

Preserving Character in Four Letters: Personality Information Compression in AIWolf Agents Using MBTI-like Labels

Ryosuke Sueyoshi

Nihon Kogakuin College, Tokyo Technical College, Japan

ai168.sueyoshi@gmail.com

Kentaro Hoshi

Global Education Center, Waseda University, Japan

drhoshiken@aoni.waseda.jp

Miyu Ueki

Meiji University, Japan

sq678223072849@gmail.com

Chaeyoung Sung

Graduate School of Artificial Intelligence and Science, Rikkyo University, Japan

seongchae35@gmail.com

Abstract

In AIWolf games, LLM-based agents may role-play characters with different personalities. The personality description may be repeatedly included in the input prompt to an LLM. ChatGPT is a well-known example of an LLM. Previous studies have shown that shorter prompts can reduce response time and computational cost. Faster responses may support more natural interaction with humans. We therefore explored shortening the prompt. In this study, we replace natural-language personality descriptions with four-letter MBTI-like labels. Age and gender are retained. The four-letter labels are not intended for human readers. The labels are not treated as psychological diagnoses. We include the labels in the prompt to the LLM as short cues for generating character-consistent utterances. The MBTI condition reduced the character count of the personality representation by 94.98%. It also reduced the character count of the overall profile by 86.80%. The total input token count was reduced by 24.59%. In a small exploratory human evaluation, we did not detect a statistically significant difference in perceived character likeness between the two conditions. In contrast, the proportion of cases in which the best-matching character was correct decreased from 51% to 33%. The proportion of cases in which the correct character was among the three best-matching characters also decreased from 84% to 74%. These results suggest that four-letter compression can preserve the perceived likeness of each character. However, it can reduce the cues that distinguish one character from another. This study shows the potential of compressed representations of personality information.

1 Introduction

Can “character likeness” be stored as external data and reconstructed when needed? If information that characterizes a person’s personality and behavior can be condensed into a sufficiently compact representation, it may become possible to store, transfer, and reuse that information across different systems, including large language models (LLMs).

Large language models (LLMs), such as ChatGPT, can generate natural-language utterances that reflect a particular character when given profiles containing information such as name, age, gender, and personality (Tseng et al., 2024). However, detailed personality descriptions increase the number of input tokens, and in dialogue settings where the same profile is repeatedly used, the model must process that information again for every utterance. Previous studies have shown that shorter prompts can reduce inference time and computational cost (Jiang et al., 2023).

Prompt compression methods such as LLMLingua (Jiang et al., 2023) and LLMLingua-2 (Pan et al., 2024) selectively preserve important information from the original text. In contrast, rather than rewriting the personality description itself into a shorter text, we use a compact label as a semantic cue intended to evoke personality-related knowledge that an LLM may have acquired during pretraining.

We apply this approach to AIWolf agents that conduct natural-language dialogue in Japanese. In the Werewolf game, agents must infer the roles and intentions of others while engaging in persuasion, cooperation, and deception; merely generating character-like utterances is insufficient

for successful gameplay. AIWolfDial has studied such natural-language Werewolf agents (Kano et al., 2024). Previous work has also investigated LLM-based agents that generate character-consistent utterances from profile information (Tanaka et al., 2024). Because the same character speaks repeatedly throughout a game, reducing personality information can provide repeated savings, and may also contribute to future real-time gameplay between humans and AI agents.

How much of a character’s likeness remains when personality information is compressed to an extreme degree? Does enough individuality remain to distinguish one character from another? In this study, only the natural-language personality description is replaced with a four-letter MBTI-like label such as INFP, while age and gender are retained. We chose MBTI-like labels not for their psychometric validity (McCrae and Costa, 1989), but because their widely recognized four-letter format provides an unusually compact representation that can convey coarse personality information and potentially serve as a semantic cue to an LLM.

We compare a Full condition using the original personality descriptions with an MBTI condition using MBTI-like labels, and investigate the following research questions.

- **RQ1 (Compression Rate):** To what extent can replacing personality information with MBTI-like labels reduce the number of characters and the number of input tokens in the utterance-generation prompt?
- **RQ2 (Character Likeness):** To what extent are utterances generated under the MBTI condition perceived by humans as being characteristic of the target character compared with those generated under the Full condition?
- **RQ3 (Character Identification):** To what extent can the original character be identified from utterances generated under the MBTI condition?

2 Related Work

2.1 Personality Conditioning in LLMs

A substantial body of work has explored role-playing with LLMs by providing them with persona settings or personality traits and prompting them to behave as specific people or characters. Tseng et al. (2024) organize persona research in the LLM era into two broad categories, role-playing and personalization, and survey the research trends

and evaluation methods in each area. Our study is closer to role-playing, in which an LLM is prompted to speak as a predefined game character.

Jiang et al. (2024) report that when LLMs are given personality profiles based on the Big Five, generated text exhibits features associated with those personalities, some of which are also perceived by human evaluators. This suggests that personality information provided in a prompt can be reflected in generated text.

2.2 Character Settings in AIWolf

AIWolfDial has investigated natural-language agents through the Werewolf game (Kano et al., 2024). In Werewolf, agents must not only maintain consistency in multi-party dialogue, but also perform reasoning, persuasion, and deception.

Tanaka et al. (2024) investigate an LLM-based AIWolf agent that uses dialogue summaries, manually defined character information, and example utterances to maintain contextual consistency and character-like speech.

2.3 Prompt Compression

LLMLingua is a prompt compression method that selects important information from long prompts to shorten the input text (Jiang et al., 2023). LLM-Lingua-2 formulates task-agnostic prompt compression as a learned token classification problem (Pan et al., 2024). These methods are fundamentally based on selecting and retaining important information from the original prompt.

In contrast, our approach does not retain the original personality description in shortened textual form, but instead maps it to a compact label such as INFP. The four-letter label itself does not contain all of the information in the original personality description, and different personality descriptions may map to the same label. Therefore, utterance generation is expected to rely not only on the information explicitly retained in the label, but also on personality-related knowledge acquired by the LLM during pretraining. In this respect, our method differs from conventional prompt compression approaches that selectively retain information from the original prompt.

2.4 Big Five and TIPI-J

The Big Five is a widely used personality model that characterizes personality in terms of traits such as Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness. The

Condition	Age	Gender	Personality information
Full	included	included	Original natural-language personality description
MBTI	included	included	Four-letter MBTI-like label only

Table 1: Full and MBTI conditions for utterance generation. All information other than the personality representation is kept identical.

Ten-Item Personality Inventory (TIPI) was proposed as a brief instrument for measuring the five Big Five dimensions using ten items (Gosling et al., 2003). Its Japanese version, TIPI-J, was developed and validated by Oshio et al. (2012).

In this study, we use the ten TIPI-J items as a reference when generating MBTI-like labels from natural-language personality descriptions.

3 Proposed Method

3.1 Full and MBTI Conditions

We compare a **Full** condition, in which the original natural-language personality description is used as is, with an **MBTI** condition, in which only the personality description is replaced with a four-letter MBTI-like label.

In AIWolfDial, each character is defined by attributes such as age, gender, and a natural-language personality description. Using Mio, one of the characters used in this study, as an example, the profile provided to the LLM for utterance generation in the Full condition is as follows.

age: 17
gender: 女性
personality: ミオは明るく元気な性格で、周りの人々を自然に笑顔にさせるタイプです。好奇心旺盛で、新しいことに挑戦するのが好き。少しおっちょこちょいなところもあり、失敗しても笑って許されることが多いです。自分の意見をしっかりと持っているが、他人の意見にも耳を傾けることができるバランスの取れた性格です。

English translation (not supplied to the model):

Mio is cheerful and energetic, and naturally makes people around her smile. She is curious and enjoys trying new things. She can be a little clumsy at times, but people often forgive her mistakes with a smile. She has her own opinions while also being able to listen to others, giving her a well-balanced personality.

In the MBTI condition, the natural-language personality field is removed and replaced with an

MBTI field containing a four-letter MBTI-like label. For Mio, the profile is as follows.

age: 17
gender: 女性
MBTI: ENFP
Here, 女性 means “female.”

By explicitly using the field name MBTI, we indicate to the LLM that the four letters represent an MBTI-like label. In the MBTI condition, no additional natural-language personality descriptions such as “extraverted” or “highly curious” are provided.

Age and gender are not compressed and remain identical between the Full and MBTI conditions; only the representation of personality information is changed.

Table 1 summarizes the two conditions.

3.2 Character Selection

We selected 10 characters from the character profiles used in the five-player track of AIWolfDial 2026.

To avoid strong biases toward particular ages or genders, we divided age into four groups: 0–12, 13–19, 20–39, and 40 years or older, and performed random sampling separately by gender. We selected one male and one female character from both the 0–12 and 13–19 groups, two male and two female characters from the 20–39 group, and one male and one female character from the 40-and-over group. The random seed was 20260817.

The 10 selected characters were Kenji, Rin, Daisuke, Mio, Ryuji, Toshio, Mei, Midori, George, and Shizue, comprising five male and five female characters.

3.3 Derivation of MBTI-like Labels

To derive four-letter MBTI-like labels from the original natural-language personality profiles, we conducted an independent LLM-based assessment. The evaluating LLM was given only the

personality field of each character; age, gender, game role, and game situation were not provided.

We used the following 10 items, based on the concepts underlying TIPI-J. For each item, the LLM assigned a score on a seven-point scale from 1 (strongly disagree) to 7 (strongly agree), and returned both the score and a brief rationale in JSON format.

1. Extraverted, enthusiastic
2. Critical, quarrelsome
3. Dependable, self-disciplined
4. Anxious, easily upset
5. Open to new experiences and ideas
6. Reserved, quiet
7. Sympathetic, warm
8. Disorganized, careless
9. Calm, emotionally stable
10. Conventional, uncreative

Reverse-scored items were transformed using $8 - x$, and the five Big Five dimensions were calculated as the mean of two corresponding items each.

$$E = \frac{x_1 + (8 - x_6)}{2}$$

$$A = \frac{(8 - x_2) + x_7}{2}$$

$$C = \frac{x_3 + (8 - x_8)}{2}$$

$$ES = \frac{(8 - x_4) + x_9}{2}$$

$$O = \frac{x_5 + (8 - x_{10})}{2}$$

Here, E denotes Extraversion, A Agreeableness, C Conscientiousness, ES Emotional Stability, and O Openness.

Four of these dimensions—Extraversion, Openness, Agreeableness, and Conscientiousness—were used to derive the four-letter label. Using 4, the midpoint of the scale, as the threshold, we defined the heuristic mapping shown in Table 2.

Dimension	> 4	< 4
Extraversion	E	I
Openness	N	S
Agreeableness	F	T
Conscientiousness	J	P

Table 2: Mapping from Big Five scores to MBTI-like letters.

Emotional Stability was not directly mapped to any of the four letters and was not used in deriving the MBTI-like label.

When a dimension used for the mapping had a score of exactly 4.0, we did not assign either side mechanically. Instead, we provided the original personality profile to the same LLM again and asked it to make a binary judgment as to whether the character was closer to the higher or lower side of that dimension. Among the 10 characters, this additional assessment was performed for Ryuji’s Agreeableness and for Toshio’s and Shizue’s Openness.

This mapping is not a psychometrically validated conversion from TIPI-J to MBTI. Rather, it is a heuristic rule defined in this study to obtain consistent, compact personality labels. The rationales generated by the LLM during the independent assessment and the resulting Big Five scores were not used in subsequent utterance generation. The only personality information provided to the utterance-generation LLM in the MBTI condition was the final four-letter MBTI-like label.

For the independent assessment, we used deepseek/deepseek-v4-flash-0731:deepinfra/fp8 via OpenRouter, with the provider fixed to DeepInfra. The temperature was set to 1.0, reasoning was enabled, and no explicit maximum output-token limit was specified. Each character was normally assessed once, with up to three retries only in cases of API errors or invalid output formatting. The same retry policy was used for the binary reassessment of dimensions with a score of exactly 4.0.

3.4 Construction of Evaluation Scenarios

The game situations used for utterance generation were constructed from actual logs of 30 games from the five-player track of AIWolfDial 2026.

The logs from all 30 games were provided to an LLM, which was instructed to select and construct 10 scenarios for use in the human evaluation. To avoid overrepresenting situations involving divination, we instructed it to include a range of situations, such as conversations at the beginning of a game, prompting other players to speak, questioning suspicious statements or voting behavior, discussing whom to eliminate, and making decisions on the final day.

To make the scenarios understandable even to participants unfamiliar with Werewolf, we avoided specialized terminology and represented each situ-

Evaluation target	Full condition	MBTI condition
Personality representation	personality: ...	MBTI: XXXX
Overall profile	Age, gender, and natural-language personality description	Age, gender, and MBTI-like label

Table 3: Comparison targets for the Full and MBTI conditions.

ation using the character’s role and approximately one sentence of contextual information sufficient for generating an utterance. Each description included only information that the target player could know at that point in the game; the true roles of other players and other unavailable information were excluded.

Each of the 10 selected characters was assigned to one of the 10 constructed scenarios. For each scenario, utterances were generated under both the Full and MBTI conditions, allowing us to vary only the representation of personality information while keeping the game situation fixed.

3.5 Utterance Generation

For each scenario, we generated one utterance under the Full condition and one under the MBTI condition, yielding 20 utterances in total across 10 scenarios and two conditions. Each utterance was generated through an independent API request. The generation model and all game-related information other than the personality representation were kept unchanged between conditions.

The utterance-generation prompt included the five players in the game, the current game situation, the target character’s profile, and the generation instructions. In the Full condition, the character profile contained age, gender, and personality; in the MBTI condition, it contained age, gender, and MBTI.

The model was instructed to speak in natural Japanese as the target character, not to introduce facts absent from the given game situation, not to reveal private role information inappropriately, to produce one natural utterance consisting of one or two sentences, and to keep it within 100 Japanese characters. It was also instructed not to explicitly mention meta-information such as MBTI, personality, profile information, or personality types in the generated utterance.

Utterance generation also used deepseek/deepseek-v4-flash-0731:deepinfra/fp8 via OpenRouter, with the provider fixed to DeepInfra. The temperature was set to 1.0 and reasoning was

enabled. The same model and generation settings were used in both the Full and MBTI conditions, with only the representation of personality information changed.

4 Evaluation Methods

4.1 Compression Rate Evaluation

We evaluate the compression effect of replacing natural-language personality descriptions with MBTI-like labels at three levels: the personality representation, the overall profile, and the full utterance-generation prompt. The comparison targets for the personality representation and the overall profile are shown in Table 3.

For the personality representation, we count not only the four-letter MBTI-like label itself, but also the MBTI: prefix that indicates its meaning. For the overall profile, we compare the number of characters in the actual profile strings supplied to the utterance-generation model, including age, gender, and personality information.

For the full utterance-generation prompt, we compare the input-token counts reported by the API during the actual DeepSeek generation requests.

For each character i , let $L_{i,\text{Full}}$ denote the value under the Full condition and $L_{i,\text{MBTI}}$ the value under the MBTI condition. We define the reduction amount D_i and reduction rate R_i as

$$D_i = L_{i,\text{Full}} - L_{i,\text{MBTI}}$$

$$R_i = 1 - \frac{L_{i,\text{MBTI}}}{L_{i,\text{Full}}}$$

For each of the 10 characters, we calculate the reduction amount and reduction rate, and report their mean, minimum, and maximum values.

4.2 Character Likeness Evaluation

For each evaluation item, the following information about the target character was presented:

- The characters involved, along with relationships or claims relevant to the evaluation at that point
- The current game situation
- The age of the target character
- The gender of the target character
- The original natural-language personality profile
- The in-game role of the target character
- One generated utterance

Regardless of the condition under which an utterance was generated, evaluators were shown the target character’s original natural-language personality profile. Thus, even for utterances generated under the MBTI condition, evaluators saw the original natural-language personality description rather than the four-letter MBTI-like label.

The evaluators were not shown the MBTI-like label, nor were they informed whether an utterance had been generated under the Full or MBTI condition.

To ensure that participants without prior experience with Werewolf could understand the situations, the survey began with brief explanations of the three roles used in the study: Villager, Seer, and Werewolf.

Participants were asked:

“Given the game situation described above, do you think this utterance is characteristic of the person described in the profile above?”

Responses were collected on a four-point scale from 1 (“Not at all”) to 4 (“Very much”), with an additional “Cannot judge” option. “Cannot judge” responses were excluded from the primary character-likeness score and reported separately. Participants also rated their confidence as High, Medium, or Low; confidence was used only as supplementary information.

To avoid direct comparison of the Full and MBTI utterances for the same context, we created two complementary survey forms, X and Y. Each form contained one utterance from each of the 10 contexts, with five Full and five MBTI utterances. The condition assignment was reversed between the two forms, so that no participant evaluated both conditions for the same context. Participants were assigned using Google Apps Script, preferentially to the form with fewer consenting responses and randomly when the counts were equal.

We recorded participant category, whether Japanese was the participant’s native language,

JLPT level for non-native speakers, and use of translation tools. Participants were asked, where possible, to complete the survey without translation tools, although their use was not prohibited.

For analysis, we calculated each participant’s mean character-likeness score separately for the Full and MBTI conditions and used these participant-level means as the unit of analysis. We report the mean and median for each condition. The within-participant difference was defined as the MBTI-condition mean minus the Full-condition mean. Cohen’s d_z was calculated as the effect size, and statistical significance was assessed using a two-sided exact sign-flip permutation test over all possible sign reversals of the within-participant differences. Because the survey used a small convenience sample, the results are interpreted exploratorily.

4.3 Character Identification Evaluation

We evaluated whether the original character could be identified from utterances generated under the Full and MBTI conditions. The evaluation set consisted of 20 utterances: one Full-condition and one MBTI-condition utterance for each of the 10 characters.

For each utterance, an evaluation model ranked the three most likely speakers from 10 candidate profiles containing age, gender, and the original natural-language personality description. Character names were anonymized as Candidate A–J. The game situation, in-game role, generation condition, and MBTI-like label were not shown.

To reduce positional effects, we randomized the candidate order using a fixed seed (20260820) and created 10 arrangements by cyclically shifting the order, so that each character appeared exactly once in each candidate position. The same arrangements were used across conditions and evaluation models.

We used four evaluation models: DeepSeek V4 Flash 0731, Xiaomi MiMo-V2.5, MoonshotAI Kimi K3, and Poolside Laguna XS 2.1. DeepSeek, which was also used for utterance generation, was treated separately as a same-model reference; the other three served as external evaluators. Each model evaluated all 20 utterances under all 10 candidate arrangements, yielding 200 evaluations per model and 800 in total. The temperature was 1.0 for all models, with reasoning enabled.

Our primary metric was Top-1 Accuracy, with Top-3 Accuracy as a supplementary metric. Accu-

Evaluation target	Full mean	MBTI mean	Mean reduction rate
Personality representation (characters)	206.50	10.00	94.98%
Overall profile (characters)	225.40	28.90	86.80%
Full prompt (tokens)	505.30	379.90	24.59%

Table 4: Compression results for the Full and MBTI conditions.

Condition	Mean	Median	Cannot judge	High confidence
Full	3.03	3.40	0.00%	28.57%
MBTI	3.38	3.25	2.86%	28.57%

Table 5: Human evaluation results on character likeness.

racy was calculated for each character over the 10 candidate arrangements, and the Full and MBTI conditions were compared using the 10 characters as the unit of analysis.

For the three external models—MiMo, Kimi, and Laguna—we additionally calculated consensus accuracy. A consensus success was recorded when at least two of the three models selected the correct character, separately for Top-1 and Top-3.

5 Experimental Results

5.1 Compression Results

As shown in Table 4, the MBTI condition reduced the personality representation by 94.98%, the overall profile by 86.80%, and the full utterance-generation prompt by 24.59% on average. The mean reduction in the full prompt was 125.40 input tokens per generation.

5.2 Character Likeness Results

There were seven valid participants: three completed Form X and four completed Form Y. Four participants were students and three were teachers or instructors. Six were native speakers of Japanese, while one was a non-native speaker with JLPT N1 certification. None of the seven participants used translation tools. Because this survey used a small convenience sample, the following human-evaluation results are interpreted exploratorily.

A total of 70 ratings were collected from the seven participants. One response was “Cannot judge,” leaving 69 responses for the primary character-likeness analysis. The proportion of “Cannot judge” responses was 0% in the Full condition and 2.86% in the MBTI condition.

For each participant, we calculated the mean character-likeness score separately for the Full and MBTI conditions. The mean of these participant-level scores was 3.03 for the Full condition and 3.38 for the MBTI condition. The median participant-level score was 3.40 for the Full condition and 3.25 for the MBTI condition. These human-evaluation results are summarized in Table 5.

The mean participant-level difference (MBTI – Full) was +0.35. Cohen’s d_z for the within-participant difference was 0.59, and a two-sided exact sign-flip permutation test yielded $p = 0.1875$. Thus, the test did not detect a statistically significant difference between the two conditions.

5.3 Character Identification Results

For the consensus of the three external models, Top-1 Accuracy decreased from 51% in the Full condition to 33% in the MBTI condition, while Top-3 Accuracy decreased from 84% to 74%. Both remained above the random baselines of 10% and 30%, respectively. Results for individual evaluation models are shown in Table 6.

The effect varied substantially across characters. Under the MBTI condition, the external-model consensus achieved 100% Top-1 Accuracy for Mio, whereas it was 0% for Midori. For Ryuji and Toshio, Top-1 Accuracy was 50% and 80%, respectively, while Top-3 Accuracy was 100% for both characters.

Evaluation model	Full Top-1	MBTI Top-1	Full Top-3	MBTI Top-3
DeepSeek	58%	28%	84%	72%
MiMo	55%	43%	80%	71%
Kimi	60%	37%	89%	78%
Laguna	27%	25%	64%	60%
External consensus	51%	33%	84%	74%

Table 6: Character identification accuracy for each evaluation model.

6 Discussion

6.1 Using Four Letters as a Semantic Key

Our approach does not assume that a four-letter MBTI-like label contains the same information as the original personality description. Instead, the label serves as a compact “semantic key” that prompts the LLM to draw on personality-related knowledge acquired during pretraining. Unlike prompt-compression methods that retain selected information from the original text, this approach cannot recover traits that are neither encoded by the label nor associated with it in the model’s pretrained knowledge. This compression therefore entails an inherent loss of detail. We did not evaluate reductions in generation time or computational cost.

6.2 Character Likeness and Identifiability Are Not the Same

Human evaluation and character identification capture different aspects of information preservation. In the small human evaluation, we did not detect a statistically significant difference in perceived character likeness between conditions, and the sample was too small to establish equivalence. In contrast, external-model consensus showed lower Top-1 and Top-3 identification accuracy under the MBTI condition. These results suggest that broad personality tendencies may remain sufficient for an utterance to appear characteristic of a target character even when finer-grained information needed to distinguish that character from others is lost.

6.3 Individuality After Compression and Its Expression in Utterances

Identification performance varied substantially across characters. Mio reached 100% Top-1 Accuracy under the MBTI condition, whereas Midori reached 0%. Ryuji and Toshio are especially informative: although their original natural-language personality descriptions differ, both were mapped

to ISTJ and therefore have identical compressed profiles in terms of age, gender, and personality representation. Their Top-1 accuracies were 50% and 80%, respectively, while Top-3 Accuracy was 100% for both. This result is consistent with the four-letter label serving as a coarse shared representation of personality information present in different original descriptions, while finer-grained information needed to distinguish between the corresponding characters is lost. At the same time, identifiability may also depend on which traits are expressed in a particular utterance and dialogue situation.

6.4 Why AIWolf Is a Meaningful Evaluation Setting

Werewolf requires agents to balance character-like expression with reasoning, persuasion, cooperation, and deception, making it a challenging setting for testing whether character likeness persists in task-oriented dialogue. Because profiles are reused across turns, compression can also yield cumulative token savings and leave more of the context window available for dialogue history and game state.

7 Conclusion

In this study, we investigated a method for substantially compressing personality information provided to LLM-based dialogue agents by replacing natural-language personality descriptions with four-letter MBTI-like labels while retaining age and gender. These four letters are not treated as psychological diagnoses, but as compact semantic keys intended to evoke personality-related knowledge acquired by LLMs during pretraining.

Using Japanese AIWolf agents, we compared a Full condition using the original natural-language personality descriptions with an MBTI condition using MBTI-like labels. We evaluated the two conditions from three perspectives: compression

rate, human judgments of character likeness, and character identification from generated utterances.

The personality representation was reduced by an average of 94.98%, the overall profile by 86.80%, and the total input-token count of the utterance-generation prompt by 24.59%. In the small exploratory human evaluation, the mean participant-level character-likeness score was 3.03 in the Full condition and 3.38 in the MBTI condition, and we did not detect a statistically significant difference between the two conditions. In contrast, character identification based on the consensus of three external models decreased from 51% to 33% for Top-1 Accuracy and from 84% to 74% for Top-3 Accuracy.

Taken together, these results are consistent with the possibility that four-letter compression affects fine-grained character identifiability more strongly than perceived character likeness. More generally, the results suggest that prompt compression may be achieved not only by selectively retaining information from the original text, but also by using compact expressions as semantic cues to knowledge that an LLM may have acquired during pretraining.

This study does not demonstrate that “character likeness” itself can be stored and reconstructed as external data. However, it provides a preliminary example of how some character-related information may be retained in a lightweight external representation and reconstructed through an LLM.

Limitations

First, four-letter MBTI-like labels are coarse many-to-one representations, and different natural-language personality descriptions may map to the same label. Fine-grained individual differences in the original profiles therefore cannot be uniquely reconstructed from the four letters alone. In particular, characters sharing the same MBTI-like label may lose information needed for individual identification. A small number of auxiliary attributes could potentially reduce this loss, but this would introduce a trade-off between compression rate, character likeness, and identifiability.

Second, the MBTI-like labels used in this study were generated through a heuristic procedure in which an LLM rated natural-language personality descriptions with reference to the concepts underlying TIPI-J, and the resulting scores were mapped to four letters using rules defined in this study. This procedure is not a psychometrically validated

MBTI assessment. The resulting labels may also depend on the knowledge and biases of the LLM used for the assessment, and different models or mapping procedures could produce different labels.

Third, the human evaluation involved only seven participants from a small convenience sample, limiting both statistical power and participant diversity. Although no statistically significant difference in character likeness was observed between the Full and MBTI conditions, this does not establish equivalence between the two conditions. The human-evaluation results should therefore be regarded as exploratory and require validation with larger and more diverse participant samples.

Fourth, the personality characteristics expressed in Werewolf utterances depend on the game situation, role, and dialogue flow. Even when a characteristic is present in the profile, it may not be observable from a single short utterance if the situation does not provide an opportunity for that characteristic to be expressed. Because our evaluation used a limited set of game situations and one utterance per condition for each character, the results may not generalize to other dialogue contexts, longer interactions, or tasks beyond Werewolf. Future work should generate multiple utterances for each character and condition to examine whether character likeness is consistently preserved across repeated utterances.

Fifth, this study evaluated compressed utterances in terms of perceived character likeness and character identification, but did not directly measure how individual personality traits in the original descriptions were preserved or changed. Consequently, the present results do not show which specific traits, such as Extraversion, Agreeableness, Conscientiousness, Emotional Stability, or Openness, are retained or lost through compression. A trait-level analysis based on the original profiles and generated utterances would be necessary to address this question. Future work could also compare the original and compressed conditions using the intermediate Big Five representations derived from the TIPI-J-inspired items used in this study. This would help examine how the compressed labels affect individual personality dimensions.

Finally, our approach is inherently model-dependent because it relies on knowledge and associations that an LLM has acquired during pretraining for MBTI-like labels. Different LLMs may

construct different character representations from the same label, and cultural biases or stereotypes associated with personality types in the training data may be reflected in generated utterances. Because utterance generation in this study used a single model and Japanese-language dialogue, the generalizability of the findings across generation models, languages, cultural contexts, and other types of semantic keys remains uncertain.

Acknowledgements

The authors received assistance from OpenAI's ChatGPT and Google's Gemini in preparing this manuscript. The first author is currently also enrolled at Tokyo Online University.

Ethical Considerations

The MBTI-like labels used in this study are applied only to predefined game characters and are not intended for psychological diagnosis or personality classification of real individuals. We also do not claim that the MBTI-like labels used here constitute scientifically validated categorical representations of personality for real people.

In the human evaluation, participants evaluated generated game utterances and were not required to provide information about their own personalities. Before participation, they were informed of the purpose of the study and how their responses would be used, and only responses from participants who provided consent were included in the analysis. Participation was voluntary, and personally identifiable information unnecessary for the analysis, such as email addresses, was not collected.

The knowledge acquired by LLMs during pre-training about MBTI-like labels may contain cultural biases or stereotypes. Therefore, the findings of this study are limited to character representation in LLM-based agents and are not intended for applications that infer the personality, aptitude, mental state, or other personal attributes of real individuals.

Supplementary Material Statement

Data Availability

The data generated in this study are publicly available. The supplementary materials are available at <https://doi.org/10.5281/zenodo.22796192>.

For data originating from third parties, including the original AIWolf game data, only materials

that can be redistributed under the applicable terms and conditions are included.

Reproducibility Materials

To support reproducibility, we have released the following materials on Zenodo to the extent permitted by redistribution rights:

- Source code and configuration files used in this study
- Outputs produced by the programs used in the experiments
- Survey responses with personal information and potentially identifying information removed

The released source code includes the procedures used for deriving MBTI-like labels, generating utterances, measuring compression rates, performing character identification, and conducting the analyses. The program outputs include the generated utterances and evaluation results necessary to reproduce the main quantitative results reported in this paper.

References

- Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. 2003. [A Very Brief Measure of the Big-Five Personality Domains](#). *Journal of Research in Personality* 37(6):504–528.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. [Person-aLLM: 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.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. [LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore.
- Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. [AIWolfDial 2024: Summary of Natural Language Division of 6th International AIWolf Contest](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12, Tokyo, Japan.
- Robert R. McCrae and Paul T. Costa. 1989. [Reinterpreting the Myers-Briggs Type Indicator from the Perspective of the Five-Factor Model of Personality](#). *Journal of Personality* 57(1):17–40.

- Atsushi Oshio, Shingo Abe, and Pino Cutrone. 2012. Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J). *The Japanese Journal of Personality* 21(1):40–52.
- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. [LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 963–981, Bangkok, Thailand.
- Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki, Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi, Tomoya Higuchi, Ryutaro Asahara, and Michimasa Inaba. 2024. [Enhancing Consistency of Werewolf AI through Dialogue Summarization and Persona Information](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 48–57, Tokyo, Japan.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA.