Who Laughs with Whom? Disentangling Influential Factors in Humor Preferences across User Clusters and LLMs

Anonymous ACL submission

Abstract

Humor preferences vary widely across individuals and cultures, complicating the evaluation of humor using large language models (LLMs). In this study, we model heterogeneity in humor preferences in Oogiri, a Japanese creative response game, by clustering users with voting logs and estimating cluster-specific weights over interpretable preference factors using Bradley-Terry-Luce models. We elicit preference judgments from LLMs by prompting them to select the funnier response and found that user clusters exhibit distinct preference patterns and that the LLM results can resemble those of particular clusters. Finally, we demonstrate that, by persona prompting, LLM preferences can be directed toward a specific cluster. The scripts for data collection and analysis will be released to support reproducibility.

1 Introduction

Large language models (LLMs) have garnered significant attention, as they are capable of creative reasoning akin to humans. Humor understanding and generation, which require contextual and nuanced understanding, provide a useful testbed for evaluating the creative capabilities. However, because humor is highly subjective and exhibits cultural dependence, the quantitative measurement and replication of funniness have long been a challenging problem in natural language processing (Loakman et al., 2025). In this study, we focus on Oogiri, a Japanese creative response game. In Oogiri, which is characterized by a question-answer style of humor, participants produce witty responses to a given prompt.

Improving humor understanding and response generation requires both clarifying human humor preferences and identifying the gaps between LLMs and humans. In this study, we analyze the factors underlying human humor preferences and examine the gaps in humor preferences between

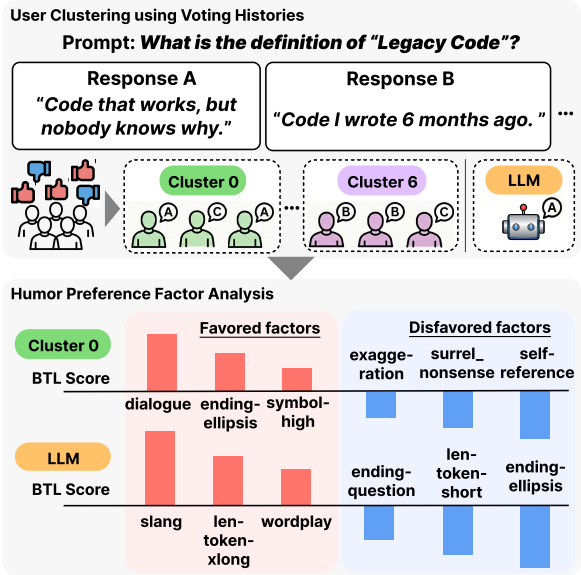


Figure 1: Overview of humor-preference factor analysis across user clusters and LLMs. Users are clustered based on their voting history, and humor-preference factors are analyzed using the Bradley-Terry-Luce model.

LLMs and humans, which have already been investigated in several studies, while two challenges still remain. First, previous analyses of human humor preferences do not consider individual preferences (Murakami et al., 2025; Sakabe et al., 2025). Previous work typically collected ratings from multiple human annotators for each humorous text and aggregated them into an "overall score" by averaging. However, in subjective evaluation tasks, including humor judgments, inter-annotator agreement is often low (Celikyilmaz et al., 2021). Thus, it is reasonable to assume that humor preferences differ across users (Chakrabarty et al., 2019; Zhang et al., 2024). Given the subjectivity of humor and its cultural dependence, humor research should go beyond aggregated preferences and account for individual preferences (Cabitza et al., 2023).

Second, the gaps between LLM and individual human humor preferences remain still unclear. Sak-

abe et al. (2025) analyzed the differences between human and LLM humor preferences, showing that humans prioritize *empathy*, whereas LLMs prioritize *novelty*. However, because their analysis relied on the "aggregated humor preferences" obtained from multiple user ratings and compared only LLMs with the overall user population, it neither accounted for individual user preferences nor examined the alignment between individual human and LLM preferences. Therefore, they remain still unclear whether the LLM humor preferences are intrinsically different from those of humans and whether they align with those of certain users.

To address these issues, we analyze the factors affecting humor preferences at the user cluster level, using the voting data collected from the Oogiri platform (Figure 1). Because each vote is associated with a user ID, we represent each user with a voting history-based vector and cluster these vectors to identify the groups with similar preferences. For each cluster, we fit a Bradley–Terry–Luce (BTL) model (Bradley and Terry, 1952) to estimate the weights of preference factors, including linguistic features (e.g., length) and humor-strategy labels (e.g., black joke). To enable a direct comparison with LLMs' preferences, we additionally construct LLM preference data by asking the model to select the funniest response for each prompt. The BTL model is then fitted to the LLM data using the same set of factors, and the resulting weights are compared with those of each user cluster. This analysis quantifies the preference gaps between LLMs and user clusters, and tests whether any clusters exhibit preference patterns aligned with those of LLMs.

Based on this user cluster-aware analysis, we investigate the following research questions:

- RQ1: What kinds of humor do different user clusters prefer?
- RQ2: Do LLMs align with the preferences of any specific user clusters?
- RQ3: Can we align LLM preferences with those of a specific user cluster?

In answering these questions, our contributions are four-fold. First, we quantified the variation in humor preferences across user clusters by estimating cluster-specific preference factors. Second, we characterized differences between LLM and human humor preferences in aggregate and identified cases in which the LLM results resemble particular clusters in terms of factor weights. Third, we demonstrated that persona prompting can increase the alignment between an LLM and a specific user cluster. Finally, we will release our data collection and analysis scripts to support reproducibility upon publication. We hope that our findings will provide useful insights for advancing humor understanding and generation that accounts for heterogeneous humor preferences.

2 Related Work

With the rapid progress in LLMs, humor understanding and generation have received renewed attention (Amin and Burghardt, 2020). Humor spans diverse forms, such as puns and irony, and corresponding datasets and methods have been developed (Ritchie, 2005; Hossain et al., 2022; Hessel et al., 2023). This study focuses on Oogiri, a traditional Japanese humor format. Oogiri is a question-answer style of humor, whereby a witty response is produced for a given prompt. Recently, online Oogiri platforms have emerged. Zhong et al. (2024) collected data from Bokete[1] to build the Oogiri-GO dataset, thereby accelerating research in this area (Sakabe et al., 2025; Murakami et al., 2025).

Oogiri is related to "Caption This" (e.g., The New Yorker's Cartoon Caption Contest), where a participant creates a funny caption for a given panel or photo (Hessel et al., 2023; Tanaka et al., 2024; Zhang et al., 2024). Although both *Caption This* and Oogiri require the generation of a witty response conditioned on a prompt, Oogiri in this work is text-based, whereas *Caption This* is multimodal. However, Zhou et al. (2025) reported that state-of-the-art LLMs often struggle with visually grounded humor, which can undermine the reliability of LLM-based preference estimation in multimodal settings. Accordingly, we focused on text-based Oogiri, deferring multimodal settings to future studies.

Quantifying the factors that shape humor preferences can help advance research on humor understanding (Murakami et al., 2025). In particular, contrasting human and LLM preferences provides a clearer picture of current LLM humor understanding and potential improvements. Recently, Sakabe et al. (2025) analyzed these differences and found that humans prefer *empathy*, whereas LLMs prefer *novelty*; however, their analysis was based on humor preferences aggregated across multiple users and thus did not address how well LLM preferences

[1] https://bokete.jp

align with those of individual users. In a related study, Chakrabarty et al. (2019) constructed user clusters from joke voting data for recommender systems but did not analyze cluster-specific preference patterns in detail. Therefore, we analyze humor-preference factors across user clusters to characterize the humor preferences of LLMs and the cluster-level differences.

3 Construction of Analytical Dataset

Existing Oogiri datasets consist of prompt-response pairs and vote counts, indicating the funniness of each response; however, they lack information on which user voted for which response, making it impossible to track individual humor preferences. Therefore, we extended the existing Japanese Oogiri dataset, the Oogiri-Corpus (Murakami et al., 2025), to construct a new dataset that includes user-level voting data for each response.

Source Dataset The Oogiri-Corpus is a prompt-response pair dataset collected from the Oogiri platform, Oogiri Sogo,[2] comprising 908 prompts and 82,536 responses, each associated with a vote count. Users can cast up to three votes per prompt on this platform. The three votes can be distributed flexibly; for example, a user may allocate two votes to one response and one vote to another. Because the user IDs were attached to the voting data, the voting histories of individual users can be tracked.

Dataset Construction Process We built the analytical dataset with user-level voting data in two steps: (1) web-crawling the votes, including user IDs for all prompt-response pairs from the source site, and (2) filtering. To ensure reliability, active users with at least 100 total votes were first selected, ensuring that their humor preferences are well represented. Only the votes from these active users were retained, recomputing each response's vote count. Responses with fewer than three votes were then removed to maintain high response quality.

Dataset Statistics The dataset contained 908 prompts, 14,389 responses, and 57,751 votes from 276 users (35.6 users per prompt on average).

4 Method

Figure 1 summarizes our analysis pipeline. To account for heterogeneity in humor preferences, we analyzed preference factors at the user-cluster level. We first constructed user representations and clustered users based on their voting histories (§4.1), then defined humor preference factors for each prompt-response pair (§4.2), and finally estimated factor weights with an BTL model (§4.3).

4.1 User Representation and Clustering

The users were clustered based on their voting histories. Each user u is represented by a sparse voting history vector $\mathbf{x}_u \in \mathbb{R}^N$, where N is the number of responses in the dataset and $\mathbf{x}_u[i]$ is the number of votes cast by u for response i (0 if never selected). To reduce the influence of responses frequently chosen by many users, we applied term frequency-inverse document frequency (TF-IDF) reweighting to obtain $\tilde{\mathbf{x}}_u$. Subsequently, to mitigate sparsity, we computed a 100-dimensional representation by applying a truncated singular value decomposition (SVD), $\mathbf{z}_u = \mathrm{SVD}_{100}(\tilde{\mathbf{x}}_u)$, and normalized it as $\mathbf{y}_u = \mathbf{z}_u / \|\mathbf{z}_u\|_2$ to control for scale differences across users. Finally, we clustered $\{\mathbf{y}_u\}$ using K-means clustering. We report the clustering configuration and robustness analyses in Appendix A.

4.2 Humor Preference Factors

We designed two main types of features that influence user humor preferences. Table 1 summarizes the humor preference factors used in this study. The first group consists of linguistic features, which are basic features extracted directly from the prompts and responses (§4.2.1). We defined 45 linguistic features. The second group comprises humor strategy types that capture more nuanced aspects of humor (§4.2.2). Specifically, we assigned 11 different humor strategy labels to each prompt-response pair, covering various humor strategies such as "black joke" and "parody." We defined these features with reference to prior research on humor factor analysis (Murakami et al., 2025) and humor theories (Morreall, 2024). An overview of each feature group is provided below, with detailed definitions available in Appendix B.

4.2.1 Linguistic Features

We designed 45 linguistic features, ranging from basic features, such as text length and part-of-speech (POS) ratios, to features based on the relationship between the prompt and response. These features were categorized into six subgroups: basic, morphological analysis, sentence-ending pattern, special symbol, writing-style, and prompt-response relational features. The first five subgroups are

[2] https://chinsukoustudy.com

Group	#	Brief definition
Linguistic Features		
Basic	11	Surface statistics of each response (e.g., character count, character type ratio).
Morphological	10	Morphological analysis features of each response (e.g., part-of-speech ratio, word count).
Special symbols	5	Usage of special symbols in each response (e.g., quotation marks, parentheses, slang phrases).
Sentence-ending	9	Sentence-ending patterns of each response (e.g., whether it ends with a symbol (?, !)).
Writing style	4	Japanese writing-style indicators in each response (e.g., polite/casual forms, exaggeration).
Relational	6	Features capturing relations between a prompt and a response. (e.g., length ratio)
Humor strategy	11	Interpretable multilabel annotations derived from humor theories to capture nuanced humor beyond linguistic features (e.g., `incongruity`, `black_joke_satire`, `self_reference`.)

Table 1: Summary of humor preference factors used in our BTL analysis. A full list of features and their definitions is provided in Appendix B.

based on the linguistic characteristics of the responses, whereas the last subgroup is based on the relationship between the prompt and response.

4.2.2 Humor Strategy Labels

To capture more nuanced aspects of humor that cannot be represented by linguistic features alone, we annotated each prompt-response pair with humor strategy labels. We defined the 11 strategy labels based on humor theories (Morreall, 2024) and prior research (Murakami et al., 2025). Theories of humor, which aim to explain the essence of humor and the mechanisms by which humans perceive humor, have been studied in fields such as psychology, sociology, and linguistics. For example, `incongruity` represents unexpected twists and surprising connections, aligning with *incongruity theory* (McDonald, 2013), whereas `black_joke_satire` encompasses dark humor and social commentary, relating to *benign violation theory* (McGraw and Warren, 2010).

Annotation Process We annotated all prompt responses using GPT-5.1. We carefully designed the system prompts, including detailed definitions and annotation examples for each label. To ensure reliability, we followed a self-consistency protocol (Wang et al., 2023), conducting three trials per response and determining the final labels by majority voting. Each response can be assigned multiple labels. We additionally validated the labels via a crowdsourced study (81.4% overall accuracy); see Appendix B.2 for details and examples.

4.3 Preference Modeling

We formalized our approach to analyze humor preference factors by building it on the DecipherPref framework (Hu et al., 2023). DecipherPref converts features into categorical and interpretable factors and applies the BTL model (Bradley and Terry, 1952; Luce et al., 1959) to analyze the factors that influence pairwise preference judgments in a summarization task (e.g., length, linguistic quality, content accuracy). In our setting, pairwise preference judgments were modeled as comparisons between the above factors, and each factor's relative strength was estimated using the BTL model.

Let $\mathcal{D} = \{(p_i, \mathbf{r}_i, \mathbf{v}_i)\}_{i=1}^{M}$ be a collection of M prompts, where p_i is the i-th prompt; $\mathbf{r}_i = (r_i^1, r_i^2, \ldots, r_i^{n_i})$ is the response list for prompt p_i; and $\mathbf{v}_i = (v_i^1, v_i^2, \ldots, v_i^{n_i})$ is the vote-count list such that $v_i^j \in \mathbb{N}$ is the vote count for response r_i^j; and n_i is the number of responses for prompt p_i.

We defined a finite vocabulary $\mathbb{F}$ of K interpretable factors to characterize the prompt-response pairs. Each factor $f \in \mathbb{F}$ represents a categorical property of a prompt-response pair. Specifically, we considered two types of factors: (i) linguistic features such as character length and punctuation patterns (§4.2.1) and (ii) humor strategy labels (§4.2.2). For continuous-valued linguistic features, we discretized the values into quartile-based categorical bins following DecipherPref (Hu et al., 2023); for example, character length is represented as `len-char-{short|medium|long|xlong}` (see Appendix B.1.2 for details). For any response r, let $f(r) \subseteq \mathbb{F}$ be the subset of factors present in r. Factors are binary or multilevel categorical variables, and in the multilevel case, each level is treated as a separate binary factor.

We derived pairwise comparisons from vote counts. For each prompt p_i, we compared all pairs of responses (r_i^j, r_i^k) where $j \neq k$; if $v_i^j > v_i^k$, we treated r_i^j as the winner and r_i^k as the loser, and discarded ties ($v_i^j = v_i^k$), as they provide no preference information. For each winner-loser pair (r_i^j, r_i^k), factor-level comparisons were extracted by defining

$$F_i^+ = f(r_i^j) \setminus f(r_i^k), \qquad F_i^- = f(r_i^k) \setminus f(r_i^j),$$

where F_i^+ and F_i^- denote the factors unique to the winner and loser, respectively. Factors appearing in both responses were ignored because they do not provide explanatory power for the preference judgment. Each factor $f \in F_i^+$ was treated as having "beaten" every factor $g \in F_i^-$. Thus, one response-level comparison yielded $|F_i^+| \times |F_i^-|$ pairwise outcomes at the factor level.

The relative strength of the factors was modeled using the BTL model. Each factor $k \in \mathbb{F}$ was associated with a real-valued parameter $\theta_k \in \mathbb{R}$. For any ordered pair of factors (k, ℓ), the probability that factor k beats factor ℓ is:

$$P(k \succ \ell) = \frac{\exp(\theta_k)}{\exp(\theta_k) + \exp(\theta_\ell)}.$$

This formulation assumes that the probability of one factor beating another depends only on the difference $\theta_k - \theta_\ell$. The parameters $\hat{\boldsymbol{\theta}}$ were estimated using the Luce spectral ranking (LSR) algorithm (Maystre and Grossglauser, 2015), which is a spectral estimator for models based on Luce's choice axiom (Luce et al., 1959). In our implementation, we used the `choix` library[3] to compute the spectral ranking. The estimated parameters $\hat{\theta}_k$ provide a ranking of the factors based on their influence on human preferences. Factors with higher $\hat{\theta}_k$ values are strongly associated with preferred responses, whereas factors with lower values are associated with less-preferred responses.

5 LLM Preference Data Collection

To analyze the gaps between LLM and human preferences for RQ2, we defined a funniest response selection task to collect LLM humor preferences.

Task Definition Let $\mathcal{P}$ denote the set of prompts. For each prompt $p \in \mathcal{P}$, the LLM is presented with a set of candidate responses $\mathcal{R}(p) = \{r_{p,1}, \ldots, r_{p,K_p}\}$ and instructed to select exactly one response that it finds the funniest.[4] The selection is recorded as index $y_p \in \{1, \ldots, K_p\}$ and the chosen response as $\hat{r}_p = r_{p,y_p}$. Repeating this procedure over prompts yields a dataset $\{(p, \mathcal{R}(p), y_p)\}_{p \in \mathcal{P}}$, in which each entry consists of a prompt, its associated response set, and the index of the response selected by the LLM as the funniest.[5] This dataset effectively captures the humor preferences of the LLM, enabling us to analyze and compare the LLM's humor preferences with human humor-preference data.

Dataset For this task, we constructed an evaluation dataset from the analysis dataset described in §3. To avoid unreliable evaluations with insufficient choices, prompts with fewer than five responses were excluded from the analysis dataset. Consequently, the evaluation dataset comprised 897 unique prompts and 14,352 responses with an average of approximately 16 responses per prompt.

Models We analyzed humor-preference variation across three state-of-the-art LLMs: Gemini 3 Pro, GPT-5.1, and Claude Sonnet 4.5.

Persona Prompting Prior research has shown that changing the prompts provided to LLMs can affect their response characteristics (He et al., 2024). In this preference collection process, we used persona prompting (Tseng et al., 2024) as a case study, to investigate how different personas influence humor preferences. Specifically, we hypothesized that explicitly assigning different personas to LLMs may influence their humor preferences. If humor preferences vary by persona, we assume that each persona can potentially align with the preferences of specific user clusters. Thus, we analyzed the relationship between the humor preferences of LLM personas and user clusters. Specifically, we defined seven personas: `{male,female}_20`, `{male,female}_45`, `{male,female}_65`, and `no_persona`. For example, the `male_20` persona represents a 20-year-old male university student well-versed in trending jokes, memes, and internet vocabulary. We also included `no_persona` as a control condition, whereby no specific persona is provided for the system prompt. To implement these personas, unique prompts were incorporated for each persona into the system prompt. We carefully designed these persona prompts to reflect the characteristics and perspectives associated with each persona. The details of the persona prompts are presented in Appendix C.

[3] https://choix.lum.li/

[4] For BTL estimation, we converted each selection into a pairwise comparison by treating the selected response as the winner and each non-selected response as the loser.

[5] To address API variability, we called the API three times per prompt with randomly permuted response orders; see Appendix C. Self-consistency of these three calls is reported in Appendix D.

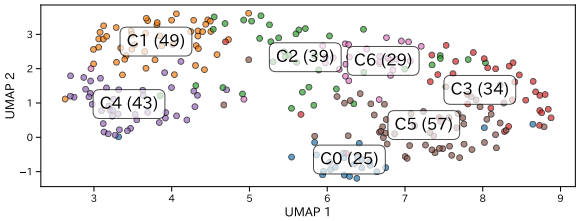


Figure 2: UMAP visualization of user clusters. Different colors represent different clusters. C0 to C6 denote the respective user clusters and the number of users in each cluster is indicated in parentheses.

6 Experimental Results

In the experiments, we analyzed the factors underlying humor preferences for each user cluster and LLM using the BTL model. This section presents the experimental results that answer the three research questions introduced in §1. We first report the results of user clustering based on voting histories (§6.1) and then present the results of the humor preference factor analysis (§6.2).

6.1 Results of User Clustering

The users were clustered based on their voting histories using K-means clustering. We identified $K = 7$ user clusters, where K was selected based on the elbow method (Thorndike, 1953) and silhouette scores (Rousseeuw, 1987). In the subsequent analysis (§6.2), we present the results based on these clusters.

For visualization, we embedded the users into two dimensions using uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) (see Figure 2). Although clusters are observable, their boundaries are not sharply separated, consistent with a low silhouette score ($s = 0.025$); nevertheless, we show that clusters exhibit distinct humor-preference factors in §6.2.1

6.2 Analysis of Humor Preference Factors

6.2.1 RQ1: Humor Preference Factors across User Clusters

To answer RQ1, we report the humor preference factors favored by each user cluster. Table 2 presents the top-3 and bottom-3 factors for each user cluster based on BTL scores, and in Figure 3 the BTL scores are visualized for a broader set of factors as a heatmap (including those for the LLM) to provide a more detailed view. Our key findings from user cluster analysis are as follows:

User Clusters Exhibit Distinct Preferences Each user cluster exhibited a distinct humor preference factor. For example, as shown in Table 2 and Figure 3, C0 prefers longer responses that contain dialogue (`dialogue`) and multiple sentences (`sentences-many`), whereas C5 tends to dislike overly long responses (`len-char-xlong`). Additionally, factors such as slang and self-deprecating humor show significant variations in preferences across clusters. For instance, C1 favors self-deprecating humor (`self_reference`), whereas C0 tends to dislike it. This indicates that the influence of humor preference factors varies according to user cluster, reflecting the subjectivity of humor preferences.

Common Preference Factors also Exist across User Clusters Some factors exhibited consistent BTL scores across user clusters. For example, wordplay-based humor (`wordplay`) and appropriate response length (`len-char-medium`) show positive scores in most user clusters (Figure 3). Similarly, overly long responses (`len-char-xlong`) show negative scores in many clusters.

Correlation Analysis of Humor Preferences between User Clusters To quantitatively assess the differences in humor preferences among user clusters, we calculated the Pearson correlation coefficients between the BTL scores of each user cluster. Figure 4 shows the Pearson correlation matrix. Weak to moderate positive or negative correlations are observed between specific clusters. For example, positive correlations are observed between C0 and C3 (0.41), whereas negative correlations are observed between C0 and C6 (-0.39). This means that users belonging to these clusters have similar or different humor preferences. These correlation results indicate differences in humor preferences among user clusters, further supporting the subjectivity of humor preferences.

6.2.2 RQ2: Humor Preference Differences between LLMs and User Clusters

Figure 3 summarizes the BTL scores of humor preference factors for the user clusters and three LLMs under the `no_persona` setting. To address RQ2, this visualization enables a direct comparison of humor preferences across models and user clusters. Our key findings are as follows:

LLMs and User Clusters Exhibit Differences and Similarities in Humor Preference Factors

Cluster	Preferred factors (Top-3; BTL)	Dispreferred factors (Bottom-3; BTL)
C0	parentheses; dialogue; sentences-many	self_reference; surreal_nonsense; length-ratio-short
C1	self_reference; ending-adjective; personification	prompt-proper-noun; exaggeration-rule; ending-ellipsis
C2	self_reference; mini_story; ending-question	prompt-proper-noun; ending-ellipsis; exaggeration-rule
C3	parentheses; ending-ellipsis; space-high	mini_story; exaggeration; prompt-verb
C4	ending-ellipsis; self_reference; parentheses	slang; exaggeration-rule; meta
C5	slang; exaggeration-rule; prompt-proper-noun	surreal_nonsense; ending-adjective; len-char-xlong
C6	surreal_nonsense; prompt-proper-noun; parody	ending-ellipsis; slang; parentheses

Table 2: Cluster-wise preferred and dispreferred humor factors estimated by the BTL model (top-3 and bottom-3). Full results of the cluster-wise BTL scores are reported in Appendix E.

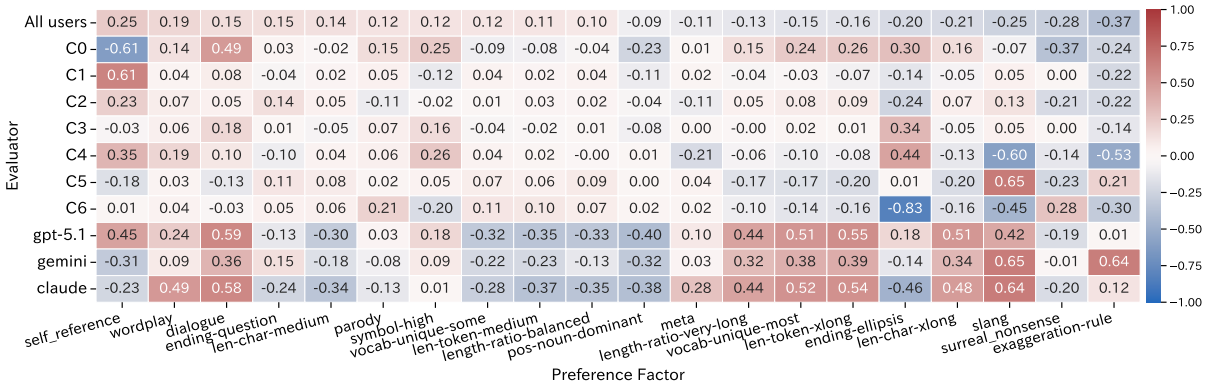


Figure 3: BTL scores of humor preference factors for each user cluster and LLM. C0 to C6 represent each user cluster. "All users" indicates the BTL scores calculated using all users without clustering. Each LLM shows the BTL scores calculated using humor preference data in the `no_persona` setting. Due to space limitations, only the top 10 and bottom 10 factors based on the BTL scores of "All users" are displayed here. The BTL scores for all factors are provided in Appendices E and F.

Based on Figure 3, we compare the BTL scores of humor preference factors between LLMs and user clusters. Overall, both differences and similarities in BTL scores exist between LLMs and user clusters across multiple factors. First, regarding the differences in preference factors between LLMs and user clusters, LLMs exhibited higher BTL scores than user clusters for overly long responses (`len-char-xlong`), responses with high vocabulary diversity (`vocab-unique-most`), and the use of slang (`slang`). Second, regarding the similarities between LLMs and user clusters, both LLMs and users commonly favor factors such as `wordplay` and `dialogue`, whereas commonly disfavored factors include responses with a high proportion of nouns (`pos-noun-dominant`) and surreal/nonsense-style humor (`surreal_nonsense`). Furthermore, focusing on specific clusters, we observed that only certain clusters exhibit humor preferences similar to those of LLMs. For instance, C5 shows high BTL scores for slang, as do LLMs; C0 also exhibits relatively high BTL scores for `len-token-xlong`, similar to LLMs. These results suggest that LLMs share humor preferences with specific user clusters.

LLM Humor Preferences Similar to Specific User Clusters but not to Overall User Preferences To quantify preference similarity between user clusters and LLMs, we computed Pearson correlations between their BTL scores and humor preference factors. Figure 4 shows the Pearson's correlation coefficients. Moderate positive correlations exist in humor preferences between LLMs and specific user clusters. For example, LLMs and cluster C0 exhibit moderately positive correlations (GPT-5.1:0.57; Claude: 0.52). Additionally, we observed weak negative correlations between the BTL scores of all users (estimated using vote data from all users without clustering) and those of each LLM (GPT-5.1: -0.22, Gemini: -0.36, Claude: -0.26). These results not only support prior findings that LLMs possess humor preferences that differ from those of humans (Sakabe et al., 2025) but also provide new findings suggesting that LLMs may share humor preferences with specific user clusters. We believe that these insights contribute to a deeper understanding of the relationship between

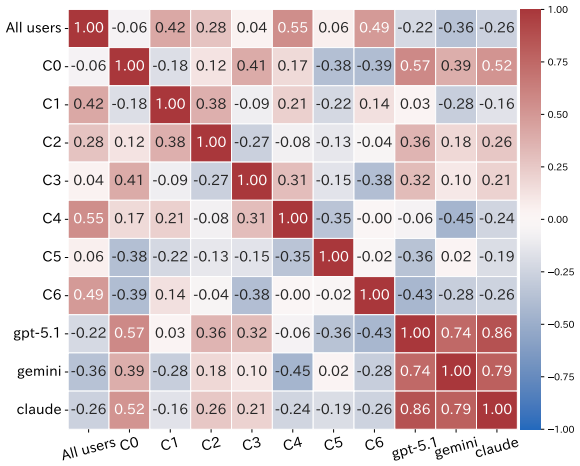


Figure 4: Pearson's correlation matrix of BTL scores between user clusters and LLMs. By comparing user clusters and LLMs, we can identify the LLMs that align with the humor preferences of specific user clusters.

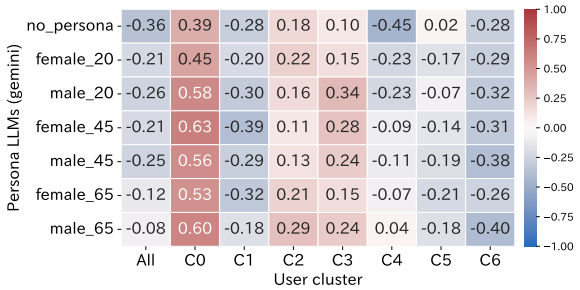


Figure 5: Pearson's correlation matrix between Gemini 3 Pro persona and user clusters, computed over the BTL scores of humor preference factors.

LLM and human humor preferences.

6.2.3 RQ3: Persona Prompting for Aligning LLM Preferences with User Clusters

To answer RQ3, we investigated whether persona prompting influences the alignment of LLMs' preference with specific clusters. The key findings are as follows:

Persona Prompting Helped LLMs Align Humor Preferences with Specific User Clusters We computed Pearson correlations between the BTL scores of humor preference factors for each persona-conditioned LLM and user cluster. Figure 5 shows the resulting correlation matrix. Persona prompting helped LLMs align their humor preferences with specific user clusters. For instance, although weak positive correlations with C0 (0.39) are observed in the `no_persona` setting, the `female_45` persona exhibits a moderate positive correlation with C0 (0.63). In another example, the `male_20` persona exhibits a positive correlation with C3 (0.34), whereas the `no_persona` setting exhibits a very weak positive correlation (0.10). These results suggest that by assigning specific personas to LLMs, their humor preferences can be made to align with those of particular user clusters, paving the way for personalized optimization.[6]

7 Discussion and Conclusion

The humor preferences of user clusters and LLMs were investigated by analyzing the BTL scores of various humor preference factors using Japanese Oogiri. Our analysis revealed that the influence of humor preference factors varies across user clusters (§6.2.1). LLMs possess humor preferences that are different from overall users but may align with specific user clusters (§6.2.2), and persona prompting can help LLMs better align their humor preferences with those of particular user clusters (§6.2.3). These findings lead to discussions below.

Preference Heterogeneity and Personalization Our preference comparison analysis between user clusters and the LLM suggests that, in subjective tasks with heterogeneous preferences, evaluating an LLM solely by its alignment with the overall user population is insufficient. Because an LLM aligns with a specific cluster rather than the overall population, reporting alignment at the cluster (or individual) level clarifies whose preferences the LLM reflects. This perspective also motivates personalized humor evaluation and generation that explicitly accounts for preference heterogeneity. Developing such methods is an important direction for future work.

Effectiveness and Limitations of Persona Prompting We demonstrated that persona prompting can help LLMs align their humor preferences with those of specific user clusters. This suggests that persona prompting may be effective in the context of humor preference, despite prior studies indicating its limited effect (Zheng et al., 2024). However, persona prompting has certain limitations. In our setting, persona prompting mainly changes alignment strength rather than switching the aligned cluster; see Appendix C.1. We leave individual-level alignment beyond persona prompting to future work.

[6]The true demographic attributes of each user cluster are unknown. Our claim is not that persona prompting replicates them but that it can align an LLM's humor preferences with specific user clusters.

Limitations

Limited Coverage of Humor Preference Factors The humor preference factors considered in this study are not exhaustive, and additional factors may influence humor preferences. For instance, non-linguistic factors such as social and cultural background may affect humor preferences (Ruch and Forabosco, 1996; Yue et al., 2016; Cao et al., 2023), but we did not model them. Future work should examine a broader set of humor preference factors.

Extrinsic Validation via Factor-Guided Humor Generation Our study focuses on analyzing preference factors and does not validate them via downstream humor generation in the Oogiri setting. Future work should test whether using the identified high-weight preference factors as guidance in the prompt can steer humor generation toward the preferences of specific user clusters, providing an extrinsic validation of our conclusions.

Focus on Text-based Oogiri Humor We focused on text-based Oogiri humor and did not consider multimodal humor content such as images or audio (Hossain et al., 2022; Hessel et al., 2023). Humor preferences for such content may differ from those for text. Moreover, prior work suggested that even state-of-the-art LLMs have limited multimodal humor understanding, which raises concerns about the reliability of analyzing LLM humor preferences in multimodal settings (Zhou et al., 2025). Future work should extend our analysis to multimodal humor content.

Limited to Japanese Oogiri Humor Our analysis is based on Japanese Oogiri dataset (Murakami et al., 2025). Because we focus on Japanese, our factors include language-specific characteristics such as character-type ratios. Consequently, our findings may not directly transfer to other languages or cultural contexts. However, we note that our user-cluster-based analysis method is language-agnostic and can be applied to other languages and cultures. Broadening the data sources across languages and cultures is an important direction for generalization.

Limited Set of LLMs We analyzed three LLMs (GPT-5.1, Gemini-3-Pro, and Claude-Sonnet-4.5), selected based on reports that they exhibit stronger humor understanding than other open-source models (Murakami et al., 2025). However, many other LLMs may have different humor preferences. Future work should analyze a more diverse set of models, including language-specialized models and models with a wide range of parameter sizes.

Unknown True Attributes of Raters in Source Data Our clustering and humor preference factor analyses rely on user voting data collected from an Oogiri platform. However, the platform does not provide verified information about raters' attributes or backgrounds. Therefore, we cannot trace how each user cluster corresponds to raters' true personas or demographic attributes. Collecting such information and clarifying these correspondences is an important direction for both personalized humor modeling and the interpretation of our analyses. In future work, we plan to analyze how user information such as demographics and personality traits (e.g., the Big Five) relates to cluster membership and humor preferences, once such data become available.

Generalizability to Less Active Users To stabilize user vectors, we restrict the analysis to users with at least 100 votes (§3). This focus on highly active users may limit the generalizability of our clusters and analyses to low-activity and cold-start users. Developing methods that can robustly model users with sparse feedback remains an important direction for future work.

Sensitivity to Missing Votes in Clustering The vote data are inherently sparse because each user votes on only a subset of prompts. As a result, user similarity estimation and clustering rely on partially overlapping observations, and the resulting clusters may be sensitive to how missing values are treated. Moreover, the missingness can be informative (e.g., users may choose not to vote on prompts they dislike), potentially conflating preference similarity with participation patterns. Future work should incorporate missingness-aware modeling and conduct robustness checks under alternative treatments of missing values, such as excluding missing values or imputing missing votes with a neutral value.

LLM vs Human Voting Setup On the Oogiri platform, humans may cast up to three votes per prompt, whereas our LLM task asks for a single funniest response per API call (we call the API three times and treat each as an independent trial). This difference in the task setup may affect the comparability of LLM and human preference estimates;

future work could align the setup (e.g., allowing the LLM to select up to three responses per prompt).

Prompt Dependency of LLM Preference Results LLM humor preference estimates may depend on the prompt template used for the funniest-response selection task (e.g., the no_persona template underlying the main LLM-vs-cluster comparisons). Our persona prompting experiments can be viewed as a sensitivity analysis over different prompts: we observed changes in the magnitude of correlations with user clusters but not a full reversal of preferences, suggesting that the reported LLM preferences are not entirely brittle to prompt wording. Nevertheless, results under a single template should be interpreted with the possibility of prompt dependence in mind.

Reliability and Variance of Humor Strategy Labels We used an LLM to annotate humor strategy labels (§4.2.2) and validated them via a crowdsourced study (1,100 instances, five annotators per instance; 81.4% judged appropriate overall). Nevertheless, quality varies across labels (e.g., `incongruity` has lower accuracy), and the crowdsourced evaluation does not guarantee that every label is equally reliable across the full dataset. Results that depend on these labels should be interpreted with this variance in mind.

Ethical Considerations

Data Source The dataset used in our study is a Japanese Oogiri dataset collected by Murakami et al. (2025). We used the dataset in accordance with its intended use and license (CC BY-NC-SA 4.0). The dataset was collected from a public website in compliance with its terms of use, and the site permits automated crawling as specified in `robots.txt`. To support reproducibility, we will release the code for data collection, processing, and analysis under the CC BY-NC-SA 4.0 license upon publication.

Privacy Although the data are publicly available, they may still contain personal information such as usernames or other identifiers. In our analysis, we took care to protect user privacy. We removed or anonymized identifiers and avoided including content that could enable re-identification.

Potential Risks Although our study is intended for research purposes, analyses of user preferences may be misused for profiling or for drawing unwarranted inferences about individuals or groups. There is also a residual risk of re-identification when working with public web data. To mitigate these risks, we (i) anonymized identifiers and do not release personal identifiers, (ii) focus on aggregate, cluster-level analyses rather than individual-level claims, and (iii) explicitly limit the scope of our conclusions as discussed in the Limitations section.

Bias and Generalizability Because our data are Japanese and culture-dependent, our findings may not generalize to other languages or cultural contexts as we discussed in the Limitations section. The observed preferences may also reflect societal biases present in the source platform.

Use of LLMs for Writing Assistance We used an LLM as a writing assistant for English proofreading and editing. All technical content, analyses, and conclusions were produced and verified by the authors.

References

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Yi Cao, Yubo Hou, Zhiwen Dong, and Li-Jun Ji. 2023. The impact of culture and social distance on humor appreciation, sharing, and production. *Social Psychological and Personality Science*, 14(2):207–217.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

Navoneel Chakrabarty, Srinibas Rana, Siddhartha Chowdhury, and Ronit Maitra. 2019. Rbm based joke recommendation system and joke reader segmentation. In *Pattern Recognition and Machine Intelligence*, pages 229–239, Cham. Springer International Publishing.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model APIs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 792–810, Singapore. Association for Computational Linguistics.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. 2023. DecipherPref: Analyzing influential factors in human preference judgments via GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8344–8357, Singapore. Association for Computational Linguistics.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? an overview of computational humour generation and explanation. *Preprint*, arXiv:2509.21175. Preprint, arXiv:2509.21175.

R Duncan Luce and 1 others. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of plackett-luce models. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 172–180, Cambridge, MA, USA. MIT Press.

P. McDonald. 2013. *The Philosophy of Humour*. Philosophy Insights. HEB Humanities E-Books.

A Peter McGraw and Caleb Warren. 2010. Benign violations: making immoral behavior funny: Making immoral behavior funny. *Psychol. Sci.*, 21(8):1141–1149.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

John Morreall. 2024. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Soichiro Murakami, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2025. Oogiri-master: Benchmarking humor understanding via oogiri. *Preprint*, arXiv:2512.21494.

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Willibald Ruch and Giovannantonio Forabosco. 1996. A cross-cultural study of humor appreciation: Italy and germany.

Ritsu Sakabe, Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi. 2025. Assessing the capabilities of llms in humor:a multi-dimensional analysis of oogiri generation and evaluation. *Preprint*, arXiv:2511.09133.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: a Japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kohtaro Tanaka, Kohei Uehara, Lin Gu, Yusuke Mukuta, and Tatsuya Harada. 2024. Content-specific humorous image captioning using incongruity resolution chain-of-thought. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2348–2367, Mexico City, Mexico. Association for Computational Linguistics.

Robert L Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18(4):267–276.

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. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xiaodong Yue, Feng Jiang, Su Lu, and Neelam Hiranandani. 2016. To be or not to be humorous? cross cultural perspectives on humor. *Frontiers in psychology*, 7:1495.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, Robert Mankoff, and Robert Nowak. 2024. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In *Advances in Neural Information Processing Systems*, volume 37, pages 125264–125286. Curran Associates, Inc.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13246–13257.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T. Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16273–16287, Suzhou, China. Association for Computational Linguistics.

Metric	Mean	Std	Median
ARI	0.420	0.076	0.420
NMI	0.558	0.055	0.555

Table 3: Stability of user clustering over 100 random seeds. Agreement (ARI, NMI) with the baseline partition (random_state=42) is reported.

Method	ARI	NMI
Agglomerative	0.430	0.523
Spectral	0.417	0.581

Table 4: Agreement with the baseline K-means partition (K=7). User embeddings and K=7 were kept the same; proximity was cosine similarity.

K	ARI	NMI
6	0.347	0.534
7	1.000	1.000
8	0.378	0.543

Table 5: Sensitivity to the number of clusters. ARI and NMI are computed against the baseline partition with K=7.

A Clustering Robustness Analyses

To assess whether our user clusters are reproducible rather than artifacts of a single setup, we report three robustness analyses: seed sensitivity, comparison with other clustering methods, and sensitivity to the number of clusters K. We treated the baseline partition as the reference and measured agreement using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

Implementation Details We used scikit-learn's `KMeans` with k-means++ initialization and `n_init=10`, and we used `random_state=42` to define the baseline partition. The number of clusters was set to $K = 7$ based on the elbow method (Thorndike, 1953) and silhouette scores (Rousseeuw, 1987).

Seed Sensitivity We ran K-means with 100 different random seeds (varying `random_state`) and computed ARI and NMI of each resulting partition against the baseline. Table 3 reports the mean, standard deviation, and median. The median ARI (0.420) and NMI (0.555) indicate moderate agreement across initializations.

Comparison with Other Methods We compared the baseline K-means partition with Agglomerative and Spectral clustering (same user embeddings and K=7, cosine similarity). Table 4 shows that both methods achieve moderate agreement (ARI $\approx$ 0.42-0.43, NMI $\approx$ 0.52-0.58) with the baseline, suggesting that the main partition structure is not specific to K-means.

Sensitivity to K We varied K to 6 and 8 and compared the resulting partitions to the baseline (K=7) using ARI and NMI (Table 5). Although some users shift clusters when K changes, NMI remains around 0.53–0.54 for $K \neq 7$, indicating that the overall structure is not extremely sensitive to small changes in K.

B Humor Preference Factors

This section provides the detailed definitions of the humor preference factors used in this study. The humor preference factors are broadly categorized into linguistic features (§B.1) and humor strategy labels (§B.2).

B.1 Linguistic Features

B.1.1 Definition

Table 6 summarizes the detinitions of the linguistics features. The linguistic features consist of six subgroups: basic features, morphological analysis features, special symbol features, sentence-ending pattern features, writing-style features, and prompt-response relational features, totaling 45 features. The first five subgroups are based on the linguistic characteristics of the responses, while the last subgroup is based on the relationship between the prompt and the response. We describe each subgroup in detail below.

Basic Features (11 features) This feature subgroup includes basic linguistic features such as the number of characters in the response and the ratios of different character types. Specifically, it comprises the following features: character count; the hiragana, katakana, kanji, alphabet, digit, punctuation, space, and symbol ratios; punctuation and sentence counts.

Morphological Features (10 features) This feature subgroup includes features derived from the morphological analysis of the responses, such as POS ratios and token counts. It consists of the following features: noun, verb, adjective, adverb, particle, and auxiliary verb ratios; token and unique token counts; lexical diversity (unique words/total words); and the presence of proper nouns.

Feature Name	Definition
Basic Features (11 features)	
len-char-{short, medium, long, xlong}	Number of characters in the response (quartile-based)
hiragana-{minimal, low, medium, high}	Ratio of hiragana characters (quartile-based)
katakana-{minimal, high}	Ratio of katakana characters (median-based, 2 levels)
kanji-{minimal, low, medium, high}	Ratio of kanji characters (quartile-based)
alphabet-{minimal, high}	Ratio of alphabet characters (median-based, 2 levels)
digit-{minimal, high}	Ratio of digit characters (median-based, 2 levels)
punct-{minimal, high}	Ratio of punctuation marks (median-based, 2 levels)
space-{minimal, high}	Ratio of space characters (median-based, 2 levels)
symbol-{minimal, high}	Ratio of symbol characters (median-based, 2 levels)
punct-count-{few, most}	Number of punctuation marks (median-based, 2 levels)
sentences-{one, many}	Number of sentences (median-based, 2 levels)
Morphological Analysis Features (10 features)	
pos-noun-{low, medium, high, dominant}	Ratio of nouns (quartile-based)
pos-verb-{minimal, high}	Ratio of verbs (median-based, 2 levels)
pos-adj-{minimal, high}	Ratio of adjectives (median-based, 2 levels)
pos-adverb-{minimal, high}	Ratio of adverbs (median-based, 2 levels)
pos-particle-{minimal, low, medium, high}	Ratio of particles (quartile-based)
pos-auxiliary-{minimal, high}	Ratio of auxiliary verbs (median-based, 2 levels)
len-token-{short, medium, long, xlong}	Number of tokens (quartile-based)
vocab-unique-{few, some, many, most}	Number of unique tokens (quartile-based)
vocab-diversity-{repetitive, very-diverse}	Lexical diversity: unique tokens / total tokens (median-based, 2 levels)
proper-noun	Contains proper nouns
Special Symbol Features (5 features)	
dialogue	Contains dialogue markers (「」)
parentheses	Contains parentheses
tilde	Contains tilde (~)
number	Contains numbers
slang	Contains internet slang (www)
Sentence-Ending Pattern Features (9 features)	
ending-period	Ends with a period
ending-question	Ends with a question mark
ending-exclamation	Ends with an exclamation mark
ending-ellipsis	Ends with an ellipsis
ending-noun	Ends with a noun
ending-verb	Ends with a verb
ending-adjective	Ends with an adjective
ending-particle	Ends with a particle
ending-auxiliary	Ends with an auxiliary verb
Writing Style Features (4 features)	
polite-style	Polite style (desu/masu form)
casual-style	Casual style (da form)
exaggeration-rule	Contains exaggeration expressions
negation	Contains negation expressions
Prompt-Response Relational Features (6 features)	
prompt-overlap-{minimal, low, medium, high}	Character overlap ratio between prompt and response (quartile-based)
prompt-kanji-share-{minimal, high}	Shared kanji ratio between prompt and response (median-based, 2 levels)
prompt-noun	Contains nouns from the prompt
prompt-verb	Contains verbs from the prompt
prompt-proper-noun	Contains proper nouns from the prompt
length-ratio-{short, balanced, long, very-long}	Ratio of response length to prompt length (quartile-based)

Table 6: Definition of linguistic features.

Special Symbol Features (5 features) This subgroup includes features based on the presence of specific symbols or patterns in a response. Specifically, it consists of indicators of whether the response contains dialogue quotes (「」), parentheses, tilde (∼), numbers, internet slang phrases such as "www" or "草". The internet slang phrases "www" and "草" are commonly used in Japanese online communities, similar to "lol" in English.

Sentence-Ending Pattern Features (9 features) This feature subgroup includes features based on POS, punctuation marks, or symbols that appear at the end of a response. Specifically, it consists of indicators of whether a response ends with a period (.), question mark (?), exclamation mark (!) or ellipsis (. . .) and whether it ends with a noun, verb, adjective, particle, or auxiliary verb.

Writing Style Features (4 features) This subgroup includes features based on the presence of specific writing styles unique to the Japanese language, such as polite, casual, exaggerated, and negated forms.

Prompt-Response Relational Features (6 features) This feature subgroup consists of features based on the relationship between the prompt and response. Specifically, it includes the following features: character overlap ratio between the prompt and response, shared kanji ratio between the prompt and response, response-to-prompt length ratio, and indicators of whether the response contains nouns, verbs, or proper nouns from the prompt.

B.1.2 Binning of Continuous Features

To estimate the relative strength of each factor in the BTL model, all features must be treated as categorical factors. In this study, following the DecipherPref framework (Hu et al., 2023), we apply binning to convert continuous-valued features into categorical factors. Specifically, we use quartile-based binning (4 levels) for continuous-valued features. Features that cannot be reliably categorized based on quartiles (e.g., due to many duplicated values and insufficient value diversity) are split into two levels using the median.

For example, the feature `len-char` (the number of characters in a response) is binned into four quartile-based categories (`len-char-short`, `len-char-medium`, `len-char-long`, `len-char-xlong`), whereas `pos-verb` (the ratio of verbs in a response) is binned into two median-based categories (`pos-verb-minimal`, `pos-verb-high`). Details of the binning scheme for each feature are provided in Table 6.

```
You are an annotator for Ogiri responses. Read the
     given prompt and response, and assign
     predefined response strategy labels accurately
     and consistently. Think carefully so you can
     explain all decision rationale, and select a
     confidence level that represents the
     reliability of your judgment.

## Task Overview
Read the given Ogiri prompt and response pairs, and
     based on the strategy label definitions below,
     annotate each response with the most
     appropriate strategy label. Multiple label
     assignments are permitted. Labels must be
     assigned based on the label definitions.

## Strategy Label Definitions
- Label: {{label_name}}
  Definition: {{definition}}
  Guidelines: {{guidelines}}
  Ambituity Notes: {{ambiguity_notes}}
  Examples: {{examples}}

... (The remaining label definitions are omitted
     for brevity.)

## Confidence Levels
- Available confidence: high, medium, low

## Output Format
{
  "items": [
    {
      "prompt_id": "string",
      "response_id": "string",
      "selected_labels": [
        {
          "reason": "string",
          "label": "string",
          "confidence": "string"
        }
      ],
    }
  ],
}

## Annotation Target
prompt_id: {{prompt_id}}
response_id: {{response_id}}
Prompt: {{prompt}}
Response: {{response}}
...
```

Figure 6: Prompt template for humor strategy labeling.

B.1.3 Implementation

For extracting linguistic features, we used SudachiPy (Takaoka et al., 2018) for morphological analysis, and Python's standard library `unicodedata` and regular expressions for character type identification. SudachiPy provides morpheme segmentation and part-of-speech tags, which we use to compute POS-based features (e.g., POS ratios and ending POS indicators) for each response. We used `unicodedata` and regular expressions to categorize character types and detect punctuation/symbol markers; for relational features, we applied the same procedure to both the prompt and response texts. For all tools, we used the default settings.

B.2 Humor Strategy Labels

This section provides the detailed definitions of the humor strategy labels, the implementation of

Label Name	Definition	Guidelines	Ambiguity Notes
wordplay	Manipulates surface linguistic features (sound, characters, syntax) to create humor through puns, double meanings, or rhythm.	Check for phonetic substitutions, puns, or rhythmic structure. Use this label when surface-level manipulation is primary.	If the humor comes solely from meaning inversion, consider incongruity. Focus on "surface-level linguistic manipulation."
shared_experience	Draws on shared everyday experiences, eliciting laughter through empathy.	Confirm empathy is primary; exaggeration is secondary.	If the humor relies primarily on impact, use exaggeration.
exaggeration	Exaggerates or downplays quantity, emotion, or scale to an extreme.	Determine whether exaggeration is the primary goal.	If empathy is core, use shared_experience.
black_joke_satire	Engages with social norms or taboos, creating humor through irony and taboo language.	Record the source and target categories (i.e., what is being referenced vs. what is being targeted). Distinguish it from exaggeration and incongruity.	If the humor primarily stems from the prompt's absurdity, consider surreal_nonsense. If it relies only on expectation reversal, consider incongruity.
surreal_nonsense	Severs contextual connections, making absurdity itself the source of humor.	Confirm that the logical leap is intentional.	Context destruction is required; eccentricity alone is insufficient.
incongruity	Uses a reversal of an expected development/premise as the punchline.	Confirm that the response sets up an expectation and then reverses it. Check whether the reason for the reversal is explicit.	If the target is the oogiri framework or the prompt itself, use meta. If ethical criticism is primary, consider black_joke_satire.
meta	Refers to contradictions in the oogiri framework, its rules, or the prompt itself, creating humor from an external perspective.	Check whether framework elements (recording, host, format) are used. Use when pointing out flaws in the prompt.	If the humor is purely meaning inversion or expectation violation, consider incongruity. If it includes an overview of the framework, prioritize meta.
self_reference	Uses the responder's own shortcomings as material, creating humor from a first-person perspective.	Check for first-person pronouns and whether the responder's own characteristics/failures are central to the punchline.	If the response primarily criticizes the framework or the prompt, prioritize meta.
personification	Gives human-like emotions or a voice to inanimate objects, creating humor through characterization.	Check whether the target is clearly personified. Judge whether the character's voice is the primary driver of the humor.	If borrowing entire settings/stories, use parody as primary, personification as secondary.
parody	Borrows or transforms settings/stories from external content, creating humor through the gap between the response and the source material.	Explicitly identify the source material. Confirm structural borrowing (not just proper nouns).	Proper nouns alone do not constitute parody. Use structural, setting, or story borrowing as the criterion.
mini_story	Depicts a short story/scene, creating humor in the conclusion.	Check whether a specific situation/scene is described and ends with a punchline. Confirm that the narrative structure is primary.	If story elements are primary, use mini_story as the main label. If the response is a one-liner or primarily wordplay, prioritize other labels.

Table 7: Humor strategy labels: definitions, guidelines, and ambiguity notes

Humor strategy label	Accuracy	Agreement rate
shared_experience	0.880	0.774
black_joke_satire	0.820	0.784
exaggeration	0.700	0.704
incongruity	0.590	0.716
meta	0.830	0.734
mini_story	0.910	0.762
parody	0.830	0.778
personification	0.840	0.778
self_reference	0.790	0.794
surreal_nonsense	0.890	0.804
wordplay	0.870	0.752
Overall	0.814	0.762

Table 8: Crowdsourced quality evaluation of humor strategy labels (1,100 instances, 5 annotators per instance, majority vote). Accuracy: proportion judged valid by majority. Agreement rate: average proportion of annotators supporting the majority decision.

feature extraction, and the quality evaluation of the annotated labels.

B.2.1 Definition and Prompt Template

Figure 6 shows the prompt template for humor strategy labeling, and Table 7 presents the definitions of each label, annotation guidelines, and instructions for handling ambiguous cases. The `{{label_name}}`, `{{definition}}`, `{{guidelines}}`, and `{{ambiguity_notes}}` parts of the prompt template are replaced with the label name, definition, annotation guidelines, and instructions for handling ambiguous cases for each label shown in Table 7, respectively. The `{{examples}}` part is replaced with specific examples of annotations corresponding to each label. Additionally, the `{{prompt_id}}`, `{{response_id}}`, `{{prompt}}`, and `{{response}}` parts of the prompt template are replaced with the IDs and texts of the prompt and response, respectively.

B.2.2 Implementation

For humor strategy labeling, we used GPT-5.1, a state-of-the-art LLM, via an API. For each API call, we set `temperature=1` and used default values for all other parameters. To mitigate variability in API outputs, we followed a self-consistency protocol (Wang et al., 2023): for each labeling prompt, we collected three responses and selected the final label by majority voting. In addition, because the annotation set was large (14,389 instances), we adopted a batch prompting strategy (Cheng et al., 2023) to reduce API costs, processing 20 samples per API call.

B.2.3 Quality Evaluation of Annotation Labels

To evaluate the annotation quality of humor-strategy labels, we conducted a crowdsourced evaluation on 1,100 instances (100 per label) via Yahoo! Crowdsourcing, a major Japanese crowdsourcing platform. Each instance was judged by five independent Japanese-speaking annotators; the final label was determined by majority vote. Annotators were shown the prompt, response, humor strategy label, and label definitions, and answered a binary question: *Is the humor-strategy label appropriate for this prompt-response pair?* We administered a comprehension test on the label definitions; only workers who passed could participate. Following the platform's payment guidelines, we set the reward to 10 yen per 25 tasks. Workers were notified that their annotations would be used for research.

Overall, 81.4% of the 1,100 instances were judged appropriate (average agreement rate 76.2%). Table 8 reports accuracy and agreement rate per label; Table 9 provides annotation examples. Quality varies across labels (e.g., `incongruity` has relatively lower accuracy), which we attribute to definition ambiguity; results should be interpreted with this variance in mind (see Limitations). In future work, improving label definitions and validation with additional annotators will be important.

C Persona Prompting

This section describes the persona prompting template and detailed implementation.

Prompt Template Figure 7 shows the prompt template for persona prompting. Table 10 lists the descriptions of each persona used in the prompt template. For example, the `{{persona_description}}` part of the prompt template is replaced with the description of each persona shown in Table 10.

Implementation We describe the implementation for collecting LLM humor-preference data using persona prompting. We used the three state-of-the-art LLMs, Gemini-3-Pro, GPT-5.1, and Claude-Sonnet-4.5, via their APIs. For each API call, we set temperature to 1.0 and used default values for the other parameters. To account for variability in API outputs, we collected three responses for each persona prompt. In our analysis, we treated each response as an independent vote and used them to

Prompt	Response	Label
Build suspense in one line.	You know why you were called in, right?	shared_experience
Tell me the most pointless trivia in the world.	I am in a bad mood when I wake up.	self_reference
What is a wedding you would hate like?	It is held every day.	exaggeration
Slimy horse racing: what is it like?	Every horse is newborn.	surreal_nonsense
We are recruiting heroes. What are the eligibility requirements?	Someone whose hometown was destroyed.	incongruity

Table 9: Annotation examples of humor-strategy labels. For visibility, Japanese prompts and responses are translated into English.

Persona	Description
no_persona	Please evaluate the Ogiri responses.
female_20	You are a 20-year-old female born in 2005. You are a university student who enjoys comedy variety shows. You are sensitive to cute and emotionally resonant things, and you like relatable content. Please evaluate the Ogiri responses with a young woman's sensibility.
male_20	You are a 20-year-old male born in 2005. You are a university student who frequently uses SNS and watches YouTube. You are well-versed in trending topics, memes, and internet slang. Please evaluate the Ogiri responses with a youthful sensibility.
female_45	You are a 45-year-old female born in 1980. You work as a company employee and have a family (husband and two children). You are knowledgeable about Showa and Heisei era comedy and current affairs. Please evaluate based on the common sense and experience you have cultivated as a working professional.
male_45	You are a 45-year-old male born in 1980. You work as a company employee and have a family (wife and two children). You are knowledgeable about Showa and Heisei era comedy and current affairs. Please evaluate based on the common sense and experience you have cultivated as a working professional.
female_65	You are a 65-year-old female born in 1959. After retirement, you enjoy pursuing hobbies. You enjoy traditional comedy such as rakugo and manzai. Please evaluate the Ogiri responses from a perspective enriched by life experience.
male_65	You are a 65-year-old male born in 1959. After retirement, you enjoy pursuing hobbies. You enjoy traditional comedy such as rakugo and manzai. Please evaluate the Ogiri responses from a perspective enriched by life experience.

Table 10: Persona descriptions used for persona prompting.

estimate BTL scores.

C.1 Effectiveness and Limitations of Persona Prompting

This subsection provides additional observations from our persona prompting experiments (§6.2.3) to clarify how persona prompts affect LLM-to-cluster alignment. We focus on whether persona prompting changes the direction vs. magnitude of correlations and on heterogeneity across models, personas, and clusters.

First, in our experiments the different persona profiles tended to maintain the same direction of correlation with each user cluster (e.g., positive with C0, C2, C3 and negative with others); we did not observe a flip from positive to negative correlation when changing the persona. Thus, persona prompting mainly modulates the magnitude of alignment rather than fully steering the model toward a different cluster, and it is insufficient on its own for full personalization. Second, effects vary by model, persona, and cluster (e.g., stronger for Gemini than for GPT-5.1 or Claude; see Appendix F). Future studies should explore methods beyond persona prompting to align LLM humor preferences with individual users more precisely.

D Self-Consistency of LLM Funniest-Response Selections

To assess how often the LLM selects the same response across the three API calls per prompt, we computed the proportion of prompts where all three calls agree (3of3), where at least two agree (2of3), and the average pairwise agreement (Pairwise) over 897 prompts (no_persona setting; average 16 responses per prompt). Table 11 reports these metrics for each model along with a random baseline (selecting one of 16 responses uniformly). Agreement is substantially higher than the random baseline, indicating that the LLM's choices are not purely random and exhibit moderate self-consistency.

```
{{ persona_description }}

Your role is to look at multiple responses to a
     given Ogiri prompt, and select the response
     you truly find funny based on your own unique
     sensibility. Your task is to choose the
     response you find funny.

## Evaluation Rules
1. Select at most 1 response that you find funny
2. If there are no responses you find funny, you
     must not select any. (0 selections is
     acceptable)
3. For the response you select, provide a brief
     explanation of why you found it funny

## Prompt
{{ prompt }}

## Response_candidates
{{ response_candidates }}

## Output Format
Please output in the following JSON format.
Do not include any text other than JSON.

{
  "selected_responses": [
    {
      "reasoning": "Reason for selecting that
           response (approximately 50 characters)"
      "response_id": "Response ID",
    }
  ]
}

Notes:
- Include only the selected response(s) in
     selected_responses (maximum 1)
- If there are no responses you find funny, set
     selected_responses to an empty array `[]`.
- Use the exact IDs listed in the response
     candidates above for response_id
```

Figure 7: Prompt template for persona prompting.

Model	3of3	2of3	Pairwise
Claude-Sonnet-4.5	0.313	0.499	0.480
Gemini-3-Pro	0.202	0.462	0.356
GPT-5.1	0.167	0.448	0.317
Random baseline	0.004	0.176	0.063

Table 11: Self-consistency of the three API calls per prompt (no_persona): proportion of prompts where all three calls agree (3of3), at least two agree (2of3), and average pairwise agreement (Pairwise). Random baseline: expected value when selecting one of 16 responses uniformly at random.

E Results of User-Cluster Preference Analysis

This section presents the results of the BTL scores of humor preference factors for each user cluster and all users. Figure 8 shows the BTL scores of humor preference factors. In this heatmap, rows represent humor preference factors, columns represent each user cluster and all users, and the color of each cell indicates the strength of the BTL score.

From the heatmap, we can visually confirm that the strengths of humor preference factors differ across user clusters. For example, we observed that `ending-ellipsis` (responses ending with an ellipsis) has high BTL scores in clusters 0, 3, and 4 (0.30, 0.34, and 0.44, respectively), while showing low BTL scores in clusters 2 and 6 (-0.24 and -0.83, respectively), indicating different trends across clusters.

F Results of LLMs' Preference Analysis

In this section, we discuss the trends in LLMs' humor preferences (§F.1), the effects of persona prompting (§F.2), the similarity between LLMs and user clusters in humor preferences (§F.3), and the differences in humor preferences among persona LLMs (§F.4).

F.1 Factors that Influence LLMs' Humor Preferences

Figures 9, 10, and 11 show the BTL scores of humor preference factors for Claude-Sonnet-4.5, Gemini-3-Pro, and GPT-5.1, respectively. In each heatmap, rows represent humor preference factors, columns represent each persona, and the color of each cell indicates the strength of the BTL score.

We observed the following trends regarding humor preference factors favored or disfavored by LLMs: Across all three models, LLMs consistently exhibit high BTL scores for overly long responses (`len-char-xlong`), the use of slang (`slang`), and dialogue punctuation (`dialogue`), while showing low BTL scores for responses of moderate length (`len-char-medium`) and responses with low vocabulary diversity (`vocab-unique-some`). Additionally, we observed that the strength of BTL scores for certain factors, such as self-deprecation (`self-reference`), varied across models.

F.2 Effect of Persona Prompting

From the heatmaps (Figure 9, 10, 11), we can visually confirm that the strengths of humor preference factors change with persona prompting compared to the no-persona setting (`no_persona`). For example, in Claude-Sonnet-4.5 (Figure 9), the BTL score for `exaggeration-rule` (responses containing exaggeration phrases) increases from 0.12 in the `no_persona` setting to 0.73 in the `male_20` persona, indicating a trend of changing humor preference factor strengths with persona prompting.

Additionally, the heatmaps show that while each persona of the models generally exhibits similar strengths of humor preference factors, some factors display different trends across personas. For instance, in Claude-Sonnet-4.5 (Figure 9), `alphabet-high` (responses with a high alphabet ratio) shows high BTL scores in the `female_20` and `male_20` personas (0.27 and 0.52, respectively),

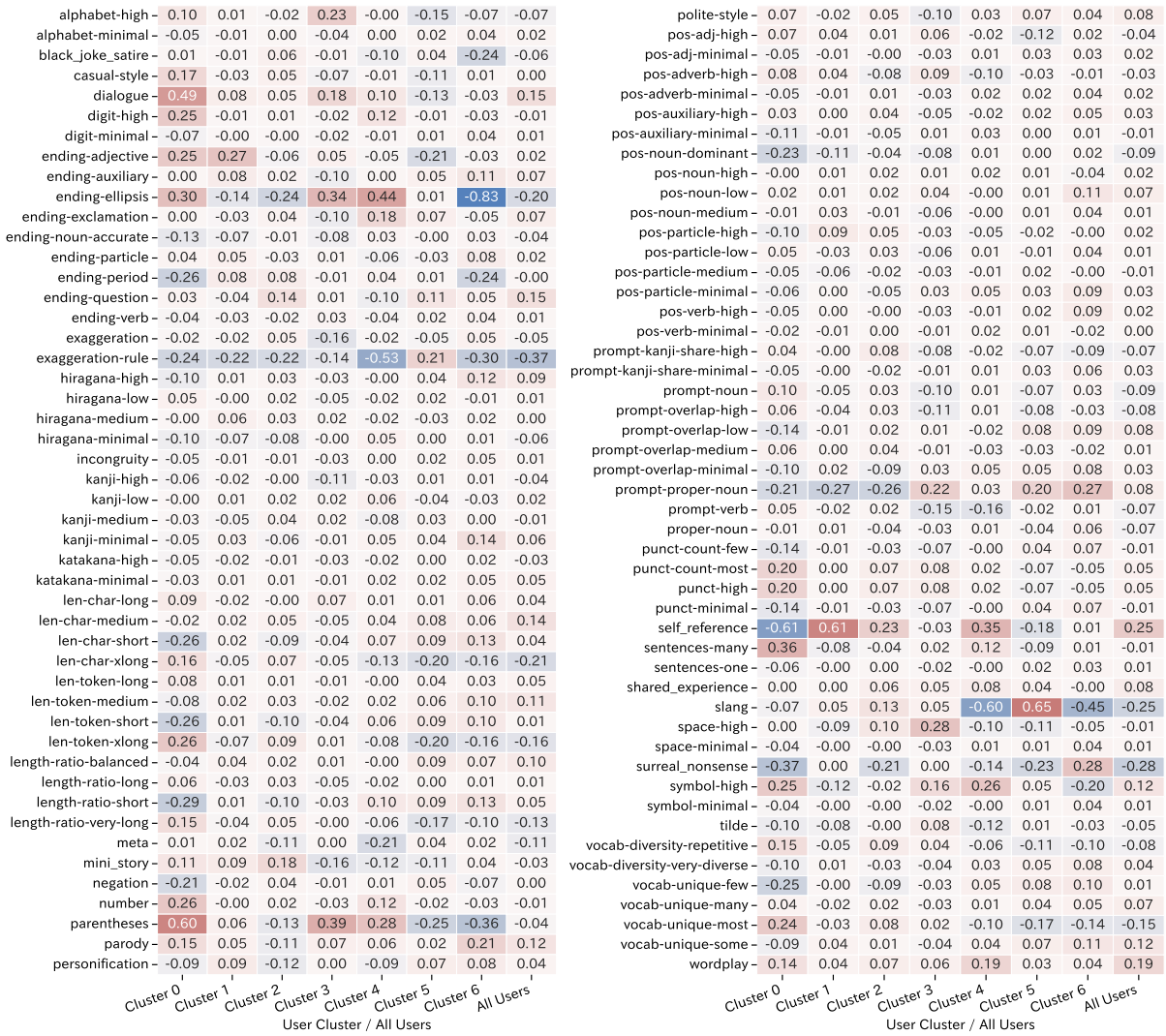


Figure 8: BTL scores of humor preference factors for each user cluster. "All Users" indicates the BTL scores calculated using all users without clustering.

while showing low BTL scores in the `female_65` and `male_65` personas (-0.26 and -0.21, respectively), indicating different trends across personas.

F.3 Humor Preference Alignment between Persona LLM and User Cluster

To address RQ3, we examined whether providing personas can influence an LLM's humor preferences and thereby align the model with specific user clusters. In this analysis, for each persona LLM, we computed the Pearson correlation coefficient between its BTL scores over preference factors and those of each user cluster. This enables a quantitative assessment of the similarity in humor preferences between persona LLMs and user clusters. Figure 12 presents heatmaps of the Pearson correlations for Claude-Sonnet-4.5 and GPT-5.1. Results for Gemini-3-Pro are already reported in the main text (Figure 5). Each row corresponds to a persona LLM, each column corresponds to a user cluster, and each cell shows the Pearson correlation coefficient between the BTL scores of the persona and the user cluster.

The heatmaps indicate that changing the persona alters the correlation with each user cluster, suggesting that the model's humor preference alignment varies across personas. For example, for GPT-5.1, the correlation with C2 increases from 0.36 under the `no_persona` setting to 0.49 under `male_45`. Similar persona-dependent shifts are observed for Claude-Sonnet-4.5 and Gemini-3-Pro, indicating that providing personas can align LLMs with the humor preferences of particular clusters.

However, we also observed cases where the correlation coefficient remains largely unchanged regardless of the persona. For instance, for GPT-5.1,

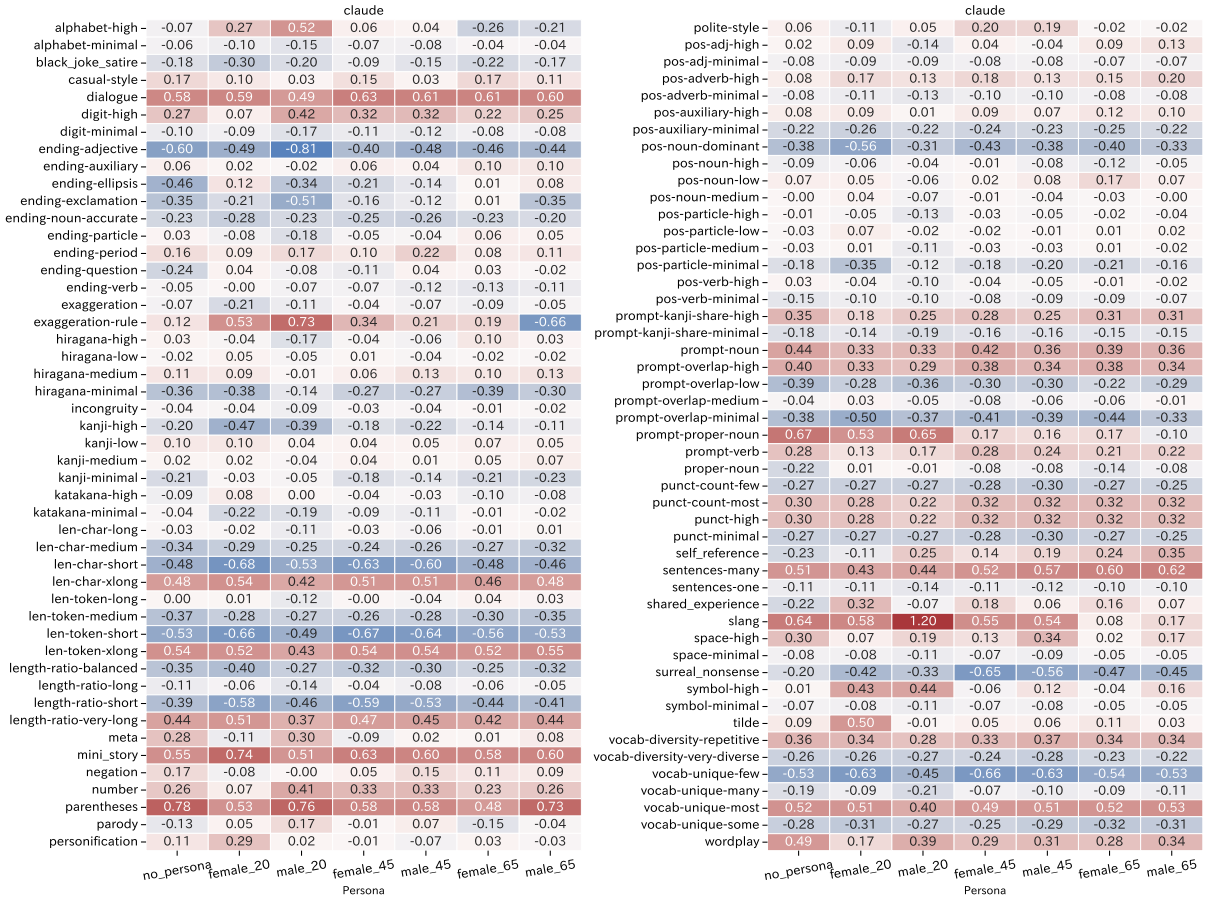


Figure 9: BTL scores of humor preference factors for Claude-Sonnet-4.5.

the correlation with C1 ranges only from -0.09 to 0.03 across personas, showing no notable variation. This suggests that persona prompting may be insufficient to align an LLM's humor preferences for certain user clusters. In the context of personalized humor evaluation and generation aimed at aligning with diverse user preferences, exploring alignment methods beyond persona prompting is an important direction for future work to better optimize for individual preferences.

F.4 Humor Preference Differences between Persona LLMs

Finally, we quantitatively analyzed whether assigning different personas to an LLM induces differences in humor preferences across persona LLMs. Specifically, we computed the Pearson correlation coefficients between the BTL scores of humor-preference factors across persona LLMs. This enables a quantitative assessment of whether humor preferences are similar across persona LLMs. For example, if the correlation coefficient between different personas (e.g., `male_20` and `female_65`) is low, this indicates that the tendencies of humor preference differ across personas; conversely, a high coefficient suggests similar humor preferences. Figure 13 shows the Pearson correlation coefficients of BTL scores across personas for Gemini-3-Pro, GPT-5.1, and Claude-Sonnet-4.5. Each row and column corresponds to a persona, and each cell represents the Pearson correlation coefficient between the BTL scores of the corresponding persona pair.

From this heatmap, we can visually confirm that the correlations between personas tend to be high on average. This suggests that even when different personas are provided, the LLM is likely to exhibit similar humor preferences. One plausible explanation is overlap in pretraining data and similar post-training procedures; however, we did not test this directly. Future work should test this hypothesis by comparing these API-only models with open-weight models and models trained on different corpora, examining whether shared training data contributes to the observed LLM-to-LLM similarity relative to the user clusters in our setting.

Another noteworthy point is that the overall pattern of inter-persona correlations differs substan-

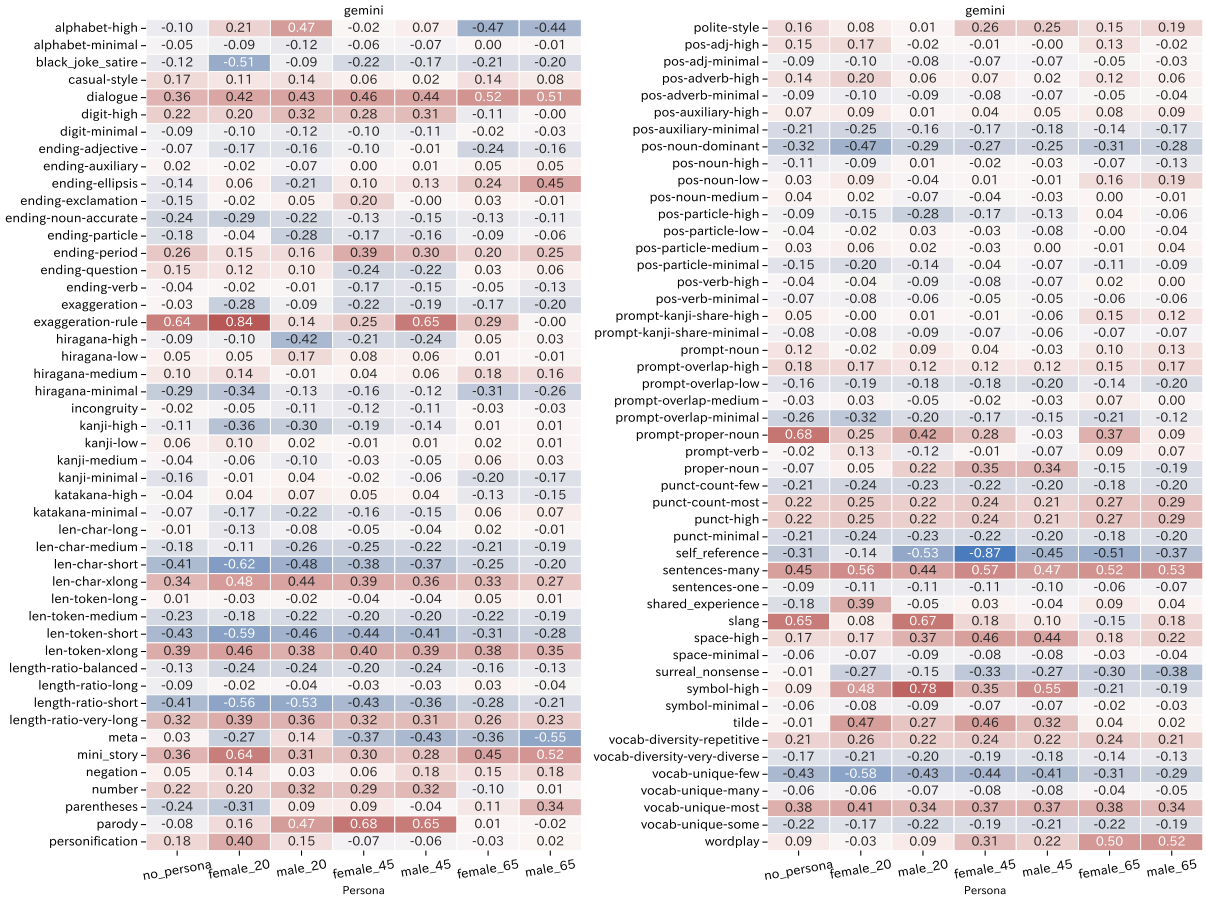
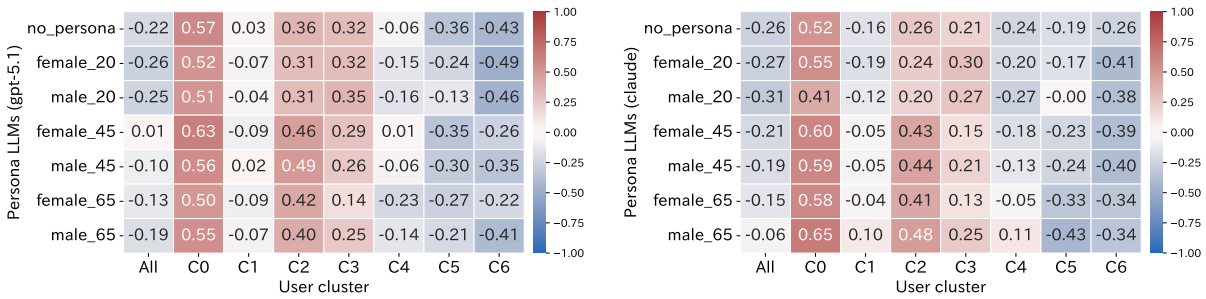


Figure 10: BTL scores of humor preference factors for Gemini-3-Pro.

tially across models. Gemini-3-Pro tends to exhibit larger variations in humor-preference correlations across different personas, whereas GPT-5.1 and Claude-Sonnet-4.5 show consistently high correlations across most persona pairs. For instance, the correlation in humor preference between `female_65` and `male_20` is 0.49 (a moderate positive correlation) for Gemini-3-Pro, but 0.83 and 0.74 (strong positive correlations) for GPT-5.1 and Claude-Sonnet-4.5, respectively. These results suggest that Gemini-3-Pro is more influenced by persona-induced changes in humor preferences, while the other two models are less affected.

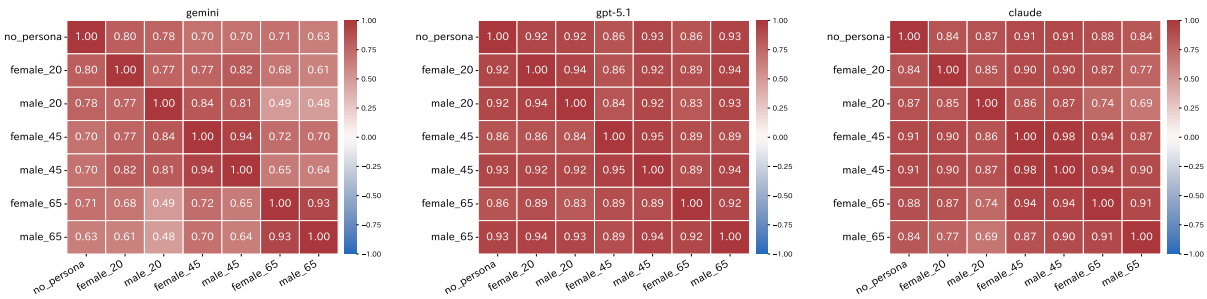
One possible explanation is the impact of reasoning effort. Gemini-3-Pro performs extensive reasoning by default in the API, whereas the other models do not. Introducing reasoning effort may make the LLM more likely to exhibit persona-specific behaviors, which could in turn change humor preferences. In future work, we plan to systematically investigate the effects of varying the scale of reasoning effort and the choice of personas. Such analyses are expected to provide more concrete evidence for the usefulness of persona prompting aimed at alignment to specific user clusters.

Figure 11: BTL scores of humor preference factors for GPT-5.1.



(a) GPT-5.1 (b) Claude Sonnet 4.5

Figure 12: Pearson's correlation matrices between persona LLMs and user clusters, computed over the BTL scores of humor preference factors.



(a) Gemini 3 Pro (b) GPT-5.1 (c) Claude Sonnet 4.5

Figure 13: Pearson's correlation matrices between persona LLMs, computed over the BTL scores of humor preference factors.