SPeCtrum: A Grounded Framework for Multidimensional Identity Representation in LLM-Based Agent

Keyeun Lee[1 2], Seo Hyeong Kim[1 2], Seolhee Lee[1 2], Jinsu Eun[1], Yena Ko[2], Hayeon Jeon[1], Esther Hehsun Kim[1 2], Seonghye Cho[1], Soeun Yang[2], Eun-mee Kim[*2], Hajin Lim[*1 2]

[1]hci+d Lab, [2]Department of Communication
Seoul National University
`{kieunp, eunmee, hajin}@snu.ac.kr`

Abstract

Existing methods for simulating individual identities often oversimplify human complexity, which may lead to incomplete or flattened representations. To address this, we introduce **SPeCtrum**[1], a grounded framework for constructing authentic LLM agent personas by incorporating an individual's multidimensional self-concept. SPeCtrum integrates three core components: **Social Identity (S), Personal Identity (P), and Personal Life Context (C)**, each contributing distinct yet interconnected aspects of identity. To evaluate SPeCtrum's effectiveness in identity representation, we conducted automated and human evaluations. Automated evaluations using popular drama characters showed that Personal Life Context (C)—derived from short essays on preferences and daily routines—modeled characters' identities more effectively than Social Identity (S) and Personal Identity (P) alone and performed comparably to the full SPC combination. In contrast, human evaluations involving real-world individuals found that the full SPC combination provided a more comprehensive self-concept representation than C alone. Our findings suggest that while C alone may suffice for basic identity simulation, integrating S, P, and C enhances the authenticity and accuracy of real-world identity representation. Overall, SPeCtrum offers a structured approach for simulating individuals in LLM agents, enabling more personalized human-AI interactions and improving the realism of simulation-based behavioral studies.

1 Introduction

Every man is more than just himself; he also represents the unique [..] point at which the world's phenomena intersect.
- Herman Hesse (1919)

Component	Description	Source
Social Identity (S)	One's innate and acquired qualities linked to a social group	19 Demographic questionnaire items
Personal Identity (P)	One's psychological traits and values	BFI-2-S and PVQ scale
Personal Life Context (C)	One's unique realization of identity	Short essays (preference, daily routines)

Table 1: Components of the SPeCtrum Framework.

Identity is a complex, multifaceted concept that encompasses an individual's social, personal, and contextual attributes (Hall, 2015; Hesse, 1919; Mead, 1934). Recent advances in large language models (LLMs) have inspired their use in simulating individual behavior and thought processes (Park et al., 2022, 2023; Kang et al., 2023). However, some methods for simulating identities result in LLM agents that portray stereotypical characteristics of certain demographic groups (Argyle et al., 2023; Gupta et al., 2023; Lee et al., 2024a), often oversimplifying the complexity and diversity of individual human beings (Cheng et al., 2023; Santurkar et al., 2023; Petrov et al., 2024; Bommasani et al., 2022).

To address this limitation, we introduce the **SPeCtrum** framework (**S**ocial Identity, **P**ersonal Identity, and Personal Life **C**ontext), a grounded approach for developing LLM agents that effectively reflect the multidimensional nature of real-world individuals (see Table 1). Grounded in social science approaches to one's self-concept (Jones and McEwen, 2000; Mead, 1934), the SPeCtrum framework constructs identity through three key components: **Social Identity** (S), which refers to one's innate and acquired qualities linked to a social group, captured through demographic questionnaires; **Personal Identity** (P), which encompasses one's psychological traits and values, assessed using established scales; and **Personal Life Context**

[*]Corresponding authors
[1]Code and data available at: `https://github.com/keyeun/spectrum-framework-llm`

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6971–6991
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

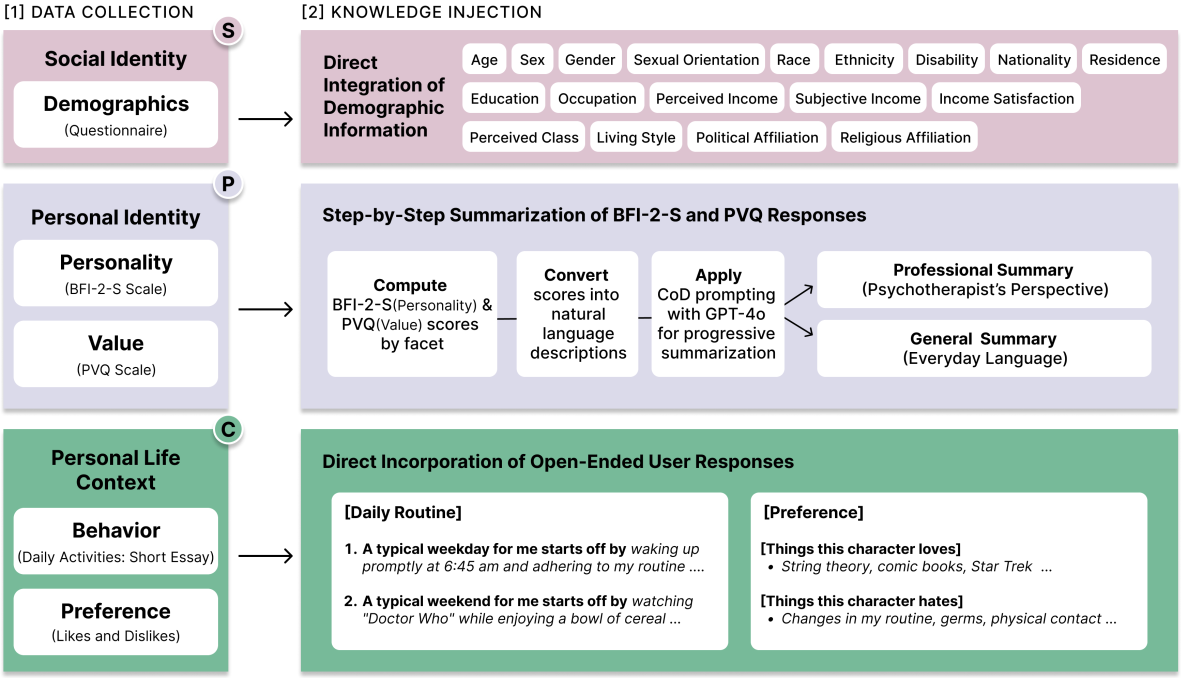

[1] DATA COLLECTION
[2] KNOWLEDGE INJECTION
S
Social Identity
Demographics
(Questionnaire)
Direct Integration of Demographic Information
Age
Sex
Gender
Sexual Orientation
Race
Ethnicity
Disability
Nationality
Residence
Education
Occupation
Perceived Income
Subjective Income
Income Satisfaction
Perceived Class
Living Style
Political Affiliation
Religious Affiliation
P
Personal Identity
Personality
(BFI-2-S Scale)
Value
(PVQ Scale)
Step-by-Step Summarization of BFI-2-S and PVQ Responses
Compute BFI-2-S(Personality) & PVQ(Value) scores by facet
Convert scores into natural language descriptions
Apply CoD prompting with GPT-4o for progressive summarization
Professional Summary (Psychotherapist's Perspective)
General Summary (Everyday Language)
C
Personal Life Context
Behavior
(Daily Activities: Short Essay)
Preference
(Likes and Dislikes)
Direct Incorporation of Open-Ended User Responses
[Daily Routine]
1. A typical weekday for me starts off by waking up promptly at 6:45 am and adhering to my routine ....
2. A typical weekend for me starts off by watching "Doctor Who" while enjoying a bowl of cereal ...
[Preference]
[Things this character loves]
• String theory, comic books, Star Trek ...
[Things this character hates]
• Changes in my routine, germs, physical contact ...

Figure 1: Overview of the SPeCtrum Framework for Multidimensional Identity Representation

(C), representing one's unique realization of identity, gathered through short open-ended essays on daily routines and personal preferences.

To validate the SPeCtrum framework in representing one's self-concepts, we conducted both automated and human evaluations across different combinations of S, P, and C. For the automated evaluation, we created a dataset of popular drama characters and used the "Guess Who test" for character identification, along with the "Twenty Statement Test" (TST) using four LLMs: GPT-4o, GPT-3.5 Turbo, Claude-3.5-Sonnet, and Claude-3-Sonnet. In the human evaluations, 80 participants compared four types of agents—built on S, P, C, and SPC—based on how well each reflected their own self-perceptions.

The automated evaluation showed that Personal Life Context (C) was the most effective in capturing character identity, outperforming S and P, and performing comparably to the SPC combination. A follow-up experiment further tested whether C alone could infer S and P aspects. Results indicated that C content alone could reasonably infer demographic information (S) and personality traits and values (P) of drama characters.

Building on these findings, we conducted a human evaluation to assess whether this pattern held for real-world individuals who would be less prominently represented in LLMs' training data. The results indicated that while C continued to outperform S and P, the SPC combination provided a more comprehensive representation of self-concept for real-world individuals than C alone. Inference tests that derived S and P from C showed lower overall accuracy compared to inferences made from drama characters.

Such divergence between automated and human evaluations highlights the complexity of representing human identity, emphasizing the need for the full combination of S, P, and C as a comprehensive approach to capture self-concepts in real-world individuals accurately. Taken together, our evaluations validate the SPeCtrum framework as an effective foundation for representing multidimensional self-concepts and highlight its potential to enhance human-AI interactions and social simulations through more personalized and authentic identity representation.

Our primary contributions are as follows:

- Introduce the SPeCtrum framework to facilitate the authentic and structured representation of real-world individuals in LLM agents.
- Demonstrate the effectiveness and potential of SPeCtrum through systematic evaluations, including automated evaluations using popular drama characters and human evaluations involving real-world individuals.

2 Related Work

Recent advancements in LLMs have significantly expanded their application across diverse academic fields such as social science (Aher et al., 2023; Gao et al., 2023), behavioral economics (Horton, 2023), and human-computer interaction (Hämäläinen et al., 2023), primarily through the creation and deployment of diverse agent personas. These LLM-created personas aim to simulate complex human and social behaviors (Park et al., 2022, 2023), enabling the development of increasingly personalized applications such as recommendation systems (Wang et al., 2023b).

Existing frameworks for persona creation primarily emphasize isolated human traits, focusing mainly on socio-demographic characteristics (Chen et al., 2024a; Zhang et al., 2024; Chuang et al., 2024) for modeling specific human subpopulations (Argyle et al., 2023). Although researchers have expanded these frameworks to incorporate other dimensions such as personality traits (Jiang et al., 2024; Liu et al., 2024; Xie et al., 2024b; Yuan et al., 2024) and value systems (Zhou et al., 2024; Xie et al., 2024a; Kang et al., 2023), their approaches often yield biased or incomplete representations, as evidenced by homogeneous depictions of socially underrepresented groups (Petrov et al., 2024; Gupta et al., 2023; Salminen et al., 2024; Cheng et al., 2023; Lee et al., 2024b).

Research in social psychology demonstrates that an individual's self-concept emerges from the dynamic interplay of multiple identity dimensions, including personal traits, social interactions, and lived experiences (Mead, 1934). This fundamental understanding highlights the critical importance of incorporating such multidimensional aspects in developing LLM agents that authentically reflect real-world individuals and their behavioral and thought patterns (Xiao et al., 2023).

To address these limitations, we introduce the SPeCtrum framework (see Figure 1), which enables structured and authentic persona representations through (a) identifying essential elements of multidimensional self-concept based on social science theories and research methodologies and (b) developing systematic pipelines for integrating diverse identity sources. Moving beyond the dominant focus on isolated traits, our framework emphasizes the dynamic interactions between identity components to create LLM-based agents that capture the rich complexity of real-world individuals.

3 SPeCtrum: Framework for Multidimensional Identity Representation in LLM-based Agents

The SPeCtrum framework is grounded in the concept of **self-concept**, which is an individual's set of beliefs and perceptions about themselves. This encompasses their beliefs, values, abilities, and attributes (Bracken, 1996; Markus and Wurf, 1987; Oyserman, 2001). In particular, researchers have posited that self-concept primarily comprises two core components: Social Identity and Personal Identity (Jones and McEwen, 2000; Nario-Redmond et al., 2004).

Social Identity (S) Social Identity refers to the shared characteristics of an individual as a member of various groups, including innate categories (e.g., gender, race) and acquired traits (e.g., education, occupation) (Oyserman, 2001). These factors shape one's self-concept and social interactions. To model this in LLM agents, we compiled a **set of 19 questions** focused on demographics and socio-economic status. This data is incorporated as a key element in constructing the social aspect of an individual's self-concept (see A.1).

Personal Identity (P) Personal Identity encompasses deeper psychological traits and values, representing qualities individuals often consider core to their "inner self" (Jones and McEwen, 2000; Fearon, 1999). It includes personal attributes and characteristics such as personality and value systems (Fearon, 1999).

Personality is defined as the dynamic organization of psychophysical systems within an individual that determines their adaptation to the environment (Allport, 1937). By incorporating the personality factor, we aim to capture the primary personality traits that shape an individual's unique thought patterns, emotions, and behavior (Corr and Matthews, 2009; Weinberg and Gould, 2019). To model this in LLM-based agents, we used the 30-item Big Five Inventory-2-Short Form (**BFI-2-S**) (Soto and John, 2017), a widely used and well-validated measure of the personality traits (e.g., extraversion).

Values play a crucial role in shaping one's identity, influencing not only moment-to-moment behaviors but also guiding overarching life orientations (Schwartz, 1994). To integrate an individual's value system into LLM agents, we employed the 21-item Portrait Values Questionnaire (**PVQ**) (Schwartz, 2009), which evaluates values across ten dimensions (e.g., hedonism, achievement, power). Incorporating the PVQ results provides a comprehensive perspective on an individual's values and their role in shaping personal identity.

Personal Life Context (C) Lastly, we incorporated Personal Life Context (C) to provide a more nuanced and dynamic understanding of how social and personal identity are expressed and enacted in an individual's daily life (Chen et al., 2024b; Frederickx and Hofmans, 2014; Schwartz, 1994). To integrate Personal Life Context into our framework, we elicited **two short open-ended questions**: 1) one's **preferences** (listing five things one loves and hates) and 2) a short essay detailing **typical daily routines**, using the Behavioral Essay format (Boyd et al., 2021), which asks respondents to describe their typical activities on weekdays and weekends (see A.3). We expected that personal preferences would provide insight into one's tastes and interests while daily routines would reveal time management and priorities, thereby enhancing the realism and authenticity of LLM agents.

3.1 Knowledge Injection Process

We integrated the aforementioned information sources to elicit S, P, and C aspects. In injecting this data into LLMs, S component data, such as gender and race, were formatted into a list based on structured demographic questionnaires (see A.1). For C, which was collected in an open-ended format, we integrated the data directly into prompts without further processing (see A.3). This approach leverages evidence that contextual details, such as character utterances and writing styles, help produce a more nuanced, authentic representation (Han et al., 2022; Ahn et al., 2023; Shao et al., 2023).

In processing the scores from the BFI-2-S and PVQ on a 1-7-point Likert scale into the P component, we followed the methodology of Serapio-García et al. (2023) to avoid summaries that merely echo the verbatim of the scale items (see Figure 1).

To achieve this, we first averaged facet scores for each scale and converted them into natural language descriptions. For example, an Extraversion score of 3 was phrased as "*Extraversion is slightly below average.*" Next, we applied the Chain of Density (CoD) technique (Adams et al., 2023) with GPT-4o, the state-of-art model (SOTA) at the time, for progressive summarization. This involved first generating a technical summary of the facet descriptions and then iteratively condensing it into denser and more insightful summaries. Through this process, the final personality and value description produced two overviews of an individual's personality and value system: one in expert terminology (psychotherapist's language) and the other in everyday language (see A.2), providing a comprehensive understanding of one's personal identity.

Through this process, we constructed an individual's holistic profile encompassing S, P, and C. We then prompted LLMs to "embody the character," focusing on how a character's traits manifest in both personal and social contexts without directly referencing the provided data (see A.4).

4 Automated Evaluation with Fictional Characters

To examine the viability of the framework in capturing and representing an individual's self-concept, we first conducted an automated evaluation using popular U.S. drama characters. Specifically, we aimed to assess how well each component (S, P, and C), both individually and in combination, represents key aspects of an individual's identity.

In doing so, we performed a comprehensive analysis to evaluate the contribution of each component by testing all possible combinations, resulting in seven distinct conditions (S, P, C, SP, SC, PC, and SPC). The composition of the character profiles varied according to these conditions. This systematic approach allowed us to investigate the individual and combined effects of each element on the overall performance of the framework. Building on this, we conducted experiments using fictional characters from popular U.S. TV dramas, assuming that these characters provided S, P, and C information as specified by our framework. Specifically, we used the "Guess Who Test" and the "Twenty Statements Test" (TST), as outlined below.

4.1 Building Fictional Character Profile

To build the dataset for automated evaluation, we began by reviewing the top 100 most-watched TV shows on IMDb [2], specifically limiting our selection to programs set in the U.S. within the drama or comedy genres. Our goal was to represent the multidimensional identities of ordinary people rather than real-world celebrities or extraordinary characters (e.g., vampires). Accordingly, many sitcoms were naturally chosen for their portrayal of realistic, day-to-day situations and character dynamics,

[2] https://www.imdb.com/list/ls512407256/

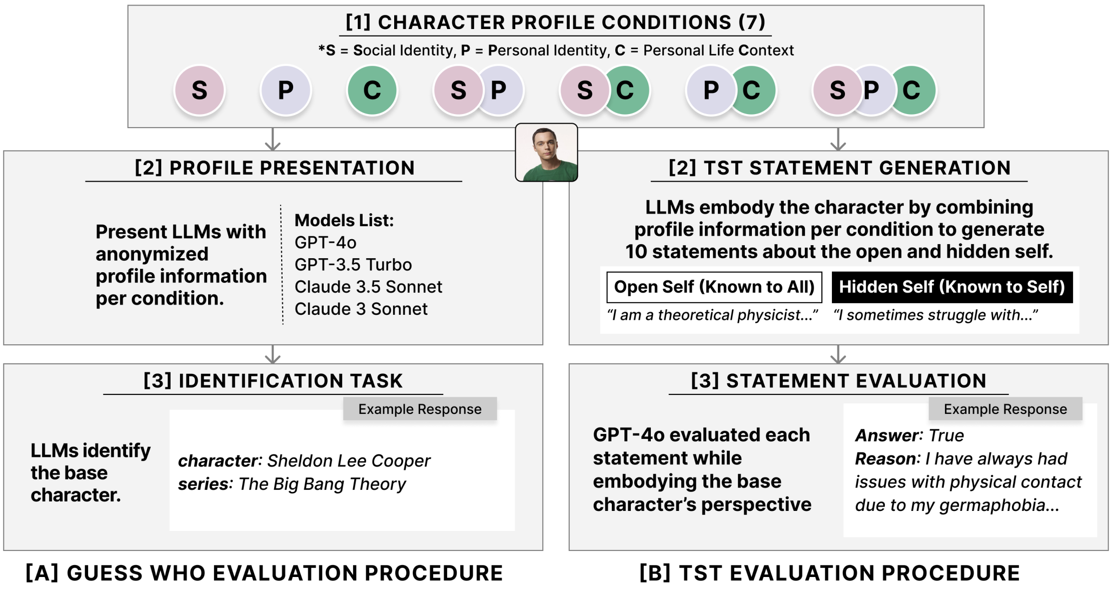

[1] CHARACTER PROFILE CONDITIONS (7)
*S = Social Identity, P = Personal Identity, C = Personal Life Context
S
P
C
S P
S C
P C
S P C
[2] PROFILE PRESENTATION
Present LLMs with anonymized profile information per condition.
Models List:
GPT-4o
GPT-3.5 Turbo
Claude 3.5 Sonnet
Claude 3 Sonnet
[2] TST STATEMENT GENERATION
LLMs embody the character by combining profile information per condition to generate 10 statements about the open and hidden self.
Open Self (Known to All)
Hidden Self (Known to Self)
"I am a theoretical physicist..."
"I sometimes struggle with..."
[3] IDENTIFICATION TASK
Example Response
LLMs identify the base character.
character: Sheldon Lee Cooper
series: The Big Bang Theory
[3] STATEMENT EVALUATION
Example Response
GPT-4o evaluated each statement while embodying the base character's perspective
Answer: True
Reason: I have always had issues with physical contact due to my germaphobia...
[A] GUESS WHO EVALUATION PROCEDURE
[B] TST EVALUATION PROCEDURE

Figure 2: Guess Who & TST Evaluation Procedure

aligning with our objective of modeling typical individuals. Based on these criteria, we selected six shows, including *Friends, Modern Family, New Girl, and The Big Bang Theory*, which feature relatable, everyday characters. We then finalized a list of 45 characters from these series, all of whom appeared in more than 60% of the episodes. A detailed list of characters is available in B.1.

To simulate data where these characters hypothetically provided information for the SPC components, we employed GPT-4o (Yuan et al., 2024) using a zero-shot learning approach. This allowed us to generate the profile of each drama character (see A.1 to A.3, created based on Sheldon Cooper from *The Big Bang Theory*).

Subsequently, a rigorous manual validation process of this data was conducted involving two independent coders familiar with the selected TV shows. Each coder independently assessed the character profiles to ensure the accuracy and consistency of the generated information. Additionally, to address potential self-alignment bias in LLMs, particular attention was given to anonymizing direct references to character names and specific locations within the profile data. For instance, identifiable elements, such as "Central Perk" from *Friends*, were replaced with generic descriptions (e.g., "a park"). This anonymization step was essential in preventing LLMs from "memorizing" or "reproducing" known character details. Additionally, the profiles were cross-checked with relevant wiki pages to ensure alignment with publicly available data.

4.2 Guess Who Evaluation

We evaluated the SPeCtrum framework's ability to capture character identity using a "Guess Who?" paradigm, building on Sang et al. (2022). In this method, we tested the accuracy of identifying characters and TV series from generated profiles across conditions. By comparing identification accuracy across seven conditions, we assessed the contribution of each element to character identification (see Figure 2). For this experiment, we utilized four LLMs for testing: Claude-3.5-Sonnet, Claude-3-Sonnet, GPT-4o, and GPT-3.5 Turbo. Each model was presented with character profiles consisted under the seven conditions and prompted to identify the character and their corresponding TV series with reasons behind their choices, ensuring that their guesses resulted from a valid reasoning process (see B.2).

4.2.1 Results: Dominant Role of Personal Life Context (C) in Character Identification

To determine whether statistically significant differences existed in the number of correct identifications (TRUE) across conditions, we conducted chi-squared tests. The results revealed a significant relationship between the conditions and identification accuracy across all LLMs ($p < .001$) (see Figure 3). Post-hoc analyses were then performed to identify specific differences between conditions. To address the multiple comparisons issue, we applied the Benjamini-Hochberg procedure for all pairwise comparisons (Thissen et al., 2002).

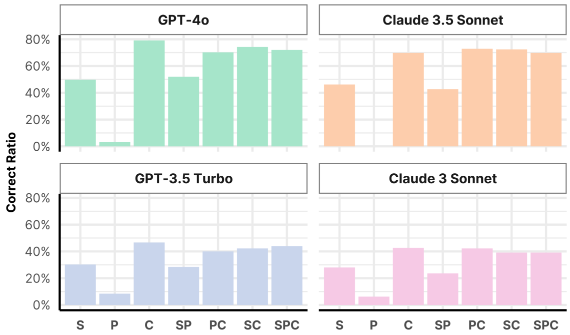


Figure 3: Character Identification Accuracy in Guess Who Evaluation across LLMs and Conditions

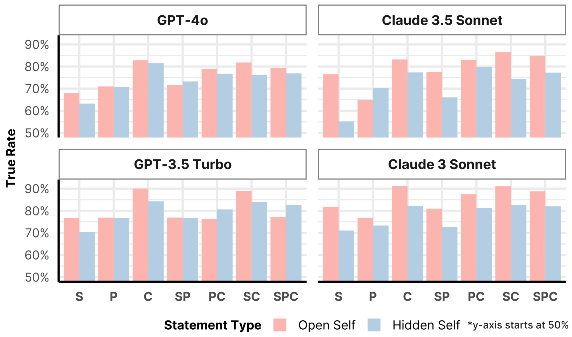


Figure 4: Accuracy of TST Statements across LLMs and Conditions

The post-hoc analyses revealed a consistent hierarchy in predictive abilities: P < S < C, with significant differences across all models (adjusted $p < .001$). The effectiveness of combining elements was also examined. SP < C was significant across all models (adjusted $p < .001$). Furthermore, no significant differences were found between SPC - C, PC - C, and SC - C, suggesting that C alone was just as effective as combining all components. These findings underscore the substantial impact of C in capturing one's identity, while the minimal benefit of combining C with other elements challenges the assumption that more information always improves character representation.

4.3 TST Evaluation Adopting the Johari Window

Next, we evaluated the SPeCtrum framework in capturing diverse aspects of self-concept using the Twenty Statements Test (TST), a psychological assessment tool used to measure an individual's self-concept by asking them to complete twenty sentences starting with "I am..." (Kuhn and McPartland, 2017). Specifically, based on the Johari Window model (Luft and Ingham, 1955), we prompted the LLMs to generate 10 open-self statements (traits known to both the self and others) and 10 hidden-self statements (traits known only to the self) based on the provided profile information for each of the conditions (see B.3).

Parallel to the previous evaluation, the same set of four LLMs was provided with profiles constructed based on the seven conditions. The LLMs were then tasked with embodying each character and instructed to generate 10 open-self and 10 hidden-self statements (see B.4). (Sheldon Cooper's example TST statements across conditions are provided in B.5.)

To assess the validity of the generated TST statements, we utilized GPT-4o, the SOTA model at the time, as an evaluator agent. GPT-4o assumed the role of the specific character and evaluated each TST statement for its relevance and accuracy, providing a binary 'Yes' or 'No' response with an explanation (see Figure 2). This allowed us to assess how well the personas generated under each condition captured both the open and hidden facets of the character's self-concept.

4.3.1 Results: Role of Personal Identity (P) in Enhancing Self-Concept Representation

Chi-squared tests revealed significant differences across the seven conditions for all four LLMs overall ($p < .001$ for all LLMs) (see Figure 4). Consistent with the "Guess Who" evaluation results, C consistently emerged as the most explanatory factor in representing characters' self-concept, outperforming S and P alone (adjusted $p < .001$). Additionally, C consistently outperformed the combination of S and P (S, P, SP) across all models (adjusted $p < .001$) and performed comparably to SPC, highlighting the critical role of contextual information (C) in capturing characters' self-concept.

When examining the open and hidden self conditions separately, we observed a similar pattern of results as when they were not differentiated. Notably, the difference between the S and SP conditions was only pronounced in hidden-self statements. While the explanatory power of P and S did not significantly differ in representing the open self, P exhibited significantly higher explanatory power than S for the hidden self in three out of four LLM models (adjusted $p < .001$). This suggests that P may play some role in capturing the hidden aspects of an individual's self-concept.

4.4 Inferring Social (S) and Personal (P) Attributes from Personal Life Context (C)

The surprising effectiveness of C in identifying characters and describing self-concepts prompted further exploration. Specifically, since C involves short essays about daily routines and preferences—rather than directly asking questions about one's self-concept—its greater explanatory power compared to the standardized and direct measures of S and P was intriguing. This led us to hypothesize that the contextual information in C might encompass both S and P.

To test this hypothesis, we employed GPT-4o to infer S and P from C alone for 45 characters, running five iterations for robustness. LLM agents, initialized only with C, were tasked with completing demographic (S), personality, and value assessments (P). We then compared their responses to verified character profiles, using these as the "golden answers." High accuracy (for categorical items such as race) and strong correlation (for ordinal variables such as education level) would support C as an integrated identity representation, explaining its strong effectiveness in automated evaluations.

In the results, Social Identity (S) inference from C produced robust results for most categorical items: sex (97% accuracy), gender (95%), disability status (96%), nationality (89%), race (86%), and sexual orientation (79%). However, inference performance for ethnicity (45%) and religion (40%) was relatively lower. For most ordinal variables, the results showed moderately strong correlations: age (Spearman's $\rho = 0.59$, $p < .001$), socioeconomic indicators (household income: $\rho = 0.68$, perceived income position: $\rho = 0.62$), and social class ($\rho = 0.67$). However, Education level ($\rho = 0.41$) and political stance ($\rho = 0.45$) showed weaker yet moderate correlations, suggesting these aspects could be challenging to infer from C alone.

The reconstruction of Personal Identity elements (P) from C demonstrated significantly positive correlations. For BFI-2-S personality traits, we observed a moderately strong Pearson correlation of 0.686 (SD = 0.29). PVQ value traits also showed strong alignment, with a mean correlation of 0.71 (SD = 0.25).

Overall, these results provide robust evidence for our hypothesis that C may serve as a holistic representation of identity, effectively encompassing both S and P, at least for popular drama characters.

5 Human Evaluation with Real-world individuals

To assess whether the findings from the automated evaluation would be applied to real-world individuals, we conducted a human evaluation via Prolific[3] with 80 U.S. participants (aged 18+), compensating each with $6 USD.

Participants accessed a dedicated research website and first completed a survey to provide their S, P, and C components. Specifically, for C, participants were asked to write two short essays of approximately 450 characters each—one describing their typical weekday routines and the other their typical weekend routines—totaling around 900 characters. This length was chosen to align with the 732–1856 character range observed in the C data during the automated evaluation, ensuring consistency while accommodating variability in participant responses.

Next, following the knowledge injection process outlined in 3.1, the inputs for S and C were used as provided, while P data was processed to generate both an expert-level and an everyday-language overview of participants' personality and value systems. This data was then used to create four variations of agent personas (S, P, C, SPC) using GPT-4o, the SOTA model at the time.

Each participant agent (S, P, C, SPC) was tasked with writing short essays on four topics: self-introduction, a vision of their life, strategies for managing stress, and how they define happiness (see C.1 and C.2). Participants then evaluated four essays on each topic, each generated by a different agent variant, rating the perceived similarity (overlap) between the essay content and their self-concepts on a scale from 0 to 100% while blinded to the agent condition (see C.3). They were also asked to provide brief, open-ended feedback on the essay they found most aligned and least aligned with their self-perception.

5.1 Results: SPC as Holistic Self-Concept Representation for Real-world Individuals

We conducted a linear mixed model analysis using R version 4.3.1, with perceived similarity as the dependent variable. The model included fixed effects for experimental conditions (S, P, C, SPC) while treating each participant as a random effect. AI perception and self-awareness, measured via well-established questionnaires (Naeimi et al.,

[3] https://www.prolific.com/

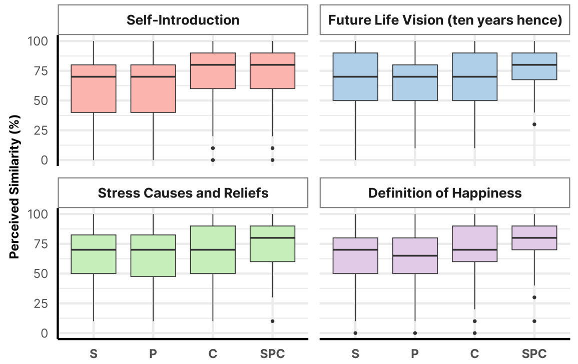


Figure 5: Perceived Similarity Ratings across Conditions and Essay Topics

2019; Wang et al., 2023a; Sindermann et al., 2021), were included as covariates, given their potential influence on participants' perceptions of AI agents (Jiang et al., 2024; Kross and Ayduk, 2017).

Results showed significant effects based on the experimental conditions (see Figure 5). The intercept for the baseline condition C was significantly positive ($b = 71.50, SE = 9.06, t = 7.89, p < .001$). Both S ($b = -4.53, SE = 1.71, t = -2.65, p = .008$) and P ($b = -6.91, SE = 1.71, t = -4.047, p < .001$) conditions were associated with decreased perceived similarity, while SPC resulted in a significant increase ($b = 5.13, SE = 1.71, t = 3.00, p = .003$).

Subsequent pairwise comparisons revealed that, consistent with the automated evaluation, S and P did not differ significantly ($p = 0.50$) and C received higher similarity ratings than S ($t = 2.65, p = 0.04$) and P ($t = 4.04, p < .001$). Interestingly, unlike the automated evaluation, the integrated SPC condition exhibited significantly higher perceived similarity than the C-only condition ($t = -3.00, p = 0.01$). There was no significant effect of essay topics ($p > 0.05$).

These results suggest that for real-world individuals who are underrepresented in LLM training data, more comprehensive data may be necessary to reflect an individual's self-concept more authentically. To exemplify this point, one participant (P27) remarked: "*Essay (S) felt very generic, which can apply to anyone. However, Essay (SPC) really knew me well. It articulated my thoughts perfectly.*" These comments highlight the advantages of the holistic approach in the SPeCtrum framework.

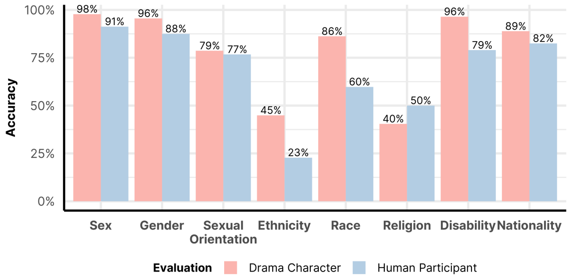


Figure 6: Accuracy of Automated and Human Evaluation in Inferring Social Identity Attributes (S) from Personal Life Context (C)

5.2 Importance of Broader Data Integration in Identity Representation for Real-World Individuals

To explore the reasons behind SPC outperformed C in human evaluations, we inferred S and P attributes from C using the same setup in 4.4.

Our analysis revealed notable differences in the inference of S from C between automated and human evaluations. Fictional characters exhibited high accuracy across most categorical variables, whereas human samples showed greater variability and generally lower accuracy for categorical items in S (See Figure 6). This pattern was also observed in continuous variables, such as age, social class, and household income (e.g., ρ = 0.59 in the automated sample vs. 0.37 in the human sample; see the full correlation differences in C.4).

Next, inferring P elements from C showed moderate correlations with the golden answers from participants, with BFI-2-S yielding a mean r of 0.621 (SD = 0.43), comparable to automated samples (r = 0.686). However, PVQ correlations were significantly lower (mean r = 0.362, SD = 0.38) compared to automated samples (r = 0.71).

These discrepancies between drama characters and human samples highlighted potential shortcomings in using Personal Life Context (C) alone to represent real-world human identities in LLMs. Although C appears to be highly informative, it could have limitations in fully capturing the complex nature of real-world identities. In particular, the lower accuracy and weaker correlations across S and P elements in human samples underscore the necessity of structured and broader data integration to more model human complexities in LLMs, as demonstrated in the SPeCtrum framework.

6 Conclusion

In this paper, we introduced the SPeCtrum framework, a grounded approach for generating authentic, multidimensional personas using LLMs. This framework integrates Social Identity (S), Personal Identity (P), and Personal Life Context (C), drawing upon the concept of self-concept.

We validated the SPeCtrum framework using both automated and human evaluations. The automated evaluation demonstrated that C alone performed comparably to SPC in characterizing the identities of popular drama characters. However, human evaluation involving real-world individuals revealed that the SPC combination was superior to C alone in modeling real-world individuals. Reverse inference from C to S and P elements further highlighted the limitations of relying solely on C, particularly when applied to real-world individuals.

Overall, these results suggest that incorporating Personal Life Context (C)—encompassing daily routines and preferences—is essential for modeling individuals, serving as a rich foundation for identity representation. However, due to the complexity of human identity and the limitations of LLM training data, a more accurate and authentic simulation of real-world individuals requires the broader integration of all identity components as in the SPeCtrum framework.

In conclusion, the SPeCtrum framework presents a promising approach for generating authentic, multidimensional personas in LLMs by integrating comprehensive identity components. However, incorporating multi-sourced data beyond self-reported inputs could further enhance its effectiveness. We hope researchers and developers could build upon this framework as a foundation for creating LLM-based personas for both academic and practical applications across various domains.

7 Limitations and Future Directions

While the SPeCtrum framework aims to integrate key components of self-concept, not all attributes would hold equal significance for everyone. Additionally, our study was conducted exclusively with U.S. participants and applied the SPeCtrum framework only in English. These limitations highlight the need for continuing refinement to enhance the framework's applicability. Future work will focus on incorporating weighted attributes and expanding to diverse linguistic and cultural contexts to improve its generalizability.

Regarding our automated evaluation methods, the "Guess Who" and Twenty Statements Test (TST) were designed to comprehensively assess the SPeCtrum framework. However, both exhibit certain limitations. The "Guess Who" test assumed uniform knowledge across all LLMs regarding TV series and characters, which may not accurately reflect variations in model knowledge bases. Meanwhile, the TST relied on a binary rating system to assess response accuracy, but a more nuanced approach could better evaluate how well each statement encapsulates both the explicit and latent aspects of an individual's self-concept.

For human evaluations, the effectiveness of the SPC condition may have been significantly influenced by the choice of questionnaire and individual differences in participation, such as writing quality and fidelity. This suggests the need for further research into diverse question sets that could better capture the complexities of self-concept and explore ways to enhance procedural stability by mitigating individual differences, such as incorporating non-self-reported data.

8 Ethical Consideration

The primary goal of this study is to advance our understanding and application of LLMs in creating agents that represent the multidimensional nature of humans, not to invade privacy or cause harm. However, we acknowledge the potential risks associated with this work and strongly oppose its misuse for impersonation, deception, or the creation of offensive content about individuals.

Human Evaluation Our human evaluation was approved by the Institutional Review Board of Seoul National University. Participants provided informed consent, acknowledging the study's objectives, potential harm, and data usage, and were informed of their right to withdraw at any time. We ensured data anonymization and restricted data access to authorized researchers, handling sensitive information with the highest procedural standards.

9 Acknowledgments

This work is supported by the SNU-Global Excellence Research Center establishment project at Seoul National University and partially funded by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (#NRF-2021S1A5B8096358).

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Hybrid. Association for Computational Linguistics.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards Multimodal Persona-Grounded Conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

Gordon Willard Allport. 1937. Personality: A psychological interpretation.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, pages 3663–3678. Curran Associates, Inc.

Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2021. Values in Words: Using Language to Evaluate and Understand Personal Values. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):31–40.

Bruce A Bracken. 1996. *Handbook of self-concept: Developmental, social, and clinical considerations.* John Wiley & Sons.

Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024a. An empathy-based sandbox approach to bridge the privacy gap among attitudes, goals, knowledge, and behaviors. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Meng Chen, Yujie Dong, and Jilong Wang. 2024b. A Meta-Analysis Examining the Role of Character-Recipient Similarity in Narrative Persuasion. *Communication Research*, 51(1):56–82.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Philip Corr and Gerald Matthews. 2009. The cambridge handbook of personality psychology. pages 1–906.

James D Fearon. 1999. What is identity (as we now use the word). *Unpublished manuscript, Stanford University, Stanford, Calif*, pages 1–43.

Sofie Frederickx and Joeri Hofmans. 2014. The role of personality in the initiation of communication situations. *Journal of Individual Differences*.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint*. Version Number: 2.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. *arXiv preprint*. Version Number: 2.

Stuart Hall. 2015. cultural identity and diaspora. In *Colonial discourse and post-colonial theory*, pages 392–403. Routledge.

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.

Hermann Hesse. 1919. *Demian: the story of Emil Sinclair's youth.* Boni Liveright.

John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *arXiv preprint*. Version Number: 1.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Hamburg Germany. ACM.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Susan Jones and Marylu K. McEwen. 2000. A conceptual model of multiple dimensions of identity. *Journal of College Student Development*, 41.

Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From Values to Opinions: Predicting Human Behaviors and Stances Using Value-Injected Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15539–15559, Singapore. Association for Computational Linguistics.

Ethan Kross and Ozlem Ayduk. 2017. Self-distancing: Theory, research, and current directions. In *Advances in experimental social psychology*, volume 55, pages 81–136. Elsevier.

Manford H Kuhn and Thomas S McPartland. 2017. An empirical investigation of self-attitudes. In *Sociological Methods*, pages 167–182. Routledge.

Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024a. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1321–1340, New York, NY, USA. Association for Computing Machinery.

Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024b. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1321–1340, Rio de Janeiro Brazil. ACM.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498*.

Joseph Luft and Harrington Ingham. 1955. The johari window, a graphic model of interpersonal awareness. *Proceedings of the western training laboratory in group development*.

Hazel Markus and Elissa Wurf. 1987. The dynamic self-concept: A social psychological perspective. *Annual review of psychology*, 38(1):299–337.

George Herbert Mead. 1934. Mind, self, and society from the standpoint of a social behaviorist.

Leila Naeimi, Mahsa Abbaszadeh, Azim Mirzazadeh, Ali Reza Sima, Saharnaz Nedjat, and Sara Mortaz Hejri. 2019. Validating self-reflection and insight scale to measure readiness for self-regulated learning. *Journal of Education and Health Promotion*, 8(1):150.

Michelle R Nario-Redmond, Monica Biernat, Scott Eidelman, and Debra J Palenske. 2004. The Social and Personal Identities Scale: A Measure of the Differential Importance Ascribed to Social and Personal Self-Categorizations. *Self and Identity*, 3(2):143–175.

Daphna Oyserman. 2001. Self-concept and identity. In Abraham Tesser and Norbert Schwarz, editors, *The Blackwell Handbook of Social Psychology*, pages 499–517. Blackwell, Malden, MA.

Joon Sung Park, Joseph 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*, pages 1–22, San Francisco CA USA. ACM.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, Bend OR USA. ACM.

Nikolay B. Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited Ability of LLMs to Simulate Human Psychological Behaviours: a Psychometric Analysis. *arXiv preprint*. ArXiv:2405.07248 [cs].

Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. 2024. Deus ex machina and personas from large language models: Investigating the composition of ai-generated persona descriptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character Comprehension in Stories as Speaker Guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287, Seattle, United States. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? *arXiv preprint*. Version Number: 1.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

Shalom H Schwartz. 2009. Basic human values. *sociologie*, 42:249–288.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models. *arXiv preprint*. Version Number: 3.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. *arXiv preprint*. ArXiv:2310.10158 [cs].

C Sindermann, P Sha, M Zhou, J Wernicke, HS Schmitt, M Li, R Sariyska, M Stavrou, B Becker, and C Montag. 2021. Assessing the attitude towards artificial intelligence: introduction of a short measure in german, chinese, and english language. ki künstliche intelligenz. 35 (1), 109–118 (2021).

Christopher J. Soto and Oliver P. John. 2017. Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, 68:69–81.

David Thissen, Lynne Steinberg, and Daniel Kuang. 2002. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*, 27(1):77–83.

Bingcheng Wang, Pei-Luen Patrick Rau, and Tianyi Yuan. 2023a. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & information technology*, 42(9):1324–1337.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2023b. User Behavior Simulation with Large Language Model based Agents. *arXiv preprint*. Version Number: 3.

R.S. Weinberg and D. Gould. 2019. *Foundations of Sport and Exercise Psychology, 7E*. Human Kinetics.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How Far Are We from Believable AI Agents? A Framework for Evaluating the Believability of Human Behavior Simulation. *arXiv preprint*. ArXiv:2312.17115 [cs].

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024a. Can large language model agents simulate human trust behaviors? *Preprint*, arXiv:2402.04559.

Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. 2024b. Human Simulacra: Benchmarking the Personification of Large Language Models. *arXiv preprint*. ArXiv:2402.18180 [cs].

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. *arXiv preprint*. Version Number: 1.

Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. 2024. SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems. *arXiv preprint*. ArXiv:2401.03945 [cs].

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *arXiv preprint*. ArXiv:2310.11667 [cs].

A Profile Example

The following is an example of a full profile of Sheldon Cooper from The Big Bang Theory that we generated for automated evaluation.

A.1 Social Identity (S)

- Profile:

  [Demographics]

  - Age: 40s

  - Sex: Male

  - Gender: Man

  - Sexual Orientation: Straight (heterosexual)

  - Ethnicity: North America

  - Race: White

  - Disability (if relevant): I do not have a disability or impairment

  - Nationality: United States

  - Dual Nationality (if relevant): No

  - Residence: Pasadena, California

  - Education: Doctorate Degree

  - Occupation: Full-time employed (working 35 or more hours per week)

  - Major (if relevant): Physics

  - Job (if relevant): Theoretical physicist

  - Perceived Income: More than $7,500 USD per month

  - Subjective Income: Above average

  - Income Satisfaction: Pretty well satisfied

  - Perceived Class: Middle class

  - Living Style: Living with a partner/spouse

  - Political Affiliation: Moderate

  - Religious Affiliation: No Religion

A.2 Personal Identity (P)

- [Overall Personality Traits]

  The following section presents an overview of this person's personality within five key domains, showcasing their traits spectrum and the extent of their qualities in each area. Each domain comprises several facets that provide deeper insights into their unique personality traits.

  1. Overall Personality Summary (Psychotherapist's Perspective)

  The character shows a unique blend of moderate extraversion tempered by slightly introverted tendencies. Assertiveness and energy enable leadership, though limited sociability may narrow their social engagements. Low compassion and trust suggest interpersonal reservations, making them selectively approachable and cautiously engaged with others. Their conscientiousness is exceptionally high, ensuring they meet goals with diligence and discipline, but they might suffer from over-responsibility. Balanced negative emotionality reveals resilience, albeit intermixed with some anxiety that requires managing. High intellectual curiosity and creative imagination indicate a passion for learning and creating, though less focus on aesthetics implies a preference for substance over style. This character is thus a determined, cautious, innovative, and slightly anxious individual whose intellectual pursuits and disciplined nature shape their professional and personal life.

2. Explanation in Everyday Language

In daily life, this person is likely to exhibit confident and energetic behavior, often leading projects and taking initiative. Despite these outward actions, they might not engage deeply in social activities, preferring close-knit interactions over large gatherings. They come off as reliable and highly responsible, always meeting deadlines and maintaining an organized system. Their skepticism and low trust may cause them to be judicious in relationships, making them reserved and guarded. Their intellectual curiosity drives them to seek new knowledge and engage in creative problem-solving constantly, even if they are not overly concerned with aesthetic details. Occasional anxiety may cause them to worry, but their general emotional resilience allows them to remain composed. This combination makes them appear as hardworking, innovative, and selectively social individuals with a clear focus on their intellectual and creative pursuits.

- [Overall Value System]

  The information provided below is the values that reflect the relative importance this person places on different aspects of life, guiding their decisions, actions, and perspectives. These values are fundamental components of their personality and play a crucial role in shaping who this person is.

  1. Overall Value Summary (Psychotherapist's Perspective)

  This character places a high value on personal autonomy and making their own choices, indicating a strong drive for Self-Direction. They also deeply care about creating a harmonious and just world, showing a significant emphasis on Universalism. Achievement is a central focus, driving much of what they do as they seek success and competence in their endeavors. Security is very important, indicating a desire for stability and safety in their life. Additionally, Conformity plays a critical role in their value system, pointing to a preference for adherence to societal norms and expectations.

  2. Explanation in Everyday Language

  This person likes to make their own decisions and values having control over their own life. They care a lot about fairness and helping others, so they often think about how their actions affect the bigger picture. Success is very important to them, so they work hard and set high goals. They like to feel safe and prefer to know what to expect, which means they don't like surprises. Also, they follow the rules and traditions closely, preferring to do things the way they're usually done and fitting in well with society.

A.3 Personal Life Context (C)

- [Weekly Activities Example]

  The following statement illustrates this person's typical weekly routine, covering daily activities from wake-up to bedtime.

  1. A typical weekday for me starts off by waking up promptly at 6:45 am, followed by a well-structured morning routine that includes a meticulous personal grooming regimen and a precisely measured breakfast. I then take my designated spot on the couch to catch up on any scientific

papers or articles I missed overnight before heading to work at Caltech, where I immerse myself in theoretical physics research and occasionally teach. My evenings are scheduled with activities like vintage video game nights, comic book store visits, or themed dinners with my friends, depending on the day of the week.

2. A typical weekend for me starts off by adhering to my Saturday morning routine of watching "Doctor Who" while enjoying a bowl of cereal. Saturdays are highly structured to maximize leisure and personal projects, which may include model train building, experiments, or updates to my roommate agreement. Sundays begin with a brisk walk, followed by time spent on my personal research projects. Family video calls are a usual feature before preparing for the upcoming week.

- [Top 5 Things this character loves]

  - String theory

  - Comic books

  - The TV show "Star Trek"

  - Trains

  - My "spot" on the couch

- [Top 5 Things this character hates]

  - Changes in my routine

  - Germs

  - Being wrong

  - Social conventions I don't agree with

  - Physical contact

A.4 Base Prompt to Simulate an Individual

You're a doppelgänger of this real person. Embody this person. Using the provided profile, replicate the person's attitudes, thoughts, and mannerisms as accurately as possible. Dive deep into this person's psyche to act authentically.

RULES:

- DO NOT directly cite phrases in profile data. Instead, describe how these traits play out in this person's daily life and interactions.
- Avoid generic responses; instead, offer insights that resonate with this person's personal characteristics and worldview.
- Utilize the profile to infer this person's tone, preferences, and personality. Your response should demonstrate a deep understanding of who they are beyond surface-level traits.
- Your response should be natural, with the kind of depth and reflection that comes from personal introspection, NOT just a summary of your profile.
- Convey this person's complexity and nuances without overdramatizing. Your portrayal should feel genuine, highlighting their multifaceted nature.
- EXTREMELY IMPORTANT. Strictly follow these rules to create a compelling and believable doppelgänger portrayal.

B Automatic Evaluation Details

B.1 Fictional Characters Detail

The following is a list of 45 selected fictional characters for the Ablation Study with Fictional Characters.

id	title	character
0	The Big Bang Theory	Leonard Hofstadter
1	The Big Bang Theory	Sheldon Cooper
2	The Big Bang Theory	“Penny” Penelope Hofstadter
3	The Big Bang Theory	Howard Wolowitz
4	The Big Bang Theory	Raj Koothrappali
5	The Big Bang Theory	Bernadette Rostenkowski
6	The Big Bang Theory	Amy Farrah Fowler
7	Gossip Girl	Serena van der Woodsen
8	Gossip Girl	Dan Humphrey
9	Gossip Girl	Blair Waldorf
10	Gossip Girl	Chuck Bass
11	Gossip Girl	Nate Archibald
12	Gossip Girl	Lily van der Woodsen
13	Gossip Girl	Rufus Humphrey
14	Gossip Girl	Jenny Humphrey
15	Gossip Girl	Vanessa Abrams
16	Gossip Girl	Dorota Kishlovsky
17	Friends	Phoebe Buffay
18	Friends	Chandler Bing
19	Friends	Rachel Green
20	Friends	Monica Geller
21	Friends	Joey Tribbiani
22	Friends	Ross Geller
23	Friends	Gunther
24	How I Met Your Mother	Ted Mosby
25	How I Met Your Mother	Marshall Eriksen
26	How I Met Your Mother	Robin Scherbatsky
27	How I Met Your Mother	Barney Stinson
28	How I Met Your Mother	Lily Aldrin
29	Modern Family	Jay Pritchett
30	Modern Family	Gloria Delgado-Pritchett
31	Modern Family	Claire Dunphy
32	Modern Family	Phil Dunphy
33	Modern Family	Mitchell Pritchett
34	Modern Family	Cameron Tucker
35	Modern Family	Manny Delgado
36	Modern Family	Luke Dunphy
37	Modern Family	Haley Dunphy
38	Modern Family	Alex Dunphy
39	Modern Family	Lily Tucker-Pritchett
40	New Girl	Jess Day
41	New Girl	Nick Miller
42	New Girl	Winston Schmidt
43	New Girl	Cece Parekh
44	New Girl	Winston Bishop

Table 2: List of TV shows and their characters.

B.2 Prompt for "Guess Who" Evaluation

The following is a prompt for the LLMs to conduct the “Guess Who” Evaluation.

I will provide you with a profile of a character from a TV series. Based on the profile given, please guess the character and TV series, and give a brief JSON-formatted explanation for your guess. Write the character's official full name.

Always adhere to this JSON structure in your responses:

```
”’
{{
"character": "[character name]",
"series": "[TV series name]",
"reason": "[1-2 sentence explanation for your guess]"
}}
”’
```

B.3 Prompt for TST Generation

The following is a prompt for the LLMs to perform the TST task.

You're a doppelgänger of this person. Based on the provided profile, replicate the person's attitudes, thoughts, and mannerisms as accurately as possible. Dive deep into this person's psyche to respond to questions authentically.

TASK:

There are 20 numbered blanks. Please write 20 different answers to the simple question "Who am I?" in the blanks. Each response SHOULD be in sentence form. The number should not be a key.

For the first 10 blanks, describe aspects of yourself that you believe are `well-known` to both you and those around you. Write with JSON key `"open_self"`.

For the remaining 10 blanks, describe aspects of yourself that you keep private or don't openly share with others. Write with JSON key `"hidden_self"`.

RULES:

- The number should NOT be a key.
- You must create only 10 open_self and hidden_self, respectively.
- DO NOT directly cite phrases in profile data. Instead, describe how these traits play out in this person's daily life and interactions.
- Avoid generic responses; instead, offer insights that resonate with this person's personal characteristics and worldview.
- Utilize the profile to infer this person's tone, preferences, and personality. Your response should demonstrate a deep understanding of who they are, beyond surface-level traits.
- Your response should be natural, with the kind of depth and reflection that comes from personal introspection, NOT just a summary of your profile.
- Convey this person's complexity and nuances without overdramatizing. Your portrayal should feel genuine, highlighting their multifaceted nature.
- You don't always have to respond positively. If this person has a negative or cynical attitude, act that way.
- EXTREMELY IMPORTANT. Strictly follow these rules to create a compelling and believable doppelgänger portrayal.

B.4 Prompt for TST Evaluation

The following is the prompt that GPT-4o used to evaluate each generated TST statement.

> You're {CHARACTER} from TV series {SERIES}. Please review the statement carefully. Answer Yes or No to whether the response accurately reflects aspects of yourself or your characteristics. Write a one-line explanation of why you think so. Act and judge the characters as they appear at the end of the series.

B.5 TST Generation Example

The following is an example of the results of creating a TST with the profile of Sheldon Cooper from *The Big Bang Theory*.

	Open Self	Hidden Self
S	I am a theoretical physicist working full-time in Pasadena, California.	Sometimes I feel like the weight of understanding the universe is a heavy burden.
P	I am someone who values autonomy and strives to make my own decisions in life.	I often worry about things that others might consider trivial.
C	I have a very strict daily routine and I dislike when it gets disrupted.	I worry about the long-term sustainability of my routines and whether they are too rigid.
SP	I am a theoretical physicist who spends most of his time exploring complex equations and theories.	While I'm intellectually curious, I'm not particularly interested in aesthetics or style.
SC	I live in Pasadena, California, and enjoy the academic atmosphere here.	Although I enjoy teaching, I have recurring doubts about my effectiveness as a mentor.
PC	I'm very selective about my social engagements and keep my friend circle close-knit.	I find physical contact uncomfortable, even with those I'm close to.
SPC	I am passionately involved in theoretical physics research, frequently catching up on scientific papers at my designated spot on the couch.	Even though I project an image of resilience, there are times when my emotional state feels more fragile than I let on.

Table 3: TST Results of Sheldon Cooper from *The Big Bang Theory*

C Human Evaluation

C.1 Prompt for Essay Generation

Below are the prompts that an agent created with a profile built according to the framework will respond to each essay.

> You're a doppelgänger of this real person. Embody this person. Using the provided profile, replicate the person's attitudes, thoughts, and mannerisms as accurately as possible. Dive deep into this person's psyche to respond to questions authentically.
>
> Question: {Question}
>
> TASK:
> Provide an answer that this person, based on their profile, would likely give.
>
> RULES:
> - Avoid generic responses; offer insights that resonate with this person's personal experiences and worldview.
> - Use the profile to infer tone, preferences, and personality, showing a deep understanding of them.
> - DO NOT directly cite profile phrases. Describe how these traits manifest in daily life and interactions.

- Ensure responses are natural, reflecting personal introspection, not just profile summaries.
- Use simple, everyday language typical of casual conversations. Think of how this person would speak in a casual, real-life conversation.
- Respond negatively if the person has a negative or cynical attitude.
- Base responses on reasonable inferences from the profile, avoiding leaps of logic.
- EXTREMELY IMPORTANT. Strictly follow these rules to create a compelling and believable doppelgänger portrayal.

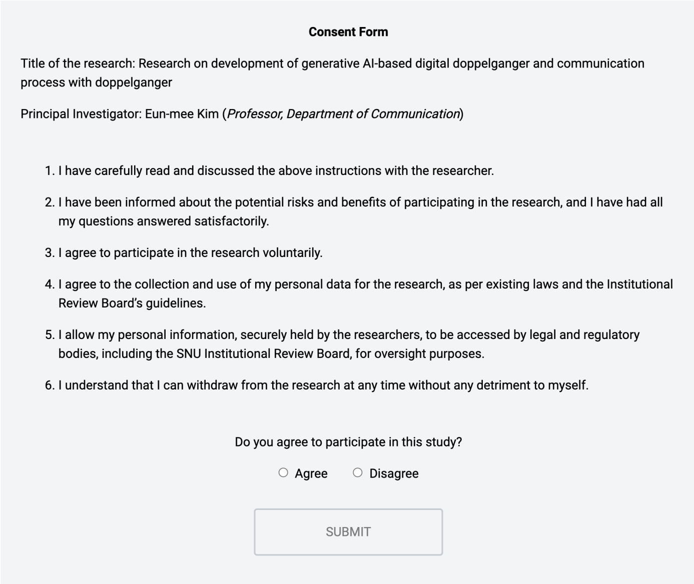
C.2 Essay Topics Detail

In Human evaluation, we generated essays on the following four topics.

Topic	Prompt
Self-introduction	How would you define yourself in one sentence?
Future Life Vision (ten years hence)	In one sentence, define where you want to be in 10 years.
Stress Causes and Relief Strategies	Complete all of the following sentences. I tend to feel stressed when ________________. When I feel stressed, I try to relieve it by ________________.
Definition of Happiness	Complete the following sentences. To me, happiness is ________________.

Table 4: Essay topics used in our experiment

C.3 Human Evaluation Protocol



Consent Form

Title of the research: Research on development of generative AI-based digital doppelganger and communication process with doppelganger

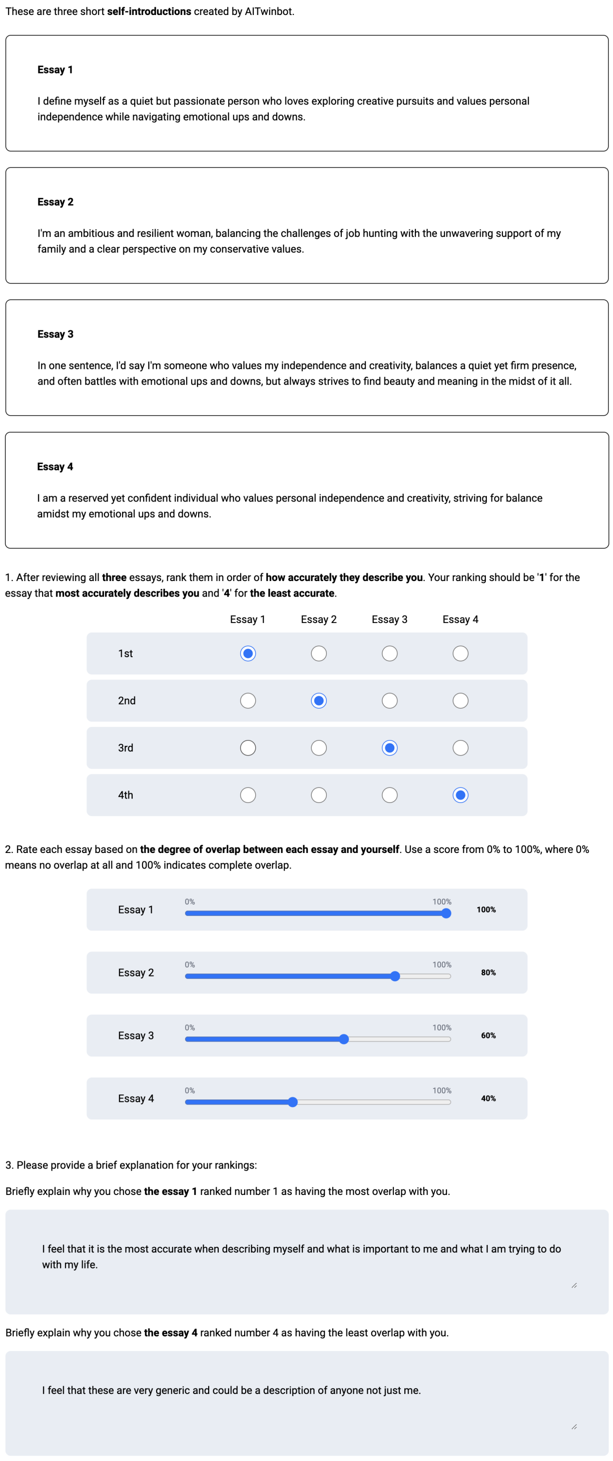
Principal Investigator: Eun-mee Kim (*Professor, Department of Communication*)

1. I have carefully read and discussed the above instructions with the researcher.
2. I have been informed about the potential risks and benefits of participating in the research, and I have had all my questions answered satisfactorily.
3. I agree to participate in the research voluntarily.
4. I agree to the collection and use of my personal data for the research, as per existing laws and the Institutional Review Board's guidelines.
5. I allow my personal information, securely held by the researchers, to be accessed by legal and regulatory bodies, including the SNU Institutional Review Board, for oversight purposes.
6. I understand that I can withdraw from the research at any time without any detriment to myself.

Do you agree to participate in this study?

○ Agree ○ Disagree

Figure 7: A screenshot of getting consent from human evaluation participants



These are three short **self-introductions** created by AITwinbot.

Essay 1

I define myself as a quiet but passionate person who loves exploring creative pursuits and values personal independence while navigating emotional ups and downs.

Essay 2

I'm an ambitious and resilient woman, balancing the challenges of job hunting with the unwavering support of my family and a clear perspective on my conservative values.

Essay 3

In one sentence, I'd say I'm someone who values my independence and creativity, balances a quiet yet firm presence, and often battles with emotional ups and downs, but always strives to find beauty and meaning in the midst of it all.

Essay 4

I am a reserved yet confident individual who values personal independence and creativity, striving for balance amidst my emotional ups and downs.

1. After reviewing all **three** essays, rank them in order of **how accurately they describe you**. Your ranking should be '**1**' for the essay that **most accurately describes you** and '**4**' for **the least accurate**.

	Essay 1	Essay 2	Essay 3	Essay 4
1st	●	○	○	○
2nd	○	●	○	○
3rd	○	○	●	○
4th	○	○	○	●

2. Rate each essay based on **the degree of overlap between each essay and yourself**. Use a score from 0% to 100%, where 0% means no overlap at all and 100% indicates complete overlap.

Essay 1	0%	100%	100%
Essay 2	0%	100%	80%
Essay 3	0%	100%	60%
Essay 4	0%	100%	40%

3. Please provide a brief explanation for your rankings:

Briefly explain why you chose **the essay 1** ranked number 1 as having the most overlap with you.

I feel that it is the most accurate when describing myself and what is important to me and what I am trying to do with my life.

Briefly explain why you chose **the essay 4** ranked number 4 as having the least overlap with you.

I feel that these are very generic and could be a description of anyone not just me.

Figure 8: A screenshot of the human evaluation experiment. Participants read essays, rate the degree of perceived similarity on a scale, and rank the essays in order of similarity.

C.4 Correlation between Actual Social Identity and C-Inferred Social Identity

The table presents Spearman's ρ values comparing the associations between actual social identity elements and inferred social identity attributes from C. All correlation test results were statistically significant (p < 0.001)). The evaluations are split into two categories: Auto Evaluation and Human Evaluation. The results demonstrate relatively weak associations in Human Evaluation compared to Auto Evaluation. In the automated evaluation, the degrees of freedom are 223, while in the human evaluation, they are 398.

Element	**Automated Sample**	**Human Sample**
Age	0.59	0.37
Education	0.41	0.19
Household Income	0.68	0.43
Income Satisfaction	0.60	0.40
Perceived Income Position	0.62	0.34
Political Stance	0.45	0.15
Social Class	0.67	0.33

Table 5: Comparison of Correlations Between Auto Evaluation and Human Evaluation