Identification and Analysis of Identity-Centric Elements of Character-Likeness from Game Scenario

Shinji Iwata[1], Koya Ihara[1], Shiki Sato[1], Jun Baba[1], Asahi Hentona[1], Masahiro Yamazaki[2], Yuki Shiotsuka[2], Takahiro Ishizue[3], Akifumi Yoshimoto[1]
[1]CyberAgent, [2]QualiArts, [3]Sumzap
{iwata_shinji, ihara_koya, sato_shiki,
baba_jun, hentona_asahi, yoshimoto_akifumi}@cyberagent.co.jp
{yamazaki_masahiro, shiotsuka_yuki}@qualiarts.jp
{ishizue_takahiro}@sumzap.co.jp

Abstract

Generating and evaluating character-like utterances automatically is essential for applications ranging from character simulation to creative-writing support. Existing approaches primarily focus on basic aspects of character-likeness, such as script-fidelity knowledge and conversational ability. However, achieving a higher level of character-likeness in utterance generation and evaluation requires consideration of the character's identity, which deeply reflects the character's inner self. To bridge this gap, we identified a set of identity-centric character-likeness elements. First, we listed 27 elements covering various aspects of identity, drawing on psychology and identity theory. Then, to clarify the features of each element, we collected utterances annotated with these elements from a commercial smartphone game and analyzed them based on user evaluations regarding character-likeness and charm. Our analysis reveals part of element-wise effects on character-likeness and charm. These findings enable developers to design practical and interpretable element-feature-aware generation methods and evaluation metrics for character-like utterances.

1 Introduction

The automatic generation and evaluation of character-like utterances are essential for applications ranging from role-playing dialogue agents (Han et al., 2022; Shao et al., 2023) to creative-writing support tools (Mirowski et al., 2023; Hoque et al., 2023). Users ultimately perceive character-likeness through observable behaviors. According to the Cognitive-Affective Processing System (CAPS) framework (Mischel and Shoda, 1995), behaviors emerge when external stimuli are processed through the internal state of an individual.

The existing study has taken the first step by defining the elements of character-likeness rooted in observable behaviors. These sets are typically validated against mainly narrative scripts or background settings (Wang et al., 2024b; Lu et al., 2024; Wang et al., 2024a; Tu et al., 2024; Wu et al., 2025). Using behavior-based evaluation methods, an automatic framework can be developed for generating or evaluating character-like behaviors for which gold data are available. However, collecting data on all possible behaviors of a character is impractical; this limits the generalizability of

Intrapersonal Elements
Personality(Cattell, 1946):
Temperamental Traits (IP-Tem.),
Dynamic Traits (IP-Dyn.), Ability Traits (IP-Abi.)
Self-Concept(Kajita, 2020):
Recognition and Definition of Current Self (IS-Rec.),
Affects and Evaluations of Self (IS-Aff.),
Possibilities and Future Imagery of Self (IS-Pos.),
Imagery of the Past Self (IS-Ima.),
Others' Perceptions of Self (IS-Oth.),
Oughtness and Ideals of Self (IS-Oug.)
Other:
Profile (IO-Pro.), Wording (IO-Wor.)
Two-person Relational Elements
Objective Relationships (T-Obj.),
Subjective Relationships (T-Sub.),
Habitual Attitudes toward Others (T-Hab.)
Group Relational Elements
Role Identity(Kaplan and Garner, 2017):
Ontological and Epistemological Beliefs (GR-Ont.),
Purpose and Goals (GR-Pur.),
Self-Perceptions and Self-Definitions (GR-Sel.),
Perceived Action Possibilities (GR-Per.)
Cultural Identity(Cristina, 2016):
Ethnicity (GC-Eth.), Nationality (GC-Nat.),
Place of Birth (GC-Pla.), Age (GC-Age.),
Sex (GC-Sex.), Socioeconomic Status (GC-Soc.),
Language (GC-Lan.), Religion (GC-Rel.),
Cultural Heritage (GC-Cul.)

Table 1: Overview of an IDentity-aware taxOnomy of 27 eLements (IDOL-27). The IDOL-27 covers various aspects of the inner states of characters, grounded in a pilot annotation study and a wide-ranging survey of the psychological literature on self-concepts and group identity.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 275–284
Aug 25–27, 2025.

the framework. On the other hand, character-like actions can be generated even in unseen contexts from an inner-state-aware model that has the character's internal model of the character. Thus, an internal model that captures the principles underlying the character's behavior needs to be constructed.

To consider the inner state of a character, we consider identity, a psychological construct that organizes the self (Erikson, 1968; Baumeister, 1987; McAdams et al., 2021) across multiple perspectives such as role, culture, and self-concept. The multifaceted nature of identity allows us to isolate internal elements along interpretable axes.

Building on this insight, we constructed an IDentity-aware taxOnomy of 27 eLements (IDOL-27) of character-likeness, as presented in Table 1, and empirically examined its properties. First, based on prior work, we enumerated elements at the intrapersonal, two-person, and group levels. Next, we collected utterances associated with each element from the smartphone game *IDOLY PRIDE*[1]. Finally, we asked the participants to evaluate the character-likeness of these utterances and analyzed the results. Because user-perceived character-likeness was the primary objective of system development, all the analyses adopted this perspective.

Our analysis revealed two key findings. First, we uncovered two classes of elements: (i) elements that consistently enhanced character-likeness and (ii) elements whose impact was strongly context-dependent. Second, we found minor character- and examinee-level deviations from the overall trend, suggesting that these tendencies could be general patterns.

Our contributions are threefold. (i) **Element Set**: We constructed a set of 27 identity-aware elements that model the inner state of a character. (ii) **Validated Properties**: We empirically validated the properties of each element through a user study. (iii) **Generality**: We demonstrated the consistency of these elements across characters and users.

2 Related Work

Character-like Utterance Generation. Recent studies have addressed the generation of character-like utterances from both dataset and model perspectives. These studies included building character-like dialogue systems from a few gold utterances (Han et al., 2022); focusing on the persona of a character (Shao et al., 2023; Peng and Shang, 2024; Wang et al., 2024c); mitigating hallucinations about the knowledge of a character (Lu et al., 2024; Sadeq et al., 2024; Tang et al., 2025); modeling temporal and situational dynamics (Ahn et al., 2024; Wu et al., 2024; Bae et al., 2025). The IDOL-27 introduced in this study contributes new insights to the current body of work; for example, it enables us to (i) identify elements that previous studies have overlooked and (ii) verify whether the generated outputs faithfully incorporate the properties of each element.

Elements of Character-Likeness. Prior work on character-like dialogue systems released datasets along with the human annotations for behavior-based metrics. Wang et al. (2024b) and Lu et al. (2024) released the Character100 and WikiROLE datasets, respectively, built from Wikipedia biographies. Wang et al. (2024a) constructed the RoleBench dataset that covered 100 distinct roles. Tu et al. (2024) constructed the CharacterEval dataset and proposed four evaluation dimensions: Conversational Ability, Role-Playing Attractiveness, Character Consistency, and Personality Back-Testing. Wu et al. (2025) constructed the RAIDEN dataset, a multi-turn dialogue benchmark. This dataset targeted self-awareness that could be further operationalized as Script-Fidelity Knowledge, Role-Cognition Boundary, Persona Language Style, and Conversational Ability. Prior work has tended to concentrate on general conversational competence and behavior-oriented elements such as consistency with narrative knowledge or settings that did not focused on the inner states of characters. Certain studies evaluated the personality of inner character using personality traits (e.g., Big Five and MBTI). Although convenient for broad classification, these taxonomies were built to group people and not encode subtle character-specific nuances. Detailed character-specific descriptions such as self-perceptions and behaviors within particular groups are necessary for an automatic framework to generate or evaluate the character-like behaviors. Our study addresses this gap by constructing elements grounded in identity, allowing for character-specific descriptions of the inner states of characters.

[1] https://idolypride.pmang.cloud/en-us/

3 List of Character-Likeness Elements

To encompass the internal states of various characters, we introduce an IDentity-aware taxOnomy of 27 eLements (IDOL-27). These elements are grounded in a pilot annotation study and a wide-ranging survey of the psychological literature on self-concepts and group identity. These elements encompass a broad spectrum of the inner attributes of characters, including personality traits, self-concepts, two-person relations, and group-level identities (e.g., role and culture). Table 1 lists the 27 elements, organized into three categories: (i) intrapersonal that represents the self-contained mechanism; (ii) two-person that captures the dyadic relationship with a specific other; (iii) group that concerns the membership of the character in groups. These three categories correspond to the categories reviewed by Brewer and Gardner (1996) in their study of the social self.

3.1 Intrapersonal Elements

Intrapersonal can be grouped into three categories: (i) personality that functions as a self-contained response mechanism; (ii) self-concept that defines personal identity (the same categories are used by Kiel et al. (2024)); (iii) other, such as wording and profile.

For personality, we adopt Cattell (1946)'s taxonomy, comprising temperamental, dynamic, and ability traits. According to Mizuta (1995), Temperamental Traits are relatively stable characteristics, Dynamic Traits are motivational elements, and Ability Traits reflect individual differences in perceptual and motor abilities.

Following Kajita (2020), we incorporate six self-concept elements. Self-concept is a long-standing psychological construct describing the way individuals perceive themselves.

To include aspects that are outside of personality and self-concept, the *other* category comprises two elements: (i) Profile that includes background attributes such as favorite foods and (ii) Wording that includes tone, phrasing, and other stylistic choices.

3.2 Two-Person Relational Elements

We define two elements at the two-person level, Subjective Relationships and Habitual Attitudes toward Others, that correspond to the cognitions/emotions and behavioral facets described by Kelley et al. (1983), respectively. In addition, we introduce Objective Relationships as an element.

Subjective Relationships capture a character's internal view of the other person. Habitual Attitudes toward Others are actions typically exhibited in interactions with specific individuals. To include aspects that are outside of these elements, we add Objective Relationships that are socially verifiable ties between two individuals that can be confirmed by a third party (e.g., parent–child and sibling relationships).

3.3 Group Relational Elements

Stets and Burke (2000) identify category/group and role as the two foundational bases of identity, reflecting our treatment of group relations. We partition the group dimension into two categories: *cultural identity*, representing the category/group basis, and *role identity*, representing the role basis. Specifically, we adopt nine cultural-identity elements proposed by Cristina (2016) and four role identity elements from Kaplan and Garner (2017).

4 Experiment

In the previous section, we presented the proposed set of identity-aware character-likeness elements. We identified the properties of each element by investigating the way users perceived character-likeness and charm in the utterances featuring these individual elements. In this section, we describe the details of the data collection method used in the investigation.

First, we collected utterances from in-game characters, each featuring a specific element (hereafter, stimulus utterances). Subsequently, we allowed human examinees familiar with these characters to subjectively evaluate each collected utterance in terms of perceived character-likeness and charm.

Characters. We selected five game characters for the analysis, selecting the combination that maximized the pairwise distances between their profile representations in the sentence embedding space. This approach allowed us to efficiently analyze a diverse set of characters. We finally selected Nagisa Ibuki, Haruko Saeki, Rui Tendo, Sakura Kawasaki, and Sumire Okuyama.

Stimulus Utterances. We used the in-game utterances of the aforementioned five characters featuring each element in the IDOL-27. To build a reliable stimulus set, we collected stimulus utter-

ances as follows. First, the first author labeled every utterance with the element that the utterance appeared to feature most saliently. Next, we requested two additional annotators to relabel across several rounds using a codebook containing element definitions, example utterances, and feedback from a pilot annotation. When at least two of the three annotators selected the same element as the most salient feature of the utterances, we extracted that utterance for the analysis of that element. [2]

Examinees. Thirteen examinees were recruited for this study. Each examinee first read the story and a summary of the game and was allowed to proceed only after passing a comprehension test on the plot and characters of the game.

Evaluation Method. We requested the examinees to rate each utterance using a visual analog scale with 101 increments (0 = "does not apply at all" and 100 = "applies perfectly") for both character-likeness and charm, and then provide free-form rationales explaining their ratings.

5 Results

In this section, we analyze the properties of each element based on the evaluation scores collected from the examinees, as described in the previous section, and their comments regarding the rationale for their evaluations. First, to obtain an overview of the way the examinees evaluated utterances featuring each element, we examined basic statistical measures across all elements (Section 5.1) and subsequently clustered these elements based on these measures (Section 5.2). Then, we decomposed these statistics by character (Section 5.3) and by examinees (Section 5.4) to investigate the properties of each element in more detail.

5.1 Overview of Examinee Rating

To obtain an overview of each element, we first computed its summary statistics. For each element featuring at least 15 evaluations, we calculated the mean (M) and sample standard deviation (SD) of the scores assigned by the examinees to the utterances, as presented in Table 2.

[2]During annotation, more precisely, annotators selected the single most salient element along with several less prominent elements when multiple elements were featured in a single utterance. Overall, we obtained overlapping labels from at least two annotators for 95% of the utterances, indicating a substantial agreement.

Here, because the neutral evaluation point for both character-likeness and charm was set at 50, if utterances featuring a particular element frequently exceeded this threshold, we interpreted that these utterances generally tended to evoke character-likeness and charm in directions aligned with user expectations. Conversely, utterances scoring below 50 featured the element in directions contrary to user expectations, potentially impairing its perceived character-likeness and charm. Hence, ensuring that featuring those elements did not contradict user expectations, is crucial.

Because the mean scores for all elements exceeded 50, which is the neutral point, we confirmed that utterances featuring any element of the IDOL-27 generally gave a positive impression of character-likeness and charm.

5.2 Clusters of Elements

We clustered the elements, as presented in Table 2, into a two-dimensional (M, SD) space. The silhouette coefficient suggested that the optimal number of clusters was $k = 2$ for character-likeness and $k = 4$ for charm.

5.2.1 Clusters for Character-Likeness

C1: low M and large SD. This cluster had a low average and a large variance, suggesting that it could either enhance or undermine character-likeness. It included Dynamic Traits, Ability Traits, Wording, Habitual Attitudes toward Others, Place of Birth, and Age. For example, 17% of the scores for Place of Birth and 9% for Wording were <50 (negative effects). By contrast, the most frequent <50 rating in C2 (see below) was 7% for Habitual Attitudes toward Others, indicating that C1 had a higher proportion of negative effects.

C2: high M and small SD. With a high mean and small variance, this cluster captured elements that almost universally enhanced character-likeness. It comprised the Temperamental Traits, Affects and Evaluations of Self, Imagery of the Past Self, Oughtness and Ideals of Self, Profile, Subjective Relationships, Purpose and Goals, Self-Perceptions and Self-Definitions, and Perceived Action Possibilities. Identity-related elements such as self-concept and role identity were concentrated in this positively perceived cluster.

5.2.2 Clusters for Charm

C1: lowest M and largest SD. This cluster exhibited the lowest mean and the largest variance of

	N	Character-Likeness	Charm
Intrapersonal Elements			
Personality			
Temperamental Traits (IP-Tem.)	150	86.7 ±17.3	81.3 ±22.4
Dynamic Traits (IP-Dyn.)	147	79.3 ±22.0	74.2 ±23.4
Ability Traits (IP-Abi.)	25	82.9 ±22.7	76.2 ±24.1
Self-Concept			
Affects and Evaluations of Self (IS-Aff.)	95	83.5 ±18.5	67.6 ±26.5
Imagery of the Past Self (IS-Ima.)	80	85.8 ±18.9	71.7 ±29.5
Oughtness and Ideals of Self (IS-Oug.)	25	88.4 ±10.0	79.6 ±21.1
Other			
Profile (IO-Pro.)	50	83.4 ±14.3	76.0 ±21.5
Wording (IO-Wor.)	150	79.0 ±22.8	76.8 ±23.3
Two-person Relational Elements			
Subjective Relationships (T-Sub.)	149	85.1 ±17.0	81.6 ±21.5
Habitual Attitudes toward Others (T-Hab.)	59	79.2 ±20.0	75.6 ±19.7
Group Relational Elements			
Role Identity			
Purpose and Goals (GR-Pur.)	25	89.8 ±14.4	89.6 ±13.2
Self-Perceptions and Self-Definitions (GR-Sel.)	60	86.8 ±19.8	82.8 ±22.8
Perceived Action Possibilities (GR-Per.)	74	84.3 ±20.3	82.0 ±20.9
Cultural Identity			
Place of Birth (GC-Pla.)	35	71.7 ±26.0	62.1 ±27.6
Age (GC-Age.)	59	76.9 ±20.7	70.7 ±25.5

Table 2: Mean ± SD for each element (N: number of scores).

the four clusters, implying the strongest potential to reduce charm and the weakest capacity to enhance it. It comprised Affects and Evaluations of Self, Imagery of the Past Self, Place of Birth, and Age.

C2: low M and large SD. Although the mean was low and the variance large, the elements in this cluster could either enhance or diminish charm. The cluster elements included Dynamic Traits, Ability Traits, Profile, Wording, and Habitual Attitudes toward Others. Only 7% of its ratings were below 50, compared with 18% in C1; therefore, its negative effect was considerably smaller.

C3: high M and small SD. The elements here showed a high mean and small variance, making this a consistently charming cluster. It included Temperamental Traits, Oughtness and Ideals of Self, Subjective Relationships, Self-Perceptions and Self-Definitions, and Perceived Action Possibilities.

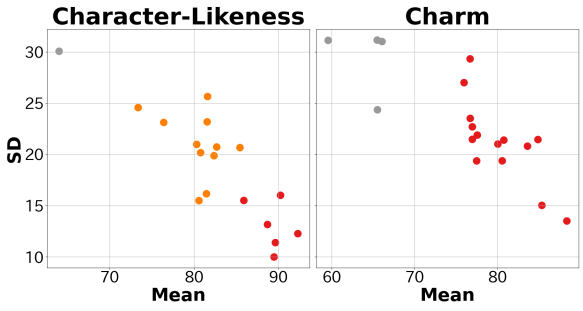

Figure 1: Clustered elements by character.

C4: highest M and smallest SD. With the highest mean and the smallest variance, C4 contributed the most to charm. A total of 36% of its ratings had a maximum score of 100, whereas C3 had only 16%. It contained only one element: Purpose and Goals.

5.3 Element-Level Analysis by Character

To examine the way the element-level properties varied across characters, we compared the ele-

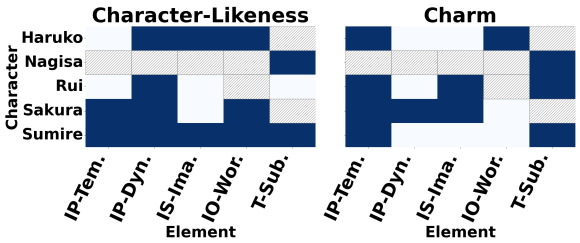

Figure 2: Heat map indicating whether the element-level properties of each character match those described in Section 5.2. Dark blue denotes a match; light blue, a mismatch; hatched cells indicate that the element is absent.

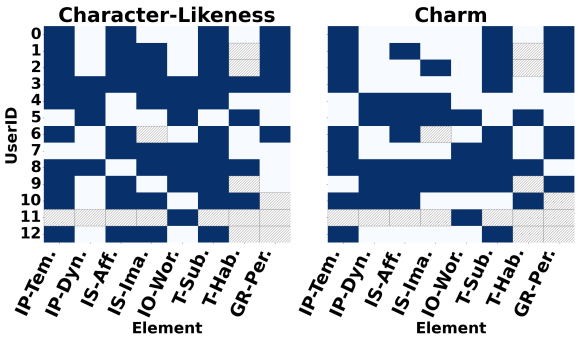

Figure 3: Heat map indicating whether the element-level properties of each user match those described in Section 5.2. Dark blue denotes a match; light blue, a mismatch; hatched cells indicate that the element is absent.

ments of each character with the global clusters reported in Section 5.2. Using the same clustering procedure, we clustered the elements of each character, including only elements that received at least 15 evaluations and were evaluated for three or more characters (Figure 1). Three clusters emerged for character-likeness: (i) extremely low mean and extremely large variance; (ii) low mean and large variance; (iii) high mean and small variance. Two clusters emerged for charm: (i) low mean and large variance and (ii) high mean and small variance. To facilitate the comparison, we merged the cluster labels into two categories: high-mean/small-variance and low-mean/large-variance. Figure 2 presents the resulting category agreement matrices.

Regarding character-likeness, more than half of the elements for every character matched the category assigned in Section 5.2, suggesting that the element-level tendencies were largely consistent across characters. For charm, the elements Temperamental Traits, Imagery of the Past Self and Subjective Relationships followed the overall trend for most characters, whereas Dynamic Traits and Wording followed this trend for fewer than half the characters. These observations revealed substantial character-specific variations. In particular, the unusually low scores on Wording and Dynamic Traits for Haruko Saeki and Sakura Kawasaki appeared to have driven this divergence. Ratings of 50 (neutral) or lower accounted for 35% and 20% of the Wording responses and for 17% and 13% of the Dynamic Traits responses, respectively. These proportions suggested that, for these two characters, expressing user-perceiving charm through these elements was difficult.

Overall, the element-level features for character-likeness were generalized across characters, whereas several charm-related elements displayed character-specific dynamics.

5.4 Element-Level Analysis by Examinee

Using the same clustering and category-based procedures described in Section 5.3, we investigated the way the element-level tendencies varied across participants. We clustered scores that satisfied the following two conditions: each participant supplied at least five ratings for the element[3], and the element was evaluated by at least seven participants. Figure 3 shows the agreement between the elements.

Regarding character-likeness, more than half the participants assigned every element to the same class obtained in the overall analysis, indicating an overall tendency. For charm, a majority agreement was observed only for Temperamental Traits, Affects and Evaluations of Self, Subjective Relationships, and Perceived Action Possibilities. The remaining elements such as Dynamic Traits, Wording and Habitual Attitudes toward Others, whose intensity or content of expression were thought to be strongly context-dependent elements. Imagery of the Past Self was thought to be associated with negatively valenced descriptions. These elements showing disagreement suggested that the perceived charms of users was strongly influenced by their subjective interpretations beyond character settings and by their personal stance toward negative expressions.

In summary, inter-examinee variation was limited for character-likeness, whereas modeling charm required accounting for participant-specific

[3]We adopted a five-utterance threshold, as per-participant splits leave few utterances per element.

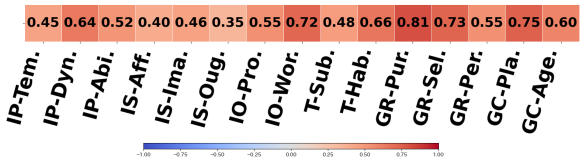

Figure 4: Pearson's r of character-likeness and charm. Positive correlations are observed for all elements.

traits and subjectivity, at least for the aforementioned elements.

5.5 Correlation of Character-Likeness and Charm

To investigate whether utterances that sound character-like were considered charming, we computed the Pearson's r between the two scores for each element (Figure 4). Although the strength of the relationship varied across elements, all r values were positive, indicating that higher character-likeness generally coincided with higher charm.

At the element level, features that could detract from charm, such as Wording and Dynamic Traits, still exhibited strong positive correlations. This pattern suggested that the magnitude of a correlation had to be interpreted independently of the element's properties.

Averaging r values within our category groups yielded the following results: Personality (0.54), Self-Concept (0.40), Other (0.63), Two-Person Relational (0.57), Role Identity (0.70), and Cultural Identity (0.67). Even when elements from different clusters were pooled into groups, we observed certain categories displayed a strong r value whereas others showed weaker values; notably, the group categories exhibited the strongest correlation.

6 Discussion

6.1 Causes of the Undermining Effect

We analyzed the free-form comments attached to the three elements that most frequently received scores below 50 to identify what undermined character-likeness or charm. Regarding character-likeness, the most negatively rated elements were Wording, Dynamic Traits, and a tie between Temperamental Traits and Place of Birth. For charm, the corresponding elements were Imagery of the Past Self, Wording, and Affects and Evaluations of Self. Below, we highlight the elements that revealed notable insights. The findings for character-likeness are as follows:

Wording. Comments such as "The character's use of onomatopoeia felt slightly off" and "The sudden overuse of Western language sounded unnatural" indicated that fine-grained linguistic choices (e.g., onomatopoeia and loanwords) could undermine the character-likeness.

Dynamic Traits. Participants remarked that the emotional intensity of certain lines conflicted with their expectations (e.g., "The character seems far too anxious" or "She is unrealistically upbeat").

Temperamental Traits. Remarks such as "I would expect the character to dislike that without wavering, so this line feels inconsistent" showed that deviations from the personality image the user held for the character were viewed negatively.

For charm, we observed systematic cues.

Imagery of the Past Self. Six of nine reasoned comments (approximately 67%) cited a negative tone (e.g., "Because the character speaks sadly, I wouldn't think it charming.")

Affects and Evaluations of Self. Five of twelve comments (approximately 42%) pointed to negative self-referential content, as in "The anxious remark is not appealing."

Overall, these findings suggested that maintaining character-likeness depended on careful lexical choices and consistency with user expectations, whereas conveying charm demanded special care whenever an utterance carried a negative affect, lest the character's charm could be impaired.

6.2 Character-Likeness vs. Charm

We examined the differences between character-likeness and charm from a system development perspective. Despite a moderate overall correlation between them, several elements scored highly on character-likeness yet poorly on charm, as revealed in Table 2. One explanation for this difference was evaluator subjectivity. As noted in Section 4, charm appeared to be influenced by subjective factors such as the personal preferences of a user.

To explore this possibility, we used GPT-4o (OpenAI, 2024) to automatically classify whether each free-text comment explicitly expressed the emotions of the examinee. After manual verification, we identified 12 character-likeness com-

ments and 52 charm-related comments that contained affective language. Representative remarks of charm included "I like how kind the character is" and "The way of expressing joy is cute" both of which directly mirrored the personal preference of the participant. Moreover, as indicated by the results presented in sections 5.4 and 6.1, utterances with negative emotions appeared to impair charm.

These results indicate that charm is a highly subjective construct, strongly dependent on the preferences and affect of each evaluator. Collectively, character-likeness and charm should be applied based on the intended uses. When the goal is to verify fidelity to the settings of a character, character-likeness is the more reliable metric, whereas dialogue agents intended for sustained user engagement should explicitly optimize for charm. In practical deployments, this insight highlights (i) the need for automatic charm estimators in addition to character-likeness evaluators and (ii) the utility of multi-stage pipelines or training curricula that first ensure character-likeness and then refine responses to maximize charm.

6.3 Intended vs. Perceived Character Profiles

We assessed how closely the official character profile and the users' character image inferred from the free-form descriptions. First, we extracted the sentence that described a character trait. Next, we labeled the relationship between each extracted description and the official profile as entailment, neutral, or contradiction. We used GPT-4o (OpenAI, 2024) for both extraction and the labeling steps.

Regarding character-likeness, the procedure yielded 404 entailments, 421 neutral cases, and 96 contradictions, whereas for charm, it produced 346 entailments, 445 neutral cases, and 98 contradictions. Neutral remarks included, "It really shows that she's soft on assertive advances and easily swayed", which imagines situational behavior not specified in the profile, whereas contradictory remarks such as, "It's also very Sakura-like that she genuinely gets down about it" highlighted a hidden duality that clashed with her cheerful persona.

In both dimensions, more than 50% of the descriptions were either not covered by the profile or absolutely contradicted it, suggesting that users and writers often had their interpretations beyond the official profile. These findings underscore the need for methods that (i) dynamically construct character profiles from the context and (ii) capture the personas that users and writers actually envision.

	IDOLY PRIDE	Dream 100
1	**IP-Dyn.**	**IP-Dyn.**
2	T-Sub.	**IO-Wor.**
3	**IP-Tem.**	**IP-Tem.**
4	**IO-Wor.**	**T-Hab.**
5	**T-Hab.**	GR-Per.

Table 3: Top-five most frequent elements in each game. The overlap percentage is 80%.

6.4 Interview with Scenario Writer

To evaluate the practical relevance of the findings reported in this paper, we conducted an interview with a professional scenario writer, and the obtained the following feedback. (i) The writer recognized the elements that our study observed to consistently improve character-likeness while creating narratives. This suggested that users might perceive the character-likeness of these elements as the writer intended. (ii) The writer indicated that they had previously been vaguely aware that negative utterances that depended on the context and user could diminish the charm of a character. Explicitly addressing this issue improved their understanding. (iii) The writer found it intriguing that users constructed their own images of the characters beyond the official profiles.

6.5 Generalizability from Cross-Game

We evaluated the generalizability of our results by examining the way the distribution of elements differed across games. In addition to *IDOLY PRIDE*, we annotated 436 utterances from five characters in the *Dream 100*[4] scenario. *IDOLY PRIDE* is an idol game featuring female characters, whereas *Dream 100* is a fantasy title centered on male characters. The two games differ markedly in world settings, character genders, and backgrounds.

Table 3 lists the five most frequent elements in each game. Four of the five elements overlapped (80%). This percentage indicated substantial commonality, even when the game genre and character gender differed. Therefore, the insights obtained in this study are likely to be useful not only for a single title but also for games across different genres and characters.

[4] https://www.yume-100.com/

7 Conclusion

We constructed an IDOL-27, an identity-aware taxonomy of 27 elements that considered the internal states of a character. In addition, we experimentally analyzed the properties of each element by collecting ratings and rationales for both character-likeness and charm through an experiment.

The analysis results revealed two classes of elements: (i) elements that consistently improved character-likeness; and (ii) elements whose influence depended on the context. This property remained stable across characters and examinees, whereas charm showed greater individual variation, particularly for emotive cues.

These findings also have practical implications at various stages of system development. For example, in prompt design, they refer to (i) providing descriptions of every element; (ii) assigning greater weight to universally important elements; (iii) flagging context-sensitive elements.

Limitations

Despite the encouraging results, this study has several limitations.

Data Imbalance. Our analysis relied exclusively on in-game utterances from the smartphone title, *IDOLY PRIDE*. Consequently, several of the 27 elements defined in the IDOL-27 scheme occurred only rarely or did not occur at all, yielding a markedly skewed distribution across the elements.

Linguistic Bias. Because the source game was written in Japanese, the present dataset was linguistically and culturally narrow. However, the study protocol itself is language-agnostic, and extending it to other languages and cultures is straightforward. Future work will therefore target English and other languages to enable broad cross-cultural, multilingual comparisons.

Number of Participants. The user evaluation experiment was conducted with a limited cohort of 13 participants.

References

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

Suyoung Bae, Gunhee Cho, Yun-Gyung Cheong, and Boyang Li. 2025. CharMoral: A character morality dataset for morally dynamic character analysis in long-form narratives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8809–8818, Abu Dhabi, UAE. Association for Computational Linguistics.

Roy F Baumeister. 1987. How the self became a problem: A psychological review of historical research. *Journal of personality and social psychology*, 52(1):163.

Marilynn B Brewer and Wendi Gardner. 1996. Who is this" we"? levels of collective identity and self representations. *Journal of personality and social psychology*, 71(1):83.

RB Cattell. 1946. Personality structure and measurement. ii. the determination and utility of trait modality. *British Journal of Psychology*, 36(3):159.

Popa Maria Cristina. 2016. Cultural identity components–romanian parents and german school. *European Proceedings of Social and Behavioural Sciences*.

Erik H. (Erik Homburger) Erikson. 1968. *Identity : youth, and crisis*. W.W. Norton.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States. Association for Computational Linguistics.

Md Naimul Hoque, Bhavya Ghai, Kari Kraus, and Niklas Elmqvist. 2023. Portrayal: Leveraging nlp and visualization for analyzing fictional characters. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, page 7494, New York, NY, USA. Association for Computing Machinery.

Eiichi Kajita. 2020. *Essays on Self-Consciousness 1: Self-Consciousness and Psychology (in Japanese)*, volume 1. TOKYO SHOSEKI.

Avi Kaplan and Joanna K Garner. 2017. A complex dynamic systems perspective on identity and its development: The dynamic systems model of role identity. *Developmental psychology*, 53(11):2036.

Harold H Kelley, Ellen Berscheid, Andrew Christensen, John H Harvey, Ted L Huston, George Levinger, Evie McClintock, Letitia Anne Peplau, and Donald R Peterson. 1983. Analyzing close relationships. *Close relationships*, pages 20–67.

Lennart Kiel, Majse Lind, Adam T. Nissen, Wiebke Bleidorn, and Christopher J. Hopwood. 2024. Incremental relations between self-understanding and social functioning beyond personality traits in young adults. *Journal of Research in Personality*, 113:104546.

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

Dan P. McAdams, Kali Trzesniewski, Jennifer Lilgendahl, Veronica Benet-Martinez, and Richard W. Robins. 2021. Self and identity in personality psychology. *Personality Science*, 2(1):e6035.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.

Zenjiro Mizuta. 1995. A study of literature on the stratum theory of parsinality (in japanese). *Bulletin of Faculty of Education, Nagasaki University. Educational science*, 49:93–109.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Letian Peng and Jingbo Shang. 2024. Quantifying and optimizing global faithfulness in persona-driven role-playing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

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. Association for Computational Linguistics.

Jan E Stets and Peter J Burke. 2000. Identity theory and social identity theory. *Social psychology quarterly*, pages 224–237.

Yihong Tang, Bo Wang, Xu Wang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. RoleBreak: Character hallucination as a jailbreak attack in role-playing systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7386–7402, Abu Dhabi, UAE. Association for Computational Linguistics.

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. 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. 2024a. 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. Association for Computational Linguistics.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024b. Characteristic AI agents via large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3016–3027. ELRA and ICCL.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. 2025. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11086–11106, Abu Dhabi, UAE. Association for Computational Linguistics.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3271–3290. Association for Computational Linguistics.