initial willingness.

- Escalation in Response to Mysteries or Ambiguity -

When confronted with mysteries, ambiguity, or unexplained phenomena, Haruhi Suzumiya exhibits a strong tendency to escalate her involvement and take decisive action. She often assumes a leadership role, directing others and organizing group investigations or activities aimed at uncovering the truth or resolving the unknown. Haruhi is quick to disregard established rules, social norms, or the preferences of others if they conflict with her pursuit of answers or excitement. Her drive to resolve ambiguity can lead her to assign roles or tasks to her peers without their consent, manipulate situations, or repurpose resources to suit her objectives. In some cases, her emotional state and dissatisfaction with the ordinary can unconsciously trigger reality-altering events or supernatural phenomena, further intensifying the situation. This pattern of escalation is characterized by her proactive, sometimes impulsive, approach to mysteries, often resulting in significant and unpredictable consequences for her environment and those around her.

- Creation and Investigation of Mysteries -

Haruhi Suzumiya exhibits a strong and persistent drive to seek out, create, and investigate mysteries, particularly those involving the unusual, supernatural, or unexplained. She frequently initiates group activities or investigations centered around uncovering secrets or resolving ambiguities, often taking charge and directing others according to her own imaginative scenarios. Haruhi is quick to act when she perceives a lack of excitement or encounters something mysterious, sometimes disregarding social norms or the preferences of her peers in pursuit of her goals.

Her dissatisfaction or boredom often serves as a catalyst for the creation or escalation of mysterious events, and she appears to possess an unconscious ability to alter reality in ways that manifest new mysteries or supernatural phenomena. This can result in large-scale or reality-altering consequences, such as time loops, alternate spaces, or the sudden appearance of unexplained entities or events, typically without her explicit awareness of her own influence.

Haruhi tends to assign investigative roles or tasks to her group members, sometimes without their consent, and will escalate her involvement or control in response to new or unexplained information. She often disregards conventional explanations, instead seeking extraordinary or supernatural outcomes, and may employ manipulation or coercion to motivate others to participate in her investigations. Her actions frequently lead to the repetition or variation of group activities, suggesting a cyclical pattern in her pursuit of mysteries. Overall, Haruhi's relentless curiosity and leadership drive the ongoing creation and investigation of mysteries within her social group and environment.

- Repetition and Cyclical Behavior -

Haruhi exhibits a notable pattern of repetition and cyclical behavior, particularly in the context of group activities and the passage of time. When faced with boredom, dissatisfaction, or the end of a time period (such as the conclusion of summer), Haruhi often initiates new events or activities to inject excitement into her environment. This drive for novelty can result in the repetition or variation of similar group activities, sometimes leading to cycles where events or time periods are experienced multiple times, such as time loops or the resetting of events. These cycles are frequently triggered by Haruhi's emotional state or unfulfilled desires, and may occur without her explicit awareness of their supernatural or reality-altering nature. Haruhi's actions in these scenarios can have significant consequences for those around her, as her unconscious influence may cause large-scale changes to the environment or timeline. This cyclical behavior underscores her tendency to seek out or create extraordinary experiences whenever ordinary circumstances fail to meet her expectations.

- Appropriation and Repurposing of Resources -

Haruhi frequently appropriates or repurposes physical spaces, resources, or even people to suit her immediate objectives, often without regard for their original purpose or ownership. This behavior is evident in her tendency to commandeer rooms, equipment, or other assets for club activities or investigations, regardless of established rules or the preferences of others. Haruhi's actions are typically driven by her own interests or the thematic needs of her current pursuits, and she rarely seeks consensus or permission before reallocating resources. This pattern of behavior reflects her assertive leadership style and her willingness to challenge social norms or boundaries in pursuit of novelty, excitement, or the fulfillment of her goals.

- Unpredictable and Large-Scale Consequences -

Haruhi Suzumiya possesses the unconscious ability to alter or reshape reality in accordance with her desires or emotional state, often without her explicit awareness. When Haruhi experiences dissatisfaction, boredom, or agitation, her actions can result in sudden and unpredictable changes to her environment, including the appearance or disappearance of people, objects, or supernatural phenomena. These reality-altering effects may manifest as large-scale consequences, such as time loops, the creation of alternate spaces, or other significant disruptions to the normal world. Haruhi frequently disregards established rules, norms, or the preferences of others in pursuit of her goals, sometimes escalating her involvement or control in response to new or unexplained information. Her tendency to act decisively and assign roles or tasks to others—often without their consent—can further amplify the impact of her actions. The full extent of these consequences is not always immediately apparent to those around her, and may only become evident as events unfold or repeat in unexpected ways.

Figure 10: Full wikified CDT for *"Haruhi Suzumiya"* (2/3)

- Interaction with Close Peer Group -

Haruhi's actions and decisions are closely intertwined with her interactions with a select group of peers, most notably the members of her club. She frequently initiates group activities, investigations, or adventures, often involving her friends regardless of their initial willingness or understanding. Haruhi tends to assert her authority within the group, assigning roles, tasks, or titles based on her own imaginative scenarios or goals, sometimes disregarding the preferences or comfort of others. Her emotional state, particularly boredom or dissatisfaction, often prompts her to seek out or create new experiences for the group, which can lead to significant changes in their environment or even reality itself. Haruhi's leadership style is proactive and sometimes coercive, utilizing manipulation or threats to motivate participation. The dynamic within her peer group is a driving force behind many of her actions, and her expectations or unfulfilled desires can unconsciously influence the experiences and circumstances of those around her, occasionally resulting in supernatural or reality-altering phenomena.

- Disregard for Established Rules and Procedures -

Haruhi consistently demonstrates a disregard for established rules, social norms, and conventional procedures, especially when they conflict with her own interests or goals. She often acts unilaterally, making decisions and taking actions without seeking consensus or considering the preferences of others. Haruhi is willing to manipulate, coerce, or threaten others to achieve her objectives, and she frequently appropriates or repurposes resources, spaces, or roles regardless of their original purpose or ownership. Her tendency to ignore boundaries and established protocols can result in significant and unpredictable consequences, including large-scale or reality-altering phenomena. This behavior is particularly evident when Haruhi perceives ambiguity, boredom, or a lack of excitement, prompting her to take direct and sometimes drastic action to reshape her environment or social group according to her desires.

- Sudden Changes in Focus or Interests -

Haruhi Suzumiya is characterized by her frequent and abrupt shifts in focus or interests, especially when she perceives her current situation as dull or unfulfilling. She often abandons previous commitments or activities in favor of new pursuits that promise greater excitement, novelty, or mystery. These sudden changes are typically driven by her emotional state—particularly boredom or agitation—which can unconsciously influence her actions and even alter reality itself. Haruhi's tendency to rapidly redirect her attention often leads her to initiate new group activities, investigations, or adventures, sometimes without consulting or considering the preferences of others. This pattern of behavior can result in significant and sometimes supernatural changes to her environment, as well as the repetition or resetting of events. Her impulsive nature and desire for extraordinary experiences make her a dynamic but unpredictable leader, frequently propelling her and those around her into new and unexpected situations.

- Participation in Traditional or Seasonal Events -

Haruhi frequently organizes and leads her group in participating in traditional or seasonal activities, such as festivals, holiday celebrations, or customary outings. She often assigns specific roles or titles to herself and other group members that reflect her personal interests or the theme of the event. Haruhi demonstrates a pattern of appropriating or repurposing physical spaces and resources to suit the objectives of these activities, regardless of their original function or ownership. She may use manipulation, coercion, or threats to motivate others to participate or to achieve her goals. Occasionally, Haruhi's actions during these events can result in the repetition or resetting of time or events, affecting the experiences of those around her. Additionally, she sometimes initiates or becomes involved in the creation or investigation of mysteries, particularly when the circumstances of the event are unusual or unexplained.

- Unconscious Response to Unspoken Thoughts or Feelings -

Haruhi Suzumiya possesses the remarkable ability to unconsciously respond to her own or others' unspoken thoughts, feelings, or desires. This often results in unexpected or even supernatural outcomes, such as the sudden manifestation of phenomena, changes in reality, or the appearance of anomalies that reflect the emotional state or hidden wishes of those around her. Haruhi herself is typically unaware of this power and does not consciously control these reality-altering effects. Instead, her subconscious reactions to boredom, dissatisfaction, or the unexpressed emotions of her peers can trigger significant and unpredictable changes in her environment, sometimes leading to large-scale or reality-bending events. This aspect of her character is central to the mysterious and often chaotic nature of the world around her, as her unconscious influence can reshape circumstances to align with her or others' unspoken needs or expectations.

Figure 11: Full wikified CDT for *"Haruhi Suzumiya"* (3/3)

G Prompts and Templates

We place the prompts used in the main experiment and analyses in Figures 12, 13, and 14 for result reproduction.

Prompt #1: Hypothesize Rules

```
# Scene-Action Pairs
{action_scene_context}

# Established Statements
{established_statement_verbalized}

# Already Proposed Common Points
{gate_path_verbalized}

# Task

Your task is to build the grounding logic for an AI system to understand the behavior of {character}, assert the AI system has no pri
or knowledge of {character}.
To do this, please propose hypotheses for the general behavior logic of {character} based on the given action-scene pairs, complete t
he task step by step:

1. What's the main feature of {character}'s behavior (e.g., identity, personality, appearance, relationship, ability, superpower mech
anism, etc.) shown in the given scene-action pairs, **other than the already established statements**?

2. Summarize {k} potential common points (grounding statements) of the actions taken by {character} in the given scenes, **which is o
ther than the already established statements**.
- The grounding statements should be general, avoiding too specific action descriptions. (except when it's a common skill of the char
acter)
- Consider the grounding statements in a general way. (e.g., identity, personality, appearance, relationship, ability, superpower mec
hanism, etc.)
- The grounding statements should be concise, informative, and general sentences.
- Never be assertive! Always make objective description of the character rather than making assertive causal relations.

3. Summarize {k} potential common points of the given scenes that trigger each behavior, **which should be different from already pro
posed common points.**
- The question should be simple, not ambiguous, and specific to a subset of scenes rather than always applicable.
- Focus on the **next action** when asking! Don't ask whether certain event is involved, instead ask whether the scene might trigger
potential behavior for {character}'s **next action**.
- Directly include "{character}'s next action" in the question!

4. Output the hypothesized scene-action triggers in the following format:
```python
action_hypotheses = [] # A list of syntactically complete statements (always mentioning {character})
scene_check_hypotheses = [] # A list of syntactically complete questions to check the given scene (always mentioning {character})
```
```

Prompt #2: Discriminate Scenes

```
Scene: {text}

Question: {question}

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

Prompt #3: 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 #4: Next Action Prediction

```
# Background Knowledge
{grounding}

# Scene
{scene}

# Question
What'll be {character}'s next action in response to the current scene? Answer a concise narration in one sentence.
```

Prompt #5: NLI 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 12: Prompts/Templates used in the main experiments for CDT.

Prompt #6: ETA (Extraction)

```
# Task

Please provide a 1000-word, narrative-style character profile for {character}.
The profile should read like a cohesive introduction, weaving together the character's background, physical description, personality t
raits and core motivations, notable attributes, relationships, key experiences, major plot involvement and key decisions or actions, c
haracter arc or development throughout the story, and other important details.
The profile should be written in a concise yet informative style, similar to what one might find in a comprehensive character guide, i
n language. Focus on the most crucial information that gives readers a clear understanding of the character's significance in the work.
The profile should be based on either your existing knowledge of the character or the provided information, without fabricating or inf
erring any inaccurate or uncertain details.

# Scene-Action Pairs

{block}

Now, based on the given scene-action pairs, please generate the character profile, starting with ===Profile===.
```

Prompt #7: ETA (Aggregation)

```
# Main Profile
{main_profile}

# New Summarized Profile (From New Episodes)
{summarized_profile}

Directly update the main profile based on the new summarized profile, keep its length in around 1000 words.
```

Prompt #8: Wikification (Propose Chapter)

```
{cdt_tree_verbalized}

# Task
The pseudo code above describes the character behavior logic of {character}. Extract the key attributes of {character} in the given co
de that can be turned into sections of character Wiki.

Output in the following python format:
```python
["...", "...", "...", ...]
```
```

Prompt #9: Wikification (Write Chapter)

```
{cdt_tree_verbalized}

# Task
The pseudo code above describes the character behavior logic of {character}. Summarize the information from the given code for "{attri
bute}" section in a character Wiki.

Output in the following python format:
- {chapter_name} -

{{Content}}
```

Prompt #10: Goal-diven Hypothesis

```
…… (Omitted - same as original hypothesis prompt)

# Task

Your task is to build the grounding logic for an AI system to understand the behavior of {character} (Current topic: "{topic}"), asser
t the AI system has no prior knowledge of {character}.
To do this, please propose hypotheses for the general behavior logic of {character} based on the given action-scene pairs, complete th
e task step by step:

1. What's the main feature of {character}'s behavior (e.g., identity, personality, appearance, relationship, ability, superpower mecha
   nism, etc.) shown in the given scene-action pairs, about the focused topic: "{topic}" **other than the already established statemen
   ts**?

…… (Omitted - same as original hypothesis prompt)
```

Figure 13: Prompts/Templates used for baselines and variants.

Prompt #11: Matching Score

```
# Scene
{scene}

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

Dimension: {dim_name}
Task: {dim_task}
Focus: {dim_focus}

Score:
A: Generally Matched - prediction aligns with ground truth on this dimension (paraphrase OK; minor differences OK).
B: Not Matched - differs, contradicts, omits the key signal, or the ground truth doesn't support judging this dimension.

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

Prompt #12: Trigger Filtering for Efficiency

```
# Task: Summarize & Compress Scene-Action Hypothesis Pairs into Top 8

You are given a list of paired hypotheses. Each pair contains:
- "scene_check_hypothesis": a question about {character}'s next action
- "action_hypothesis": a general behavioral grounding statement about {character}

Input pairs:
{paired_hypotheses}

## Goal
Produce a rewritten, deduplicated, and compressed set of **exactly 8** pairs that capture the **most important** and **most general
** behavioral grounding logic for {{character}}.

Rewriting is allowed and encouraged to increase:
- generality
- coverage across different subsets of scenes
- clarity
- non-assertiveness

## Selection Principles (prioritized)
1. **Coverage**: The 8 pairs should collectively cover the widest range of distinct behavioral patterns and distinct scene triggers.
2. **Centrality**: Prefer pairs that reflect recurring or core behaviors across many scene-action pairs.
3. **Specificity without overfitting**: Keep statements general; only keep a specific skill/ability if it appears repeatedly and br
oadly.
4. **Non-redundancy**: Each of the 8 pairs must represent a meaningfully different behavior/trigger from the others.
5. **Pair coherence**: The scene_check_hypothesis must plausibly test for the corresponding action_hypothesis (do not mismatch the
m).

## Dedup & Merge Rules
- You may merge multiple similar input pairs into one rewritten pair.
- If two candidate pairs overlap heavily in either the action or the scene question, combine them into a single more general pair.
- Do not preserve original wording when a clearer/general rewrite is possible.

## Constraints to Preserve
### scene_check_hypothesis
- Must be a **single, simple question**
- Must explicitly contain the exact phrase: "{character}'s next action"
- Must target **scene conditions** that could trigger that next action
- Must be applicable to a **subset** of scenes (not a universal always-true condition)

### action_hypothesis
- Must be a **single, concise sentence**
- Must be **non-assertive** (use “may”, “tends to”, “often appears to”, “is described as”, “is observed as”, etc.)
- Must not invent backstory or assume prior knowledge

## Output Format (JSON only)
Return exactly 8 pairs:

```json
{{
  "top8_pairs": [
    {{
      "scene_check_hypothesis": "...",
      "action_hypothesis": "..."
    }},
    {{
      "scene_check_hypothesis": "...",
      "action_hypothesis": "..."
    }},
    ...
  ]
}}
````

Quality Checklist (must satisfy)

* Exactly 8 pairs.
* No two action_hypothesis items mean the same thing.
* No two scene_check_hypothesis questions ask the same trigger.
* Each scene_check_hypothesis clearly tests for {character}'s next action.
* Each action_hypothesis is general, grounded, and non-assertive.
```

Figure 14: Prompts/Templates used for experiments in appendices.
```

H Character & Artifact Background Information

We place concise descriptions of artifacts and characters used in our experiments from Table 14 to 17.

I 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.

Artifact	Concise Abstract
The Melancholy of Haruhi Suzumiya	A fast-paced school comedy–mystery in which the whims of the eccentric Haruhi unknowingly distort reality. Everyday club activities are interwoven with supernatural anomalies, while Kyon's pragmatic narration provides stability amid escalating chaos.
K-On!	A gentle, music-focused slice-of-life narrative portraying the everyday experiences of the Light Music Club. Rather than plot-driven conflict, the story highlights friendship, routine, and subtle personal growth through shared musical practice and leisure.
Fullmetal Alchemist	A serious fantasy adventure following two brothers who turn to alchemy to reclaim what they lost in a forbidden ritual. The story blends action with ethical dilemmas, examining themes of sacrifice, human value, and the consequences of power within a turbulent political world.
JoJo's Bizarre Adventure (Part 3)	A worldwide supernatural journey in which Jotaro and his companions battle enemies using Stand abilities. The narrative is defined by inventive combat, strategic confrontations, and flamboyant character interactions that balance intensity with humor.
Spy × Family	A genre-blending family comedy revolving around a secret agent who forms a fake household to complete a covert mission. Espionage, action, and humor coexist with heartfelt moments, as hidden identities collide with genuine emotional bonds.
Death Note	A psychological cat-and-mouse thriller in which a gifted student gains the power to kill by writing names in a supernatural notebook. The narrative explores justice, morality, and ego through intense intellectual duels between equally brilliant adversaries.
A Game of Thrones	A large-scale political fantasy centered on competing noble houses locked in cycles of alliance, betrayal, and warfare. Personal ambition and moral uncertainty unfold against the backdrop of an approaching existential threat beyond human conflicts.
Avatar: The Last Airbender	An epic coming-of-age tale following Aang as he learns to master elemental powers to restore balance to the world. The series combines humor and action with emotional development, emphasizing responsibility, forgiveness, and personal identity.

Table 14: Concise story descriptions of artifacts used in our experiments (Fine-grained Fandom Benchmark part).

Band	Concise Description
Poppin'Party	A bright, guitar-driven pop-rock band formed by high school friends, defined by upbeat melodies and an earnest, forward-chasing spirit. Their story emphasizes friendship, first dreams, and the steady growth that comes from practicing together and performing as a team.
Afterglow	A straight-ahead rock band built on long-standing childhood bonds, carrying a raw, livehouse energy and a no-frills attitude toward music. Their narrative centers on loyalty, everyday honesty, and the tension between staying the same and growing up without breaking the group's core.
Pastel*Palettes	An idol-style unit whose polished, cheerful image is sustained by behind-the-scenes effort, discipline, and constant on-the-job learning. The band's arc focuses on professionalism, teamwork under pressure, and the contrast between staged perfection and genuine personal struggle.
Roselia	A gothic, high-intensity band pursuing a "perfect" sound through rigorous practice, strong ambition, and uncompromising standards. Their storyline highlights artistic pride, conflict born from high expectations, and the hard-won trust required to perform at the highest level.
Hello, Happy World!	A flamboyant, joy-first band that treats performance as a mission to make the world smile, blending showmanship with playful chaos. Comedic set pieces coexist with sincere warmth, as the members' eccentricities ultimately reinforce a shared commitment to spreading happiness.
Morfonica	A melodic rock group distinguished by the prominence of violin, combining refined textures with youthful sensitivity and introspective emotion. Their narrative explores confidence, talent gaps, and self-acceptance, as the band learns to transform insecurity into a cohesive musical identity.
RAISE A SUILEN	A hard-hitting rock/electronic hybrid centered on precision, speed, and stage dominance, built through deliberate recruitment and relentless rehearsal. The story foregrounds professionalism, creative control, and the friction—and eventual cohesion—of strong personalities striving for the same peak.
MyGO!!!!!	A volatile, emotion-forward rock band whose sound is shaped by conflict, vulnerability, and the members' difficulty with honesty and connection. Their arc focuses on miscommunication, fragile belonging, and the intense catharsis of turning personal pain into music and mutual commitment.

Table 15: Concise band descriptions used in our experiments (Bandori Conversational Benchmark part).

Haruhi	Haruhi	An impulsive, hyperactive high school girl whose restless curiosity and odd worldview trigger the story's cascade of bizarre events.
	Kyon	A sardonic, level-headed student who narrates events and acts as Haruhi Suzumiya's reluctant yet stabilizing partner.
	Nagato	A silent, unreadable SOS Brigade member marked by extraordinary intellect and mysterious, otherworldly roots.
	Koizumi	An always-smiling transfer student and esper who aids the Brigade while carefully guarding critical secrets.
	Asahina	A timid, kind upperclassman conscripted into the SOS Brigade as their cute, enigmatic "mascot," frequently dragged into their antics.
K-On!	Yui	The bubbly, scatterbrained lead guitarist of the light music club, whose boundless energy—and sweet tooth—keeps the band moving.
	Ritsu	The boisterous, prank-loving drummer whose playful antics and casual leadership keep the group upbeat and united.
	Mio	A shy but highly capable bassist, gentle at heart and blessed with sharp musical sensitivity.
	Mugi	Tsumugi Kotobuki, a kind, affluent keyboardist who loves pampering her friends and making club life feel luxurious.
	Azusa	A hardworking, gifted junior guitarist who soon becomes essential to the club's tight sound and practice habits.
FMA	Edward	A gifted, stubborn young alchemist who journeys to recover his and his brother's bodies after a catastrophic transmutation.
	Alphonse	A gentle, big-hearted boy whose soul dwells in a hulking suit of armor, traveling with his brother to regain what they lost.
	Winry	A talented automail mechanic and the Elrics' childhood friend, renowned for her technical skill and steadfast compassion.
	Roy	A charismatic, driven State Alchemist and master of flame, intent on reshaping the military from the inside.
	Ling	A charismatic, relentless prince from Xing who pursues immortality while carrying a heavy duty to his nation.
JOJO	Jotaro	A stoic, seemingly unshakable high schooler and "Stardust Crusaders" lead, famed for Star Platinum and iron resolve.
	Polnareff	A bold, flamboyant French swordsman who allies with the Crusaders, fighting through the swift Stand Silver Chariot.
	Joseph	A fast-thinking, over-the-top Joestar whose schemes and bravado—"Your next line is…"—repeatedly flip battles in his favor.
	DIO	A magnetic, utterly ruthless vampire whose towering ambition and cry of "Za Warudo!" cement him as a legendary foe.
	Kakyoin	A composed, analytic ally in "Stardust Crusaders," battling with Hierophant Green, a Stand that attacks with emerald blasts.
	Avdol	A wise, steadfast Egyptian Stand user whose Magician's Red commands fierce flames and unshakable backing.
	Iggy	A grumpy Boston Terrier Stand user with a fondness for coffee gum, whose reluctant heroics turn out to be vital.
DN	Light	A brilliant, idealistic student who acquires the Death Note and resolves to reshape the world through absolute, lethal justice.
	L	An eccentric, reclusive genius detective whose unconventional methods and sharp intuition pit him directly against Kira.
	Near	A calm, analytical prodigy who succeeds L, relying on detached logic and meticulous planning to pursue the truth.
	Misa	A devoted idol and second Kira, driven by love and gratitude, whose impulsive loyalty complicates the deadly mind games.
	Mello	A volatile, fiercely competitive successor to L who embraces risk and criminal alliances to outmaneuver his rivals.
S×F	Loid	An elite undercover agent who assembles a fake family for a high-stakes mission, balancing espionage with improvised parenthood.
	Yor	A soft-spoken civil servant secretly working as a lethal assassin, struggling to reconcile her double life with domestic normalcy.
	Anya	A cheerful, telepathic child who knows everyone's secrets, holding the family together through innocence and quiet insight.
AGOT	Tyrion	The razor-witted youngest Lannister, Tyrion navigates Westerosi politics with wit, nerve, and dark humor despite a lifetime of scorn for his size.
	Daenerys	An exiled Targaryen princess who starts as a hesitant pawn and evolves into a determined, power-claiming ruler.
	Cersei	An ambitious, scheming queen whose beauty conceals a ruthless devotion to her family and grip on power.
	Jaime	The notorious Kingslayer—charming, deadly, and deeply conflicted—whose sworn duties and loyalties are tangled and fraught.
	Robb	The dutiful heir of Winterfell, pushed too soon into command and responsibility by his family's misfortune.
	Eddard	The resolute Lord of Winterfell, a man of stern honor who serves as Warden of the North.
	Arya	A fiercely independent Stark girl who casts off courtly roles in favor of freedom, training, and the blade.
	Catelyn	The determined Lady of Winterfell, driven by fierce maternal loyalty and a firmly practical mind.
	Sansa	The elder Stark daughter, cherished for grace and manners, whose romantic dreams collide with brutal reality.
	Jon	Eddard's brooding illegitimate son, raised at Winterfell and driven by questions of identity, duty, and quiet resolve.
	Bran	A curious young Stark whose devastating fall thrusts him onto an unforeseen and fateful journey.
ATLA	Aang	The final Airbender and hesitant Avatar, playful at heart yet burdened with restoring balance to a world in war.
	Katara	A determined, compassionate waterbender from the Southern Tribe who grounds the group and refuses to tolerate injustice.
	Sokka	A wisecracking, inventive warrior whose boomerang skills and ingenuity repeatedly end up saving the day.
	Zuko	An exiled Fire Nation prince, driven by a burning quest for honor that gradually turns into a search for a new self.

Table 16: Simple background information of characters in our experiments (Fandom Benchmark part).

PoPiPa	Kasumi	An upbeat, starry-eyed vocalist–guitarist whose impulsive enthusiasm pulls people together and kicks off the band's journey.
	Tae	A free-spirited lead guitarist with strong technique and quirky instincts, often drifting at her own pace yet boosting the band's sound.
	Rimi	A shy, gentle bassist who grows braver through performance, bringing careful support and warm sincerity to the group.
	Saaya	A dependable drummer with a caring, family-first mindset, acting as the band's steady backbone in both practice and life.
	Arisa	A sharp-tongued but reliable keyboardist whose practicality and quick thinking keep the band organized, grounded, and moving forward.
AG	Ran	A blunt, prideful vocalist–guitarist who values authenticity, carrying the band's straightforward rock spirit and stubborn resolve.
	Moca	A laid-back lead guitarist with a mischievous streak, masking keen observation and musical confidence behind casual teasing.
	Himari	A bright, encouraging bassist and nominal leader, energizing the group with optimism while trying to hold everyone together.
	Tomoe	A reliable, big-sister drummer who supports others through calm strength, stepping up whenever the band needs stability.
	Tsugumi	A kind keyboardist with a gentle, practical touch, often mediating tensions and keeping the group's everyday rhythm intact.
PasuPare	Aya	A relentlessly earnest vocalist who chases the idol dream through effort and persistence, learning confidence by doing the work.
	Hina	A cheerful, genius guitarist who loves "fun" above all, acting on bright ideas with little hesitation and lots of momentum.
	Chisato	A cool, realistic bassist with strong professionalism, frequently reining in chaos while protecting the group's long-term direction.
	Maya	A drummer with deep audio-gear passion and technical know-how, becoming animated when music setups and stage craft are involved.
	Eve	A sincere keytarist devoted to "bushido," whose wholehearted intensity and kindness can be both inspiring and unexpectedly disruptive.
Roselia	Yukina	A fiercely driven vocalist who pursues a "perfect" sound, pushing herself and others with uncompromising standards and focus.
	Sayo	A serious, disciplined guitarist who relies on hard work over flair, expressing care through responsibility and relentless practice.
	Lisa	A warm, attentive bassist who acts as the band's emotional glue, balancing high ambition with everyday empathy and reassurance.
	Ako	A high-energy drummer with a dramatic, chuuni-tinged flair, bringing loud confidence while still craving recognition and growth.
	Rinko	A shy, soft-spoken keyboardist with exceptional skill, gradually building courage through supportive bonds and shared performances.
HHW	Kokoro	A wealthy, fearless optimist who treats making people smile as a mission, turning wild ideas into surprisingly sincere action.
	Kaoru	A theatrical guitarist who plays the "prince" role with flourish, using charm and melodrama to lift the mood around her.
	Hagumi	A sunny, energetic bassist with an athletic, straightforward vibe, often charging ahead with honest excitement and big smiles.
	Kanon	A timid but kind drummer who constantly pushes past fear, finding bravery through small steps and friends who believe in her.
	Misaki	A pragmatic, overworked coordinator (and DJ) who keeps the group functional, often acting as the lone realist amid cheerful chaos.
Monica	Mashiro	A sensitive vocalist and lyricist who struggles with insecurity, slowly learning to voice her feelings through song and companionship.
	Touko	A flashy, extroverted lead guitarist who loves attention and momentum, bringing brightness while occasionally stirring trouble by impulse.
	Nanami	A multi-talented bassist fixated on being "normal," masking inner conflict with humor and adaptability across many situations.
	Tsukushi	A hardworking drummer and leader who tries to be dependable, persisting through clumsiness with determination and care for the team.
	Rui	A cool, perfection-driven violinist and composer who prioritizes results, gradually confronting the role of emotion and trust in music.
RAS	CHU^2	A demanding genius DJ/producer who builds the band with strict control and ambition, driving everyone toward a professional-level stage.
	LAYER	A sharp, charismatic bassist–vocalist whose powerful presence and steady musicianship anchor the band's sound under intense expectations.
	LOCK	A young, earnest guitarist who grows through pressure and mentorship, balancing admiration with the need to prove her own worth.
	MASKING	A fearless, high-impact drummer who thrives on adrenaline and volume, powering performances with wild confidence and physical intensity.
	PAREO	A devoted keyboardist with a shy core and idol-like polish, channeling loyalty and effort into supporting the band's vision.
MyGO	Tomori	A withdrawn, highly sensitive vocalist and lyricist who clings to "words" for connection, turning pain and longing into songs.
	Anon	A social, image-savvy rhythm guitarist who wants to belong and be seen, learning sincerity as her confident front gets tested.
	Raana	A freewheeling lead guitarist with a mysterious, playful calm, following curiosity and sound first while ignoring most social rules.
	Soyo	A gentle, composed bassist who tries to keep harmony, often caught between caring intentions and the pressure of unresolved history.
	Taki	A blunt, intense drummer and composer whose strict standards hide protectiveness, expressing concern through sharp honesty and persistence.

Table 17: Simple background information of characters in our experiments (Bandori Benchmark part).