CharacterHub: Open-Domain Character Profiling for LLM Role-play via Deep Search

Anonymous ACL submission

Abstract

Building high-quality character profiles is a foundational prerequisite for developing immersive Role-Playing Language Agents (RPLAs). However, existing profiling methods primarily rely on literature-based extraction or LLM-based generation, which suffer from limited media coverage, high manual costs, and a propensity for factual hallucinations. To address these bottlenecks, we propose CHARACTERHUB, an automated character profiling framework powered by deep search agents. Unlike traditional extractive pipelines, our framework autonomously navigates open web sources to retrieve and aggregate heterogeneous information across multiple dimensions. This agentic approach offers unparalleled scalability, extending high-fidelity profiling beyond literary figures to anime, games, and user-generated characters, without human intervention. To rigorously validate our method, we establish an automatic evaluation protocol using large-scale, human-curated data from Fandom[1] as gold reference. Experimental results demonstrate that our dataset achieves strong alignment with reference sources, notably reaching a 83.13% Support Score in the critical personality dimension, while attaining nearly twice the information density of Fandom references. The dataset and associated code resources have been publicly released at https://github.com/Shadow-alpha/DRCharater.

1 Introduction

Recent advances in Large Language Models (LLMs) (Yang et al., 2025; DeepSeek-AI et al., 2025) have facilitated widespread applications across various domains. Among them, Role-Playing Language Agents (RPLAs) (Chen et al., 2024) have gained increasing attention for their ability to construct character-centric interactive AI systems capable of consistently portraying diverse personas or characters. RPLAs enable a wide range of applications, including virtual companions (Sheehy et al., 2024), game NPCs (Park et al., 2023), interactive storytelling (Sun et al., 2023), and educational simulators (Zhao et al., 2025). To build a comprehensive RPLAs system, we require rich types of structured or unstructured character data. Within this data hierarchy, the character profile constitutes the most foundational level, which naturally leads to the research problem of character profiling.

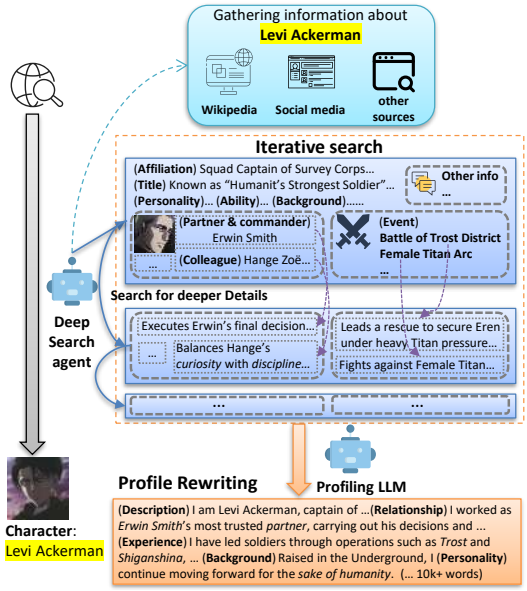

Figure 1: Character Profile Construction Pipeline of CHARACTERHUB. A deep search agent gathers multi-source information about a target character, which is then synthesized by a profiling LLM into a unified character profile.

However, constructing high-quality character profiles remains a major bottleneck for RPLA systems. Existing efforts in character profiling can be broadly categorized into two paradigms: **literature-based extraction** (Li et al., 2023; Yuan et al., 2024) and **LLM-based generation** (Zhou et al., 2024, 2025). The former 1) relies on complex pipelines

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

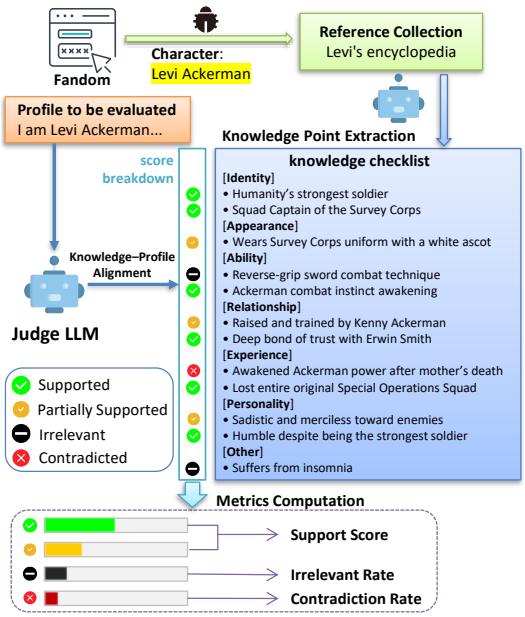

Figure 2: Character profile evaluation pipeline of CHARACTERHUB. Reference character information is extracted from Fandom sources into structured knowledge points, which are then compared against a generated profile by a judge LLM to assign consistency labels and compute evaluation metrics.

to harvest data from long-form narratives but is often resource-intensive; 2) results in sparse profiles due to implicit textual cues like motivations, and 3) fails to generalize beyond characters with extensive written records, like game, animation and virtual characters. The latter offers greater flexibility but frequently suffers from factual hallucinations and a lack of grounding in established lore. Ultimately, these extractive and static approaches struggle to adapt to the evolving universe of characters across diverse media.

To address this challenge, we propose CHARACTERHUB, an automated character profiling framework based on deep search agents that autonomously acquire character information from open web sources. Instead of relying solely on raw narrative texts, our framework actively retrieves structured and semi-structured descriptions from various sources, such as wikis, fan communities, and multimedia-related websites. Through explicit iterative search, the agent constructs comprehensive character profiles covering multiple dimensions such as personality, experiences, relationships, and abilities. Following this pipeline, we further curate a comprehensive character profiling dataset, also named CHARACTERHUB. Overall, our pipeline offers two major advantages: (1) **Broad character applicability**, as it can be applied to a wide spectrum of characters, ranging from literary figures to game, virtual, and user-generated characters; (2) **Flexible and comprehensive profiling**, as it dynamically integrates heterogeneous sources to produce multi-dimensional and fine-grained character representations.

Furthermore, to assess the quality of our profile, we propose an automatic evaluation protocol to assess character profiles. To rigorously assess the quality of the acquired profiles, we design a set of fine-grained evaluation metrics that measure factual correctness, attribute coverage, and semantic consistency. To build this benchmark, we crawl large-scale character pages from Fandom and treat them as gold references for quantitative comparison. Experimental results demonstrate that our automatically acquired profiles achieve strong alignment with human-curated sources, particularly in the most critical Personality dimension where we achieve a Support Score of 83.13%, while exhibiting substantially higher information density—nearly twice that of Fandom references.

In summary, this work makes three key contributions:

1) We propose a scalable, search-agent-based framework for automatic character profile construction across heterogeneous media domains, along with an LLM-based evaluation protocol that derives knowledge-level metrics from Fandom-grounded references.

2) We release a high-quality and diverse character profile dataset CHARACTERHUB, tailored for character-centric RPLA research, covering a wide range of profile dimensions.

3) Extensive experiments, including dimension-wise analysis, ablation studies, and human evaluation, demonstrate the high quality of the proposed dataset and its clear advantages over directly using raw Fandom content.

2 Related Work

Character Profiling Character profiling is the task of extracting attributes of given character to represent a persona. Current methodologies generally follow two paradigms: **literature-based extraction** and **LLM-based generation**.

The first paradigm focuses on harvesting character knowledge from from authoritative sources, including 1) film or animation scripts, such as *ChatHaruhi* (Li et al., 2023) and *RoleLLM* (Wang

et al., 2024), 2) literature and fictional works, such as *CoSER* (Wang et al., 2025), *HPD* (Chen et al., 2023) and *CroSS* (Yuan et al., 2024), 3) Wiki or Baidu, such as *Character-LLM* (Shao et al., 2023) and *CharacterEval* (Tu et al., 2024). These extractive methods are often limited by high manual costs, data sparsity in implicit narratives, and a scope restricted to established literary works.

Conversely, the second paradigm leverages the generative capabilities of LLMs to synthesize profiles through 1) self-alignment, like *DITTO* (Lu et al., 2024), 2) instruction-guided expansion, like *CharacterGLM* (Zhou et al., 2024), *CharacterBench* (Zhou et al., 2025), and 3) multimodal synthesis, like *MMRole* (Dai et al., 2025). Although scalable, it frequently struggles with factual hallucinations and inconsistency.

Deep Research The paradigm of information retrieval has evolved from static **RAG** (Lewis et al., 2020) to dynamic **Agentic Deep Research Architectures**, which utilize **Tool Use** (Schick et al., 2023) and **Function Calling** (Li et al., 2024) to interact with external search APIs, effectively bypassing the model's knowledge cutoff. Central to this evolution is the **ReAct** paradigm (Yao et al., 2023), which interleaves reasoning traces with task-specific actions, enabling autonomous query planning and iterative information synthesis. This capability allows our search-agent framework to target specific character wikis across the open web.

Recent work further extends agentic search into large-scale web data acquisition systems, such as *AutoData* (Ma et al., 2025) and *Agent Ohana* (Zhang et al., 2024), which design multi-agent or unified pipelines for open-domain data collection and agent training. In comparison, our work focuses on *character-centric profiling* for role-play applications. Rather than optimizing crawling scale, we formalize the transformation from heterogeneous evidence to compact character attributes.

3 Dataset Construction

We construct a large-scale, franchise-aware character corpus through (1) Character Collection with Franchise-Aware Merging, resulting in 1,109 high-precision personas; and (2) Profile Generation, where we employ our CHARACTERHUB pipeline to synthesize comprehensive first-person character profiles for RPLA conditioning.

3.1 Character Collection

Data Source and Coverage We assemble a large-scale, multi-domain character corpus by integrating three popularity rankings: AniList GraphQL API, MyAnimeList Jikan API, and a community-curated Character.AI leaderboard. The combined raw set contains 2,139 distinct fictional personas spanning anime, manga, video-games, movies, and western comics, providing both Japanese and English name variants.

Selection and Deduplication To ensure high precision, we implement a franchise-aware merging algorithm. Characters are first grouped by their canonical work (e.g., *Jujutsu Kaisen*, *Genshin Impact*). Within each group, we compute name similarity using *SequenceMatcher*[2]. Two entries are merged only if they (i) originate from different sources, (ii) share an exact franchise match, and (iii) exceed a similarity threshold of 0.85. This strategy successfully consolidates alternate spellings (e.g. "Satoru Gojo" vs. "Gojo Satoru") while preserving cross-franchise homonyms like "Zoro" from *One Piece* and *Demon Slayer*. This process collapsed 32 duplicates, yielding 2,107 unique characters.

Quality Control We filter out characters whose franchise information is missing or cannot be matched to a hand-crafted alias table covering high-impact series, resulting in 1,109 characters that are successfully processed by the CHARACTERHUB pipeline to obtain complete character profiles. The resulting dataset contains character names, normalized franchises, category tags, and cross-platform popularity scores, and serves as the backbone for all character-driven experiments in the remainder of the paper.

3.2 CHARACTERHUB Pipeline

We propose a fully automated search-agent pipeline for open-ended and scalable profile construction for arbitrary characters. As illustrated in Figure 1, the pipeline consists of two stages: character-driven iterative search and profile rewriting.

Character-Driven Iterative Search Given a character name c_i, the system activates a search agent to retrieve character-related information from open web sources. The agent automatically formulates diversified search queries by combining the

[2] https://docs.python.org/3/library/difflib.html

character name with profiling-oriented keywords (e.g., personality, background, story), and issues them to general search engines and wiki platforms.

To improve coverage and reduce ambiguity, the pipeline adopts an iterative search strategy. At iteration k, the retrieved information R_i^k is summarized and fed back to guide the next round of query generation and retrieval, producing refined information R_i^{k+1}. This process repeats for K iterations, progressively expanding under-represented profile dimensions and enabling more comprehensive and fine-grained information acquisition.

Output Profile Rewriting The refined deep-retrieved information $\{R_i^K\}$ is then provided to a profiling model to generate the final character profile P_i in a first-person narrative form. The profiling prompt instructs the LLM to integrate heterogeneous facts into a coherent persona, resolve potential cross-source conflicts, and rewrite the character description as a unified self-introduction. This design is motivated by RPLA requirements, where first-person self-consistent profiles better support immersive role-playing than third-person factual descriptions.

Finally, for each character c_i, the pipeline outputs $\{c_i, R_i^1, \ldots, R_i^K, P_i\}$. The entire process is fully automatic, requires no manual annotation, generalizes to arbitrary characters without task-specific retraining, and scales naturally to web-scale character discovery, making it well suited for large-scale open-domain character profile construction.

4 Evaluating Character Profiles

To systematically assess the quality of constructed character profiles, we adopt an **LLM-as-a-Judge** evaluation framework grounded in **external fandom-based reference**. The key idea is to treat large-scale fandom wikis as rich yet noisy factual references, extract fine-grained knowledge points from them, and employ an LLM to judge their consistency with the generated character profiles.

As illustrated in Figure 2, The evaluation pipeline consists of four stages: (1) reference collection, (2) knowledge point extraction, (3) knowledge–profile alignment, and (4) metric computation.

4.1 Reference Collection

To establish a reliable evaluation benchmark, we construct an **external reference** for each character by combining the Fandom Wiki API with the Google Search API.

We first perform **webpage matching** to identify the most authoritative source. Given a character name, the Google Search API is used to retrieve candidate URLs, with priority given to pages within the Fandom Wiki domain. When multiple candidates are available, we select the page with the highest textual relevance and the most comprehensive content structure.

After identifying the target page, we conduct **web crawling and content extraction**. Specifically, we extract the **infobox** to obtain concise factual attributes (e.g., gender, affiliation, occupation) and parse the **main content** into a structured dictionary that maps section headings (e.g., *Background*, *Abilities*) to their corresponding narrative paragraphs, preserving the original thematic organization.

Finally, the extracted data is passed through a **normalization** pipeline to remove citation markers, footnotes, navigation artifacts, corrupted Unicode characters, and formatting noise. The cleaned content is standardized into a JSON reference file, serving as a unified factual reference for subsequent knowledge extraction and quantitative evaluation.

4.2 Knowledge Point Extraction

Given the cleaned fandom reference for each character, we employ an LLM to automatically extract **fine-grained knowledge points**, which serve as atomic evaluation units.

Formally, for each character, the LLM outputs a knowledge checklist: $\mathcal{K} = [k_1, k_2, \ldots, k_N]$. Each knowledge point k_i is represented as a structured triple: $k_i = \{s_i, e_i, type_i\}$, where s_i denotes the extracted knowledge statement, e_i denotes the evidence from the input text, and $type_i$ denotes the type of knowledge.

Inspired by previous studies (Yuan et al., 2024; Zhou et al., 2025), we facilitate a **multi-dimensional evaluation** by assigning each point a type $type_i \in$ {*Identity, Appearance, Ability, Relationship, Experience, Personality, Other*}. Detailed definitions for each dimension are provided in Appendix E.1.

4.3 Knowledge–Profile Alignment

To quantitatively assess the factual consistency between synthesized profiles and canonical lore, we adopt an **LLM-as-a-Judge** framework to evaluate the alignment between the generated character

profile and Fandom knowledge points.

For each character, the evaluator LLM is provided with two inputs: the generated profile P and a set of extracted knowledge checklist $\mathcal{K} = [k_1, k_2, \ldots, k_N]$ derived from Fandom. The model evaluates each knowledge point $k_i \in \mathcal{K}$ independently and outputs a judgment pair (y_i, r_i), where y_i denotes an alignment label and r_i provides a brief textual justification.

The alignment label y_i is selected from four categories reflecting different degrees of consistency. A knowledge point is labeled 1) **Supported**(Sup) if it is explicitly stated or clearly entailed by the profile; 2) **Partially Supported**(Par) if only some aspects are confirmed while others are missing or ambiguous; 3) **Irrelevant**(Irr) if the profile contains no information related to the knowledge point; 4) **Contradicted**(Con) if the profile includes statements that directly conflict with or negate it.

The resulting alignment is represented as a structured mapping:

$$\mathcal{A} = (k_i, y_i, r_i)_{i=1}^{N}, \tag{1}$$

which serves as the basis for computing quantitative evaluation metrics. This fine-grained formulation enables localized analysis of factual accuracy and coverage across different profile dimensions.

4.4 Evaluation Metrics

Based on the LLM-generated alignment labels, we define three quantitative metrics to reflect different aspects of profile quality.

Support Score. This metric measures overall factual consistency between the profile and knowledge:

$$SS = \frac{|\mathtt{Sup}| + 0.5 \cdot |\mathtt{Par}|}{|\mathcal{K}|}. \tag{2}$$

Irrelevant Rate. This metric reflects **coverage deficiency**, i.e., how much knowledge is missing from the profile:

$$IR = \frac{|\mathtt{Irr}|}{|\mathcal{K}|}. \tag{3}$$

Contradiction Rate. This metric directly captures **factual inconsistency and hallucination risk**:

$$CR = \frac{|\mathtt{Con}|}{|\mathcal{K}|}. \tag{4}$$

5 Experiment

5.1 Settings

Models for Profile Construction For the iterative search stage, we adopt Gemini-2.5-Pro (Comanici et al., 2025) as the core model in our search-agent–based pipeline, where iterative information gathering is explicitly orchestrated by the agent. In parallel, we employ Doubao-seed-1.6 (Seed et al., 2025) as a representative model of the deep research paradigm, which is prompted only at the initial search step and relies on its internal reasoning to implicitly perform iterative information seeking, serving as a comparative setting without explicit agent control. For profile rewriting stage, we use Claude-Sonnet-4 due to its strong capability in narrative writing and role-playing generation.

Judge Models For profile evaluation, we adopt a unified LLM-as-a-Judge setup. For knowledge point extraction from the Fandom reference, we use Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025). The same model is further used to assess the consistency between each extracted knowledge point and (1) the generated character profile P_i, and (2) the deep-retrieved information R_i^K (DRinfo). All reported metrics are computed by averaging over all knowledge points across all characters.

Ablation Study To examine the contribution of iterative search, we conduct an ablation study by removing the iterative search mechanism while keeping all other components unchanged. The details can be found in Appendix F.1.

Human Annotation Study We conduct a small-scale human annotation study to validate the reliability of the LLM-based evaluator. Results show strong agreement with human judgments. Details are provided in Appendix D.

5.2 Main Results

In the experiments, we wish to answer three research questions: RQ1) Can our method construct high-quality character profiles? RQ2) How does performance vary across different profile dimensions? RQ3) What are the advantages over directly crawling Fandom pages?

5.2.1 Can our method construct high-quality character profiles?

The proposed Search Agent-based method is highly effective in constructing high-quality character profiles with high accuracy and low

Type	Source	Gemini			Doubao		
		SS	IR	CR	SS	IR	CR
Personality	DRInfo	81.67	10.00	0.87	78.04	14.16	1.78
	Profile	83.13	9.95	1.58	80.67	10.55	3.51
Ability	DRInfo	61.37	28.69	1.47	53.93	33.52	4.33
	Profile	58.86	31.91	2.05	50.42	39.49	4.14
Experience	DRInfo	60.42	27.02	5.28	50.23	33.02	10.43
	Profile	60.63	27.00	5.90	50.44	32.36	12.18
Relationship	DRInfo	68.32	19.68	3.06	56.05	28.16	7.51
	Profile	68.94	19.71	3.25	56.58	29.00	7.84
Identity	DRInfo	74.66	11.80	2.77	66.21	13.38	8.16
	Profile	73.46	13.22	3.15	65.25	14.90	9.41
Appearance	DRInfo	57.62	23.25	3.06	42.50	26.60	12.76
	Profile	55.83	25.27	4.25	42.12	30.13	13.78
Other	DRInfo	39.72	50.31	2.38	42.36	46.36	3.81
	Profile	40.57	48.58	3.69	42.54	44.67	6.78
Overall	**DRInfo**	66.54	22.08	2.55	58.36	26.06	6.56
	Profile	66.16	22.65	3.18	58.14	27.04	7.62

Table 1: Evaluation results across seven dimensions and overall performance. Scores are presented as percentages with the % symbol omitted. DRInfo means the raw information retrieved by search agent, Profile means the rewrited profile.

hallucination rates. As shown in Table 1, the Gemini based pipeline achieves a total Support score of 66.54% for raw info and 66.16% for the final profile. These high scores, coupled with a remarkably low overall Contradict rate of 3.18%, demonstrate that the search-agent approach accurately retrieves knowledge consistent with the Fandom reference.

The Profile Rewriting process introduces only marginal information loss, demonstrating the stability of the persona conversion phase. For Gemini, the Support score shows only a marginal decrease of 0.38% (66.54% to 66.16%) during the rewriting phase, while the Contradict rate remains stable under 4%.This indicates that the rewriting LLM effectively preserves the factual integrity of the retrieved knowledge. This minor trade-off between semantic density and narrative style is acceptable given the significant improvement in the naturalness of the first-person perspective.

Our method significantly outperforms the Doubao-based Deep Research architecture. Gemini maintains an overall Support score that is approximately 8% higher than Doubao, while Doubao's overall Contradict rate is more than double that of Gemini. This performance gap can be attributed to the search behavior: Gemini's tool usage tends to prioritize global information sources like Fandom, whereas Doubao frequently relies on Chinese-centric platforms such as Bilibili and Baidu Baike, which may contain narrative discrepancies or different levels of detail compared to the Fandom reference.

5.2.2 How does performance vary across different profile dimensions?

Knowledge extraction performance is strongest in the Personality dimension, which is considered the most critical component for RPLA. Specifically, the Gemini-based pipeline achieves its highest Support score in the Personality dimension, reaching 81.67% for raw info and 83.13% for the final profile. This high performance ensures that the core behavioral logic and linguistic style of the characters are captured with high precision, laying a solid foundation for immersive role-playing experiences.

The Appearance and Experience dimensions present a higher risk of hallucination, likely due to the granular nature of the required information. Doubao's Contradict rates for Experience and Appearance reach 12.18% and 13.78%, respectively, while Gemini also sees its highest contradiction rates in these areas. These dimensions often involve minute physical details (e.g., eye color, specific accessories) or complex chronological events, which are more susceptible to model confusion or information overlap.

Stability during the rewriting phase varies by dimension, with technical Ability descriptions

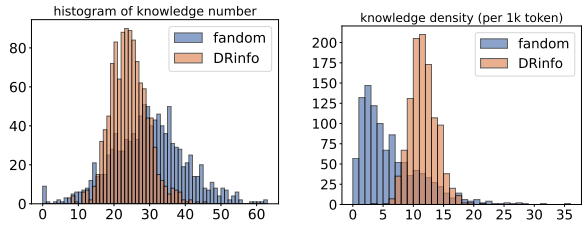

Figure 3: Comparison between Fandom and DRInfo (Gemini) in knowledge quantity and density per character. Left: histogram of the absolute number of extracted knowledge points. Right: histogram of knowledge density, measured as knowledge points number per 1k tokens. DRInfo yields more compact and information-dense character profiles than raw Fandom content.

being more prone to conversion loss. In the Ability dimension, Gemini's Support score dropped from 61.37% to 58.86% after rewriting. This indicates that converting descriptive, often technical ability mechanics into first-person dialogue is more linguistically challenging for the LLM than converting straightforward identity or personality facts.

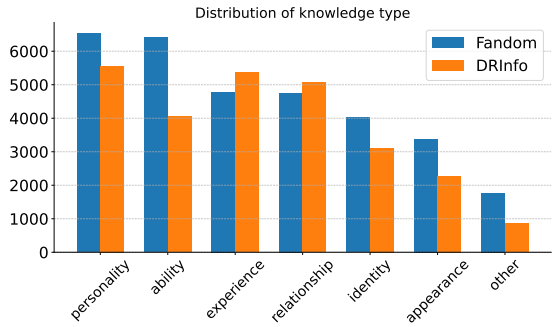

Figure 4: Comparison of knowledge type distributions between Fandom and DRInfo. Fandom exhibits a skewed distribution dominated by personality and ability-related knowledge, whereas DRInfo provides more balanced coverage across experience and relationship-centric dimensions.

5.2.3 What are the advantages over directly crawling Fandom pages?

Directly relying on Fandom introduces popularity bias, whereas our method alleviates it. As illustrated in Figure 3, Fandom pages exhibit substantial imbalance in content completeness across characters: while popular protagonists often have rich and detailed entries, characters from niche intellectual properties or peripheral roles are frequently under-documented, resulting in sparse and fragmented knowledge distributions. In contrast, our approach provides more consistent coverage across characters, which alleviates this dependency on character popularity and yields more stable and informative profiles even for under-documented characters.

Our dataset achieves substantially higher knowledge density. The right histogram of Figure 3 shows that DRInfo consistently yields a higher number of knowledge points per tokens compared to Fandom. This suggests that, despite containing fewer raw tokens, DRInfo encodes character information in a more compact and structured form. In contrast, Fandom pages often include lengthy narrative descriptions, meta information, or loosely related content that dilutes information density.

Our dataset produces a more balanced distribution of knowledge types and enriches event-centric and relational knowledge. As shown in Figure 4, knowledge points extracted from Fandom are heavily skewed toward *personality* and *ability*, reflecting the narrative and fan-driven nature of wiki pages, which often emphasize descriptive traits and combat abilities over structured life events or relational context. In contrast, our dataset exhibits higher counts in *experience* and *relationship* categories, which are particularly important for role-playing and long-horizon character behavior, as they ground the character in narrative history and social context rather than isolated attributes.

These advantages make our dataset more suitable for scalable and generalizable character-centric applications. The improved compactness, higher information density, and reduced popularity bias demonstrate that our approach offers clear advantages over directly using Fandom pages. To further examine the comparison fairness, we conduct a reverse evaluation analysis. The detailed results are reported in Appendix F.2.

5.3 Downstream Role-Playing Evaluation

While previous experiments evaluate the intrinsic quality of the constructed character profiles, an important question is whether these representations are beneficial for downstream role-playing applications. To answer this, we conduct a simulated Role-Playing LLM Agent (RPLA) evaluation.

5.3.1 Experimental Setup

We randomly sample 200 characters from our dataset. For each character, we simulate a dialogue interaction between two LLM agents. A **character LLM** plays the target character and is provided with character information as grounding knowledge. A **user LLM** acts as a regular user

Character Info	Cons.	Pers.	Back.	Inter.
DRInfo	**4.881**	4.634	3.812	4.574
Profile	4.871	**4.644**	**3.851**	**4.752**
Fandom Raw	4.752	4.584	3.802	4.505
Fandom Rewrite	4.820	4.440	3.540	4.690

Table 2: Downstream role-playing evaluation results (average scores). **DRInfo** refers to the raw information retrieved by the deep-search agent before profile rewriting; **Profile** denotes our rewritten first-person character profile; **Fandom Raw** is the original Fandom wiki text; **Fandom Rewrite** represents Fandom text rewritten into first-person form using an LLM.

who is interested in the character and is only given the character name. Each pair engages in approximately 4 dialogue turns.

The generated dialogues are evaluated by a strong judge model along four dimensions:

- **In-Character Consistency (Cons.)**
- **Personality Faithfulness (Pers.)**
- **Background Alignment (Back.)**
- **Interaction Quality (Inter.)**

Each dimension is scored from 1 (very poor) to 5 (excellent). Detailed descriptions of each dimension are provided in Appendix E.2. We use **Qwen3-Max** for both the character LLM and the user LLM, while **GPT-5** serves as the judge model.

5.3.2 Results and Analysis

Curated character representations consistently outperform raw fandom text for downstream role-playing. Both **Profile** and **DRInfo** achieve higher scores than the Fandom baseline across most evaluation dimensions. This indicates that representations produced by our deep-search pipeline provide more suitable grounding for role-playing agents than directly using raw wiki pages. In particular, even the intermediate DRInfo representation already improves over Fandom Raw in three out of four metrics, suggesting that iterative retrieval effectively filters irrelevant content and aggregates character-centric knowledge.

Iterative search primarily benefits knowledge grounding, while first-person rewriting further improves conversational grounding and interaction quality. Compared with DRInfo, the rewritten **Profile** achieves the highest scores in *Personality Faithfulness*, *Background Alignment*, and *Interaction Quality*. The improvement in interaction quality is especially notable (4.752 vs. 4.574), indicating that transforming heterogeneous knowledge into a coherent first-person narrative helps the agent better internalize character identity and produce more natural dialogue responses. These results support the design choice of rewriting retrieved knowledge into a persona-oriented profile.

Naively rewriting raw wiki pages is insufficient and may even degrade character grounding. Although **Fandom Rewrite** improves *Inter.* compared to Fandom Raw, it significantly reduces scores in *Pers.* and *Back.*. This suggests that directly rewriting long and loosely structured wiki pages may distort or omit important contextual details, consistent with the information loss phenomenon discussed in Section 5.2.1. In contrast, our pipeline aggregates and organizes information through iterative search before generating the final profile, leading to more reliable and coherent character grounding for role-playing agents.

Overall, the results demonstrate that the CharacterHub pipeline provides effective knowledge grounding for downstream role-playing agents. By combining iterative search with persona-oriented rewriting, the pipeline produces character representations that are both information-dense and dialogue-friendly, leading to consistent improvements across all evaluated dimensions.

6 Conclusion

In this paper, we present a fully automated, search-agent-based framework for large-scale character profile construction across heterogeneous media domains. By leveraging iterative web search and evidence refinement, our approach acquires comprehensive, fine-grained character information without manual annotation or domain-specific retraining. To rigorously assess profile quality, we introduce an LLM-as-a-Judge evaluation framework grounded in external fandom-based knowledge, enabling localized and quantitative analysis of factual alignment and coverage.

Extensive experiments demonstrate that our constructed profiles achieve higher knowledge density, broader attribute coverage, and more balanced representation across profile dimensions compared to fandom-based references, particularly for long-tail and under-represented characters. We further release a high-quality character profile dataset to facilitate future research on role-playing language agents and character-centric generation. We believe this work provides a scalable foundation for open-domain character profiling and reliable evaluation in character-driven AI systems.

Ethical Statement

This work relies exclusively on publicly available web content and fandom wikis, and does not involve any private, sensitive, or personally identifiable information. The constructed dataset focuses on fictional characters from animation and related media, and is released strictly for research purposes.

We acknowledge that fan-curated sources and large language models may contain factual inaccuracies, subjective interpretations, or popularity-driven biases. In addition, while the dataset targets fictional characters, there remains a potential risk of reinforcing stereotypes or misrepresentations embedded in source materials or model generations. We therefore encourage researchers and practitioners to apply critical scrutiny when using the dataset, especially in interactive or user-facing systems.

We also note that large language model–based assistants were used to support language polishing and presentation refinement during manuscript preparation. These tools were not used to generate experimental data, annotations, or evaluation results, and all technical content and conclusions were authored and verified by the researchers.

Overall, we believe the risks associated with this work are limited and manageable. By openly discussing these limitations and releasing the dataset for academic use, we aim to support transparent, responsible research on character-centric language agents and character profiling.

Limitations

Despite its advantages, our approach has several limitations. First, the quality of retrieved evidence depends on the availability and reliability of web sources; for extremely obscure characters, search results may still be sparse or noisy. Second, while the LLM-as-a-Judge framework enables scalable evaluation, it may exhibit systematic biases, such as confusing missing information with contradictions, which we partially mitigate through human evaluation but do not eliminate entirely. Third, our current pipeline focuses on textual character profiling and does not incorporate multimodal evidence such as images or videos, which could further enrich character understanding. Addressing these limitations is left for future work.

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. *Transactions on Machine Learning Research*. Survey Certification.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 13 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Zekun Li, Zhiyu Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Dong, Adithya Sagar, Xifeng Yan, and Paul A. Crook. 2024. Large

language models as zero-shot dialogue state tracker through function calling. 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 8688–8704. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7828–7840. Association for Computational Linguistics.

Tianyi Ma, Yiyue Qian, Zheyuan Zhang, Zehong Wang, Xiaoye Qian, Feifan Bai, Yifan Ding, Xuwei Luo, Shinan Zhang, Keerthiram Murugesan, Chuxu Zhang, and Yanfang Ye. 2025. Autodata: A multi-agent system for open web data collection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

ByteDance Seed, :, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, Yufeng Yuan, Yu Yue, Lin Yan, Qiying Yu, Xiaochen Zuo, Chi Zhang, Ruofei Zhu, Zhecheng An, and 255 others. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *Preprint*, arXiv:2504.13914.

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

Lisa Sheehy, Stéphane Bouchard, Anupriya Kakkar, Rama El Hakim, Justine Lhoest, and Andrew Frank. 2024. Development and initial testing of an artificial intelligence-based virtual reality companion for people living with dementia in long-term care. *Journal of Clinical Medicine*, 13(18):5574.

Yuqian Sun, Zhouyi Li, Ke Fang, Chang Hee Lee, and Ali Asadipour. 2023. Language as reality: A co-creative storytelling game experience in 1001 nights using generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, October 08-12, 2023, Salt Lake City, UT, USA*, pages 425–434. AAAI Press.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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

Jianguo Zhang, Tian Lan, Rithesh R N, Zhiwei Liu, Weiran Yao, Juntao Tan, Yihao Feng, Thai Quoc Hoang, Tulika Manoj Awalgaonkar, Liangwei Yang, Shelby Heinecke, Huan Wang, Juan Carlos Niebles, Silvio Savarese, and Caiming Xiong. 2024. The

agent ohana: Designing unified data and training pipeline for effective agent learning. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Jiajia Zhao, Jingru Zhang, and Yuhe Lu. 2025. Enhancing design historical education through AI virtual characters role-playing narratives in serious games. *Int. J. Gaming Comput. Mediat. Simulations*, 17(1):1–20.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. Characterbench: Benchmarking character customization of large language models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 26101–26110. AAAI Press.

A Experimental Setup

The hyperparameters of profile construction and evaluation can be see in Table 3. For evaluation, we deploy Qwen3-235B-A22B-Instruct-2507 using data-parallel inference on eight NVIDIA A800 GPUs with 80GB memory each. We choose $K = 2$ in our experiments for cost-efficiency and scalability considerations.

B Cost Analysis

Table 4 reports the average computational cost of our deep-search pipeline per character when using different search models. Gemini achieves lower overall runtime and produces more concise evidence summaries, resulting in fewer output tokens. All statistics are averaged over all 1109 characters.

C Dataset Statistics

This section provides a brief statistical overview of the CHARACTERHUB character dataset used throughout the paper. The dataset contains 1,109 high-precision characters spanning 399 franchises across multiple media domains. Most characters in the dataset are drawn from anime franchises, with the remainder covering games, movies, and novels *etc*. Each character is associated with a comprehensive first-person profile covering seven core dimensions. On average, profiles in CHARACTERHUB exhibit substantially higher information density than their Fandom references, while maintaining strong alignment with human-curated references. Table 5 summarizes key statistics of the dataset.

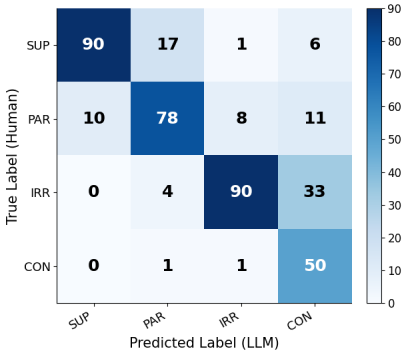

Figure 5: Confusion matrix comparing human annotations and LLM-based judgments for Knowledge–Profile Alignment.

D Human Annotation Analysis

D.1 Instructions for Annotators

To examine the alignment between LLM-based judgments and human judgments, we conduct a small-scale manual annotation study on a subset of knowledge–profile pairs. All annotations are performed internally by the authors for validation purposes.

For each instance, annotators are provided with a generated character profile and a single knowledge point extracted from the Fandom-based ground truth. Annotators are asked to assess the factual relationship between the two using the same four labels as in the LLM evaluation: *Supported*, *Partially Supported*, *Irrelevant*, and *Contradicted*.

Judgments must be based only on the information explicitly contained in the given profile. Annotators should not rely on external knowledge, future events, or information from other versions or spin-offs unless such content is explicitly mentioned in the profile.

All annotated content concerns fictional characters from publicly available sources and involves no personal or sensitive information. As annotations are conducted by the authors without external participants, no recruitment, compensation, or risk disclosure is required.

D.2 Results

Due to the inherent label imbalance, we uniformly sample 100 instances for each label, resulting in 400 manually annotated cases in total.

Overall, the LLM-based evaluator shows strong consistency with human judgments. As shown in Figure 5, the confusion matrix indicates high agreement on the supported (SUP) and irrelevant (IRR) categories. Most SUP instances are correctly identified by the LLM, and IRR cases are accurately filtered, demonstrating that the evaluator reliably distinguishes grounded profile content from unrelated information. However, its most prominent failure mode lies in misclassifying *Irrelevant* knowledge as *Contradicted*, indicating that the model sometimes interprets missing evidence as explicit conflict. More detailed analyses can be found in the Appendix F.3.

Most disagreements occur between supported and partially supported cases, reflecting inherent annotation ambiguity. A noticeable portion of SUP instances are predicted as PAR and vice versa, suggesting that the boundary between full

Parameter	Gemini	Doubao	Qwen-plus	Qwen3-235B
Temperature	1.0	1.0	0.85	0.7
Top-p	0.95	0.8	0.8	0.8
Top-k	64	50	50	20

Table 3: Hyperparameter settings used for different models.

Metric (per character)	Gemini	Doubao
Avg. LLM Calls	2	1
Output Tokens	3,919	5,295
Avg. Runtime	91s	170s

Table 4: Cost and efficiency comparison of the deep-search pipeline using different search models.

Statistic	Value
Number of Characters	1109
Number of Franchises	399
Media Domains	Mainly Anime
Avg. Profile Length (words)	10372
Avg. Profile tokens	2187
Avg. Knowledge Points per Character	24
Information Density (per 1k token)	11.68

Table 5: Statistical overview of the CHARACTERHUB dataset.

and partial support is often fuzzy and depends on coverage granularity rather than factual correctness. This observation aligns with the design of our Support Score, where partially supported knowledge is assigned a weight of 0.5, appropriately capturing its intermediate reliability between fully supported and unsupported content.

Contradicted cases are rare and are detected with high recall. The evaluator correctly identifies the majority of contradicted instances, with very few being confused with other categories. This indicates that the LLM is conservative and robust in identifying factual conflicts between knowledge points and character profiles.

These results justify the use of LLM-as-judge for large-scale evaluation. Given its strong alignment with human judgments on clear cases and reasonable behavior on borderline ones, the LLM-based evaluator provides a reliable and scalable alternative to manual annotation for evaluating character profile quality.

E Detailed Descriptions

E.1 Knowledge Type Dimensions

To provide a more granular understanding of the evaluation framework, this section details the seven semantic dimensions used for character knowledge categorization. These dimensions are designed to cover the full spectrum of a character's persona, ranging from static factual identifiers to latent psychological traits. Each category serves as a distinct axis for evaluating the comprehensiveness and factual integrity of the generated profiles.

Identity Identity refers to factual and relatively stable attributes that define who the character is. This includes the character's name, origin, birthplace, nationality, species or race, as well as formal titles, ranks, aliases, nicknames, professions, and social or institutional affiliations such as organizations, teams, clans, or factions. These attributes typically serve as the core identifiers of a character and are often explicitly stated in encyclopedic sources.

Appearance Appearance captures the character's physical and visual traits, including body features, facial characteristics, clothing style, colors, accessories, and other distinctive visual elements. This category focuses on how the character looks, and is commonly used in visual media such as animation, films, and games.

Ability Ability describes the character's competencies and capabilities, including skills, powers, techniques, magical abilities, combat styles, and special talents. This category encompasses both innate abilities and learned skills, and reflects what the character is capable of doing in practice, especially in action-driven narratives.

Relationship Relationship covers the character's social connections and interpersonal ties, such as family members, friends, allies, rivals, enemies, romantic partners, or other significant associations. These relationships define the character's position within a social network and are central to narrative development and character interaction, a distinction commonly emphasized in character-centric narrative analysis.

Experience Experience represents the character's past actions and life events, including important battles, missions, training experiences, achievements, failures, and major story arcs. Unlike iden-

tity, which is relatively static, experience is event-driven and reflects what the character has done over time, contributing to narrative progression and character growth.

Personality Personality characterizes the character's internal traits and behavioral tendencies, such as temperament, motivations, desires, goals, attitudes, emotional patterns, moral stance, and habitual ways of thinking or acting. These attributes are often implicit and inferred from behavior or dialogue rather than explicitly stated, and have been widely studied in computational character modeling.

Other Other is a residual category for miscellaneous information that does not fit the above dimensions, such as personal preferences, hobbies, or trivial habits.

E.2 Role-playing Evaluation Dimensions

We adopt four dimensions to evaluate the quality of role-playing dialogues, focusing on different aspects of character fidelity and interaction quality.

In-Character Consistency This dimension evaluates whether the agent consistently maintains the character's identity throughout the conversation, without breaking role or exhibiting meta-awareness.

Personality Faithfulness This dimension assesses whether the agent's tone, emotional style, and communication manner accurately reflect the character's personality traits.

Background Alignment This dimension examines whether statements about the character's life events, relationships, and abilities are consistent and free of contradictions within the dialogue.

Interaction Quality This dimension measures whether the agent's responses are contextually appropriate, coherent, and directly responsive to the user's inputs.

F Additional Analysis

F.1 Ablation Studies

To examine the contribution of iterative search in character profile construction, we conduct an ablation study by removing the iterative search mechanism while keeping all other components unchanged, as shown in Table 6.

Iterative Search benefits profile quality. As shown in Table 6, removing the iterative search mechanism leads to a consistent drop in Support Score across all knowledge dimensions. The overall Support Score decreases markedly, indicating that a single-round search fails to sufficiently cover diverse and fine-grained character attributes. This confirms that iterative querying and self-refinement are essential for expanding profile coverage and reducing missing information.

The performance degradation is more pronounced on dimensions that require richer contextual evidence. When iterative search is disabled, dimensions such as *Ability*, *Experience*, and *Relationship* exhibit larger declines compared to more factual attributes like *identity* or *appearance*. This suggests that action-oriented, narrative-dependent knowledge benefits more from multi-round evidence aggregation, whereas surface-level attributes can often be captured with fewer search steps. In contrast, more stable attributes such as Personality and Identity show relatively smaller performance declines.

F.2 Reverse Evaluation Analysis

To further address the concern regarding the comparison with Fandom as a reference source, we conduct an additional **reverse evaluation (RE)** analysis to examine the bidirectional alignment between DRInfo and Fandom.

F.2.1 Experimental Settings

In our original evaluation setting (denoted as **FE**), we extract atomic knowledge points from Fandom and evaluate whether the information in DRInfo *supports*, *partially supports*, *is irrelevant to*, or *contradicts* these reference knowledge points. Based on these alignments, we compute the Support Score (SS), Irrelevant Rate (IR), and Contradiction Rate (CR), as defined in Section 4.3.

In the reverse setting (**RE**), we switch the roles of the two sources: knowledge points are extracted from DRInfo and then aligned against the corresponding Fandom text. This reverse evaluation allows us to examine whether the information contained in DRInfo is also covered by Fandom and to analyze the relationship between the two knowledge sources from both directions. Table 7 reports the results across different knowledge categories.

F.2.2 Results and Analysis

Several observations can be drawn from the results.

Type	SS (%) ↑		IR (%) ↓		CR (%) ↓	
	Full	w.o. IS	Full	w.o. IS	Full	w.o. IS
Overall	66.54	64.15 (↓2.39)	22.08	24.22 (↑2.14)	2.55	2.23 (↓0.32)
Personality	81.67	80.78 (↓0.89)	11.78	12.22 (↑0.44)	0.87	0.83 (↓0.04)
Ability	61.37	57.48 (↓3.89)	28.69	32.42 (↑3.73)	1.47	1.78 (↑0.31)
Experience	60.42	57.80 (↓2.62)	27.02	30.23 (↑3.21)	5.28	4.08 (↓1.20)
Relationship	68.32	65.63 (↓2.69)	19.68	21.88 (↑2.20)	3.06	2.60 (↓0.46)
Identity	74.66	72.32 (↓2.34)	11.80	13.31 (↑1.51)	2.77	2.83 (↑0.06)
Appearance	57.62	55.94 (↓1.68)	23.25	24.19 (↑0.94)	3.06	3.08 (↑0.02)
Other	39.72	37.29 (↓2.43)	50.31	53.71 (↑3.40)	2.40	2.11 (↓0.29)

Table 6: Ablation study on iterative search. "Full" denotes the complete pipeline with iterative search, while "w.o. IS" removes the iterative search step. Relative changes compared to the full pipeline are shown in parentheses.

Type	SS_{FE}	SS_{RE}	IR_{FE}	IR_{RE}	CR_{FE}	CR_{RE}
Personality	81.67	86.76	11.78	9.29	0.87	0.26
Experience	60.42	73.50	27.02	19.21	5.28	2.55
Relationship	68.32	77.65	19.68	13.57	3.06	1.12
Ability	61.37	83.77	28.69	10.94	1.47	0.69
Identity	74.66	86.50	11.80	8.52	2.77	1.09
Appearance	57.62	81.35	23.25	8.38	3.06	1.04
Other	39.72	72.42	50.31	18.59	2.38	3.12

Table 7: Bidirectional knowledge alignment results between Fandom and DRInfo. FE denotes the original evaluation where Fandom knowledge points are aligned against DRInfo, while RE denotes the reverse evaluation where DRInfo knowledge points are aligned against Fandom.

First, the reverse evaluation consistently achieves higher Support Scores and lower Irrelevant and Contradiction Rates across most knowledge categories. This indicates that Fandom provides **broad coverage of character information** and contains most of the factual knowledge present in DRInfo.

Second, the lower support scores observed in the FE setting suggest that DRInfo does not attempt to exhaustively replicate the entire Fandom content. Instead, it performs a **selective and condensed integration of character knowledge** collected from multiple sources during the deep search process.

Finally, contradiction rates remain very low in both directions, demonstrating that DRInfo maintains strong **factual consistency** with the Fandom reference rather than introducing conflicting information.

Overall, the bidirectional evaluation reveals complementary characteristics between the two sources: Fandom provides extensive coverage of character knowledge, while DRInfo focuses on integrating and condensing relevant information into a more compact representation.

F.3 Case Study

To better understand the failure modes of automated character profile construction and evaluation, we conduct a case study focusing on knowledge points that are judged as Contradicted by the LLM-based evaluator. We choose this subset for analysis because contradicted cases represent the most explicit and verifiable form of errors, where the extracted or generated knowledge directly conflicts with canonical evidence, making them particularly suitable for qualitative inspection. By examining these samples in detail and comparing LLM judgments with human annotations, we aim to both find common reasons of factual inconsistency in character profiling and assess the reliability and limitations of LLM-as-a-Judge in identifying such errors.

F.3.1 Contradicted Types

From those cases in which both the LLM judge and human annotators consistently identify knowledge points as *Contradicted*, we observe five contradiction types: (1) *Cross-version Confusion* refers to contradictions caused by mixing information across different canons, adaptations, spin-offs, or alter-

Knowledge Point	LLM Evidence	Contradiction Type
Haruhi Suzumiya (Nyoron! Churuya-san): Haruhi created an alternate world due to her dissatisfaction with reality.	She does not create alternate worlds in this series; the world remains static and mundane.	Cross-version Confusion
Illyasviel von Einzbern (Carnival Phantasm): She was attacked by Zouken, leading to Berserker's corruption.	In this version, Berserker remains under Illya's control.	
Jousuke Higashikata (JoJolion): He exploits Crazy Diamond's restoration ability for tactical advantage.	His Stand is Soft & Wet, not Crazy Diamond.	Ability Misassignment
Ein (Cowboy Bebop): Ein is a Cardigan Welsh Corgi.	The text explicitly identifies Ein as a Pembroke Welsh Corgi.	Objective Attribute Error
Yuu Ishigami (Kaguya-sama): He is the Student Council Vice President and former Treasurer.	The text states he is the treasurer, not Vice President.	Temporal Inconsistency
Lelouch Lamperouge (Code Geass): He resurrects after death and returns to protect the world.	The text presents his death as final; no resurrection is mentioned.	Fabricated Narrative Event

Table 8: Representative **contradicted** knowledge points where LLM-judge and human annotations agree. Each row is separated by horizontal rules for clarity, and all entries are vertically centered.

nate versions of the same character, where events or attributes valid in one version are incorrectly projected onto another. (2) *Ability Misassignment* occurs when a character is attributed with abilities or powers belonging to a different character or continuity. (3) *Objective Attribute Error* captures fine-grained but verifiable factual mistakes, such as incorrect species, titles, or fixed descriptors. (4) *Temporal Inconsistency* captures cases where a knowledge point is factually correct in a later narrative stage but conflicts with the profile due to an implicit mismatch in temporal scope. (5) *Fabricated Narrative Event* denotes hallucinated plot developments that directly contradict canonical story outcomes.

As illustrated in Table 8, these contradiction types manifest as clear, localized conflicts with fandom reference. Taking *Cross-version Confusion* as an example, *Haruhi Suzumiya* is described as having created an alternate world due to her dissatisfaction with reality; however, this event occurs in other derivative or alternate-universe narratives and does not take place in *Nyoron! Churuya-san*. In the case of *Yuu Ishigami*, he does eventually become the Student Council Vice President later in the narrative; however, the generated profile only includes his present role as Treasurer.

F.3.2 Bias of LLM-Judge

From those cases where the LLM judge labels as *Contradicted* while human annotators assign a different label, we conclude 3 major bias types exhibited by the LLM judge. (1) *Absence-as-Contradiction Bias*, where missing or unspecified information in the profile is incorrectly treated as a factual contradiction. (2) *Temporal Misinterpretation Bias*, where the LLM fails to distinguish between current, future, or alternative timeline facts. (3) *External Lore Intrusion Bias*, where the LLM relies on external or non-local canon knowledge not present in the evaluated profile.

As shown in Table 9, a dominant source of disagreement is *Absence-as-Contradiction Bias*, arising from the LLM judge's tendency to treat missing information as explicit contradiction. For instance, in the cases of *Hei* and *Rei Kiriyama*, the profiles simply omit the described events, which human annotators correctly label as Irrelevant, whereas the LLM judge interprets absence as negation.

G Prompt Templates

Specifically, Table 10 present the prompts for the iterative search stage, where the first prompt initializes diversified character-oriented queries and the second prompt refines subsequent searches by incorporating summarized evidence from the previous round. Table 11 shows the prompt used for profile rewriting, which integrates multi-source ev-

Knowledge Point	LLM Evidence (Excerpt)	Human Label	Bias Type
Hei (Darker than Black): Hei struggled with the moral conflict of potentially having to kill his sister Bai.	No indication he ever considered or was required to kill her.	Irrelevant	Absence-as-Contradiction
Rei Kiriyama (March Comes in Like a Lion): He lacks common sense, shown by attempting to register his engagement.	The text does not mention any attempt to register an engagement or lack of common sense related to legal procedures.	Irrelevant	
Nagisa Furukawa (Clannad): She has two M-shaped ahoge inherited from both parents.	Ahoge is described, but no M-shape or inheritance is mentioned.	Partially Supported	
Hiroto Suwa (Orange): In an alternate timeline, Hiroto marries Naho and has a child.	Marriage occurs in the original timeline, not the alternate one.	Supported	Temporal Misinterpretation
Asuka Langley Souryuu (Evangelion 2.0): Her mother committed suicide after an EVA experiment.	Rebuild version presents a different backstory.	Irrelevant	External Lore Intrusion
Princess Leia (Star Wars): She survives in space using the Force.	Such an ability is neither mentioned nor canonically supported.	Irrelevant	

Table 9: Representative cases where the LLM judge marks knowledge points as **Contradicted** while human annotators assign other labels. The disagreements reveal systematic LLM judge biases.

First Search Prompt: Please help me collect all relevant knowledge about `{{entity}}` and integrate it into a detailed encyclopedic document. Use markdown format. Be truthful, comprehensive, and thorough, without omitting important information. You must verify the authenticity of information and the credibility of sources. Do not include or fabricate false information. You should include the character's personality (very important), background, physical description, core motivations, notable attributes, relationships, key experiences, major plot involvement and key decisions or actions, character arc or development throughout the story, if there is any information about these aspects.
Iterative Search Prompt: Please continue searching to verify and expand on the knowledge you have already collected, such as searching for closely related people, events, objects, and other entities.

Table 10: First-round and iterative search prompt used in the CHARACTERHUB pipeline.

idence into a coherent first-person character self-introduction tailored for role-playing agents. For evaluation, Table 12 details the prompt for extracting fine-grained knowledge points from fandom reference, while Table 13 presents the LLM-as-a-Judge prompt used to assess the alignment between individual knowledge points and the generated profiles. Together, these prompts operationalize the full automatic pipeline from information acquisition to quantitative profile evaluation.

Prompt: Please completely rewrite all the above information from `{{entity}}`'s first-person perspective. Ensure that the information is comprehensive and accurate. You need to focus on the character's personality, which you could also analyze based on the characters' experiences. Besides, include include the character's, background, physical description, core motivations, notable attributes, relationships, key experiences, major plot involvement and key decisions or actions, character arc or development throughout the story, and other important details, if they appear in the information that you obtained.

Table 11: Prompt for rewriting refined evidence into a first-person character profile.

Prompt: You will receive a piece of character data (in JSON format) crawled from the Fandom website. The data includes the character's basic information (infobox) and textual descriptions (content).
Please parse the input JSON and extract all **knowledge items** related to the character, and focus on *meaningful*, *distinctive*, and *role-relevant* information that contributes to understanding or role-playing this character, labeling each item with its type.
Each knowledge item should represent a semantically complete factual statement or a clearly defined attribute. For example: "He is a passionate young man determined to XXX" or "He is humanity's strongest soldier."
Output the results in a unified structured format.

Please IGNORE any information that is irrelevant to character role-playing, including:
- Field echo information that only restates a label or name (e.g. "His name is XXX.").
- Voice actors, performers, production companies, publishers, episode or chapter numbers (e.g. "voiced by ...").
- Language, romanization, or spelling variants (e.g. Kanji, Romaji, kana notation).
- Generic or obvious facts that do not distinguish this character from ordinary people (e.g. "He is human" or "He is male"), unless they have clear narrative or psychological significance (e.g. "He is human but can transform into a giant"—at this point, "human" information has semantic value).
- Statistical or external information (e.g. number of appearances, popularity rankings, fan titles).
- External commentary or meta descriptions about the character (e.g. "first appeared in episode X").
- Any meta-level information not belonging to the character's in-universe context.

Output Format
Please output your result in JSON format as follows:

```
{
  "entity": "Character Name",
  "knowledge_points": [
    {
      "type": "identity | appearance | ability | relationship | experience | personality | other",
      "evidence": "Original supporting text or sentence from the input",
      "knowledge": "Extracted knowledge statement"
    },
    ...
  ]
}
```

Rules
1. Each knowledge item should be semantically clear, self-contained, and useful for **understanding or role-playing the character**.
2. Merge and deduplicate overlapping or equivalent facts. Keep only one representative version of each unique fact. Do not output two items that share the same meaning even if worded differently. If two statements express similar meaning, combine them into one richer description.
3. **Do NOT invent or infer** beyond the text. If something is only hinted at or ambiguous, exclude it.
4. Ensure the output is valid JSON with complete and logically consistent fields.
5. If certain categories are not present, you may omit them — do not fill with "unknown".
6. You must only use the following categories for `type`:
- identity: factual identity, origin, birthplace, nationality, species, titles, nicknames, ranks, aliases, professions, affiliations with organizations, teams, clans, groups, and roles or statuses associated with them.
- appearance: physical traits, looks, body, clothing, colors, and distinctive visual features.
- ability: skills, powers, techniques, magic, fighting capability, special talents, or combat style.
- relationship: family members, friends, rivals, lovers, allies, enemies, or any significant social or character connections.
- experience: key life events, battles, training, missions, achievements, historical background, or things the character has done.
- personality: temperament, behavior patterns, motivations, desires, goals, attitudes, emotional tendencies, moral stance, or thinking style.
- other: anything that does not reasonably fit into the above categories. Do **not** create or use any other category.

Input Data:
`{{input text}}`

Table 12: Prompt for extracting structured knowledge points from reference character descriptions.

Prompt: You are an expert evaluator for role-playing character consistency.
Task
You will receive two items:
1. `knowledge_list`: a JSON array of knowledge objects about a character, each describing a factual aspect of the character. Each object has:
- "id": a unique integer identifier for the knowledge item,
- "knowledge": a concise factual statement about the character.
2. `text`: a piece of character-related text. This may be:
- a first-person profile written from the character's perspective,
- a third-person descriptive text about the character,
- or structured JSON data crawled from a source such as Fandom.
Treat all three forms uniformly as sources of factual claims.

Your goal is to determine, for each knowledge item in `knowledge_list`, determine whether the provided `text`:
- **supported**: The text explicitly expresses or clearly implies this knowledge point.
- **partially supported**: The text supports some but not all aspects of the knowledge point (e.g., mentions part of a list or partial causation).
- **irrelevant**: The text does not mention or relate to this knowledge point.
- **contradicted**: The text states or implies something that conflicts with this knowledge point.
Do NOT use any labels other than the four above.

Important Rules
1. Focus on **semantic meaning**, not wording. Example: if the knowledge says "He is calm and collected" and the text says "I rarely lose my temper," that is **supported**.
2. A **contradicted** label should be used only if the text explicitly or implicitly rejects or reverses something of the fact.
3. Missing details or partial lack of coverage should be labeled **irrelevant**, not contradicted.
4. **Do NOT treat narrative perspective or tense as evidence.** Do NOT interpret first-person narration ("I") or present tense ("I am", "I do") as evidence that the character is alive or active in the current timeline. Only mark a contradiction if the text explicitly denies or reverses the fact.
5. When uncertain, choose the most reasonable interpretation based on the content of the given text.
6. For each evaluation, provide a brief quote or reasoning from the text that justifies your judgment.

Example
knowledge_list:

```
[
  {"id": 1, "knowledge": "Eren was born in the Shiganshina District."},
  {"id": 2, "knowledge": "Eren is calm and patient."}
]
```

text:

```
"I grew up behind the walls of Shiganshina, dreaming of the world beyond.
I've always been reckless and driven by anger — patience was never my strength.
But since that day, I've carried the Titan power within me."
```

Expected Output:

```
[
  {
    "id": 1,
    "knowledge": "Eren was born in the Shiganshina District.",
    "evaluation": "supported",
    "evidence": "He says he grew up in Shiganshina."
  },
  {
    "id": 2,
    "knowledge": "Eren is calm and patient.",
    "evaluation": "contradicted",
    "evidence": "He describes himself as reckless and impatient."
  }
]
```

Input
`knowledge_list: {{knowledge_list}}`

`text: {{character_text}}`

Table 13: Prompt for knowledge–profile alignment using an LLM-as-a-Judge framework.