An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs

Manuj Malik and **Jing Jiang**
Singapore Management University
{manujm,jingjiang}@smu.edu.sg

Kian Ming A. Chai
DSO National Laboratories
ckianmin@dso.org.sg

Abstract

There are recent efforts to "personalize" large language models (LLMs) by assigning them specific personas. This paper explores the writing styles of such persona-assigned LLMs across different socio-demographic groups based on age, profession, location, and political affiliations, using three widely-used LLMs. Leveraging an existing style embedding model that produces detailed style attributes and latent Dirichlet allocation (LDA) for broad style analysis, we measure style differences using Kullback-Leibler divergence to compare LLM-generated and human-written texts. We find significant style differences among personas. This analysis emphasizes the need to consider socio-demographic factors in language modeling to accurately capture diverse writing styles used for communications. The findings also reveal the strengths and limitations of personalized LLMs, their potential uses, and the importance of addressing biases in their design. The code and data are available at: https://github.com/ra-MANUJ-an/writing-style-persona

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities to perform a wide range of tasks via text generation. Examples include question answering, summarization, logical reasoning, and code generation (BIG-bench authors, 2023). To further unlock the potential of LLMs, recently there has been much interest in "personalizing" LLMs through system prompts that instruct LLMs to behave like a specific character or a given persona (e.g., Shao et al. (2023)). Following Gupta et al. (2024), we refer to these customized LLMs as *persona-assigned LLMs*.

Despite the enthusiasm in personalizing LLMs, currently we have limited understanding of how well these role-playing LLMs perform. Some recent attempts have evaluated their abilities to answer interview questions (Shao et al., 2023), to imitate the speaking styles of the assigned roles and to have role-specific knowledge and memory (Wang et al., 2024), and to pass the Turing Test (Aher et al., 2023; Ng et al., 2024). Recent studies have found intrinsic bias in these persona-assigned LLMs (Aher et al., 2023; Gupta et al., 2024). However, evaluation of persona-assigned LLMs is still largely underexplored.

In this work, we aim to understand the writing styles of persona-assigned LLMs where the personas represent different socio-demographic groups. To the best of our knowledge, style analysis of text generated by persona-assigned LLMs has not been carefully studied. Stylometric analysis of human-written text is a well-studied topic. Previous work has studied subtasks including authorship attribution, authorship verification, and authorship profiling (Neal et al., 2017). However, stylometric analysis of machine-generated text, especially text written by persona-assigned LLMs, is new. We believe that analyzing the stylistic features of text generated by persona-assigned LLMs and comparing them with that of human-written text allows us to examine the behaviours and performance of persona-assigned LLMs from a different perspective that complements existing work on the evaluation of persona-assigned LLMs.

Specifically, we want to understand whether a persona-assigned LLM writes in a style similar to a human with the same persona, and if there are substantial differences, how the style differences can be characterized. We approach these research questions by collating both human-written text and LLM-generated text of a set of socio-demographic personas and comparing their differences in style. For stylometric analysis, we leverage an interpretable style embedding model called LISA (Patel et al., 2023) but propose an LDA-based method to derive eight coarse-grained styles from the original 768 style attributes produced by LISA. With this

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19369–19388
November 12-16, 2024 ©2024 Association for Computational Linguistics

tool, we are able to characterize the writing styles of persona-assigned LLMs and compare them with their counterparts from human-written text. Our extensive experiments reveal that although LLMs' writing styles are not drastically different from those of humans from the same socio-demographic groups, some distinct differences can be observed. Additionally, we observe clear style differences between the three LLMs we study, which suggest that they are suitable for different application scenarios.

In summary, this paper makes the following contributions: (1) We develop a stylometric analysis method based on LISA (Patel et al., 2023) and LDA to facilitate the analysis of writing styles of persona-assigned LLMs. (2) We empirically analyze the writing styles of three popular LLMs when they are assigned different personas, and compare them with those of real Reddit comments. (3) Our experiments offer interesting observations of persona-assigned LLMs' writing styles, which we hope will inspire and guide future development and application of role-playing LLMs.

We come from a more layman understanding of persona based on the organ structure of subreddit, rather than a psychographics analysis. We do not claim to have understood or analyzed personas on sub-reddit in depth. Rather, the paper is motivated by trying to understand how well persona-assigned LLMs generated texts that match texts given by such layman-understanding of persona. In fact, the notion of persona learnt by LLMs could be from similar texts.

2 Data

We collate two kinds of data for our study. First, we need to collect a corpus of text written by humans of different socio-demographic personas. This human-written corpus allows us to derive a "style profile" of each socio-demographic persona we want to study. Next, we want to collect a corpus of text generated by LLMs that have been assigned these socio-demographic personas. By comparing the style profiles of the LLM-generated text with those of the human-written text, we can assess LLMs' abilities to write in a style that matches their assigned socio-demographic persona.

Socio-demographic personas. We consider 35 socio-demographic personas in four commonly studied categories: age, location, profession, and political affiliation. Under each category, we manually curate a set of diversified personas, partially based on what we are able to observe or obtain from the subreddits in Reddit, because we will use Reddit as our main data source. For age, we aim to cover all age groups ranging from young adults to seniors. For location, we aim to cover representative English-speaking cities and countries across different continents. For profession, we select a set of representative professions that have an obvious subreddit community. For political affiliation, we try to cover a wide range of political ideologies in a political spectrum. The complete set of personas can be found in Table 1. Despite our effort to diversify the personas we use, they are not meant to be comprehensive or exhaustive.

Human-written text. To collect human-written text from different socio-demographic personas, we choose to use Reddit comments, largely because it is relatively easy to find subreddits that are representative of the different socio-demographic personas we consider. We manually identify a set of subreddits that are both popularly visited by users and can be mapped to one of our personas. The complete mapping from the subreddits we use to the personas they represent can be found in Appendix A.1. For each subreddit, we randomly select 100 posts, and for each post, we randomly select 10 comments. This process yields a total of 35,000 Reddit comments.

LLM-generated text. Because our focus is on stylometric analysis and comparison, we want our LLM-generated text to be topically similar to the human-written text that we have collected from Reddit. To do so, we give the same set of subreddit posts that we have used (100 for each socio-demographic persona) to an LLM together with a persona instruction that asks the LLM to behave according to that persona. We adopt the persona instructions designed and validated by Gupta et al. (2024). Specifically, Gupta et al. (2024) evaluated ten persona instructions and chose three of them that passed an effective test. While these original persona instructions were designed to answer questions, in our case we want to prompt LLMs to generate Reddit-like comments for us to study their writing styles. Therefore, we modify the persona instructions by Gupta et al. (2024). An example of our persona prompt for an LLM to generate 10 comments in response to a post is shown in Figure 1. The full set is in Appendix B.

Category	Count	Personas
Age	4	a GenZ, a Millennial, a GenX, a Baby Boomer
Location	14	**North America:** New York City, Los Angeles, Canada, Chicago, Texas **Europe:** Paris, Berlin, London, Scotland, Manchester **Oceania:** Australia **Asia:** Singapore, Mumbai, South Korea
Profession	10	a journalist, an architect, an engineer, a finance manager, a photographer, a teacher, a lawyer, a chef, a nurse, a doctor
Poli. Affi.	7	a conservative, a liberal, a libertarian, a progressive, a socialist, an anarchist, a centrist

Table 1: Socio-demographic personas used in our study.

> Take the role of a person from New York City. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform. It is critical that you stay true to the language styles of this role. Here are the details:
>
> **Title:** Millionth Cyclist on Manhattan Bridge
> **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop.
> I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge.
> Whadya think, cool moment I passed up on? Or headache avoided?
>
> Please write comments without any additional details and put them in a form of a list.

Figure 1: A prompt template for data generation using LLMs.

3 Method for Stylometric Analysis

Stylometry is the study of the stylistic features of text. Early work usually uses manually identified features and frequency-based methods such as counting function words (e.g., Rosenthal and McKeown (2011); Bergsma et al. (2012a)). Modern neural methods learn hidden style representations through proxy tasks such as style transfer (e.g., Shen et al. (2017)) and fake news detection (e.g., Schuster et al. (2020)). Although these neural methods deliver stronger results, their style representations are uninterpretable. For our study, we want to characterize the writing styles of LLM-generated text in an interpretable manner. To this end, we adopt a recently proposed interpretable style embedding model called LISA (Patel et al., 2023). LISA produces a 768-dimensional style vector s. Each dimension takes value in $[0, 1]$ and corresponds to a *style attribute* that has a textual description such as "the author uses a simple language", "the author uses a negative tone", and "the author uses offensive language".

Although each style attribute is interpretable, using 768 of these to characterize the style of any text can still be hard to understand. Moreover, we observe that there are many similar or redundant style attributes among the 768 dimensions. Therefore, we use a component analysis method to first identify a few principal coarse-grained styles. Specifically, we use latent Dirichlet allocation (LDA, Blei et al. 2003), which can be interpreted as a multinomial analogue of principal component analysis (Buntine, 2002). We can then project any collection of text onto a lower-dimensional vector $\mathrm{s}' \in [0.0, 1.0]^C$, where C is the number of coarse-grained styles.

We now present the details of our stylometric analysis method. Figure 2 illustrates of our approach.

Identification of coarse-grained styles. We want to identify C coarse-grained styles from the 768 style attributes. We opt to use the LDA method to identify C latent "style topics". First, we take a collection of Reddit comments that are not part of the comments used to represent different personas. Specifically, two social media datasets are used: 'go_emotions'[1] (Demszky et al., 2020) and a subset of the Reddit MUD dataset[2] (Khan et al., 2021; Andrews and Bishop, 2019), totaling 140,727 comments. Details of this dataset can be found in Appendix A.2.

Next, we use LISA to process each comment into a 768-dimensional vector. The value for each style attribute represents the probability of the style being expressed within the comment, and we keep the 20 most prominent styles for each comment. This choice is supported by the data: for most of the 140,727 comments, their 20 most prominent styles all have LISA probabilities of 1.0; only 8,286

[1] https://huggingface.co/datasets/google-research-datasets/go_emotions

[2] https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits

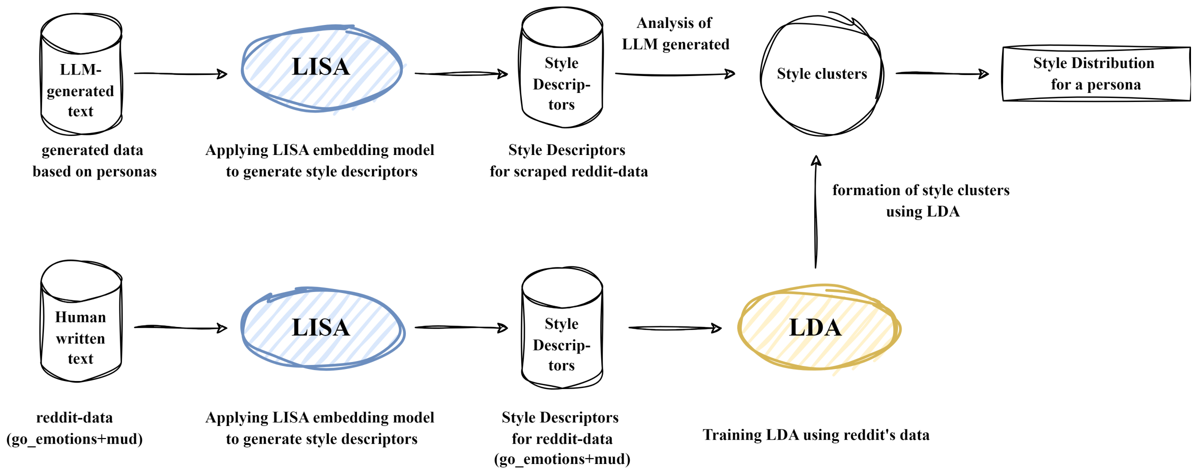

Figure 2: **Methodology for Analysing LLM Generated Text.** First, human-written Reddit data is processed using the LISA embedding model to create style descriptors, which train a Latent Dirichlet Allocation (LDA) model to form style clusters. Concurrently, LLM-generated text, categorized by defined personas, is processed using the LISA embedding model to create its own style descriptors. This text is then analysed with the style clusters to identify the style distribution for a specific persona. This approach combines LLM and LDA to assess the performance of the LLM.

comments have their 20 most prominent styles with LISA probabilities from 0.72 to 0.95; and the median number of styles with a LISA probability 1.0 is 24.

An example comment and the descriptions of its 20 style attributes are provided in Figure 3. We use the 20 style attributes — not their descriptions — for each comment as "words" to create a synthetic document. We eventually obtain a collection of about 140K synthetic documents, each with 20 "words" out of 768 possible "words".

We run LDA on these documents to derive C topics. We have tried six different values of C from 5 to 20, and we find $C = 8$ to be suitable by manual inspection (three of the six topic clustering values (5, 8, 10) are presented in the Appendix C). Each topic is then treated as a coarse-grained style, and we use the top-20 words (which are LISA style attributes) to represent each topic. To obtain coarse-grained styles, we use ChatGPT to generate a meaningful label from the LISA style descriptions of the 20 words in each topic. This gives eight styles: 'Cheerful', 'Simple', 'Judgmental', 'Inquiry', 'Analytical', 'Direct', 'Unenthusiastic', and 'Professional'. This LDA model will be used subsequently to analyse each persona's writing style.

Profiling of texts from human and LLMs. With the LDA model of 8 coarse-grained styles obtained above, we can project each socio-demographic persona's text collection onto a 8-dimensional vector. Recall that for each socio-demographic persona in Table 1, we have collated a set of Reddit comments, and we can also generate a set of comments by a persona-assigned LLM. Given such a set $\mathcal{D}$ that represents a persona, for each comment $d \in \mathcal{D}$, we use LISA to produce a 768-dimensional vector and select the 20 most prominent style attributes. We combine all the style attributes from all the comments in $\mathcal{D}$ to derive a single document. Applying the trained LDA model on this document, we obtain an 8-dimensional vector representing a distribution over the 8 coarse-grained styles. Because most language models aim to emulate the writing style of social media users, the probability distribution is skewed towards a 'Simple' style, making it like a *background* style. To help us better examine the differences across different personas and different LLMs, we therefore remove this *background* style descriptor, and renormalise the probabilities among the remaining 7 coarse-grained styles.

4 Experiment Results and Analysis

LLMs. We choose to experiment with the following three LLMs: `GPT-3.5-Turbo`, `Mixtral-8x7B-Instruct`[3] (Jiang et al., 2024), and `Llama-3-70B-Instruct`[4] (Meta, 2024).

[3]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[4]https://huggingface.co/meta-Llama/Meta-Llama-3-70B-Instruct

Example from the training corpus
`That was the funniest thing so far this season. Sam SCREECHING and stabbin' wights all around in battle fury while more fall on him like throw pillows.`

Associated 20 Style Descriptors, ordered by score
`'The author uses uncommon phrases.', 'The author uses descriptive words.', 'The author uses colorful language.', 'The author uses an energetic style.', 'The author uses a clever play on words.', 'The author is vivacious.', 'The author is using words to create a vivid and engaging atmosphere.', 'The author is using vivid descriptions.', 'The author is using punctuation to create a sense of tension and suspense.', 'The author is using male pronouns.', 'The author is intense in their writing.', 'The author is dramatic.', 'The author is captivating.', 'The author has a distinct and memorable style.', 'The author is creating a sense of anticipation and excitement.', 'The author is using a playful style.', 'The author is describing a current event.', 'The author uses victorious language.', 'The author is using a lighthearted tone.', 'The author uses singular subjects.'`

Figure 3: An example comment in the training corpus with its 20 style descriptors ordered by score.

These are chosen based on our budget, and compute and memory constraints. These LLMs are also at the forefront of both open-source and closed-model applications. The last two are also open-source models that have been instruction-tuned to improve conversational ability and task completion, enabling more natural and coherent dialogue.

Measuring distributional discrepancy. Since the style profiles are expressed in topic distributions, we measure their similarities with the Kullback–Leibler (KL) divergence.

We calculate pairwise KL-divergences between Reddit's distribution and other distributions, then take the average depending on the prompt type. For baseline prompts without any persona, $n = 1$, while for baseline prompts with human persona or persona prompts, $n = 3$. Thus, the result represents the average pairwise KL-divergence.

The average KL divergence between a probability distribution P (Reddit's distribution) and a list of distributions $Q_1, Q_2, \ldots, Q_n$ is given by:

$$D_{\mathrm{KL}}^{\mathrm{avg}}(P \parallel Q_1, Q_2, \ldots, Q_n) = \frac{1}{n}\sum_{j=1}^{n} D_{\mathrm{KL}}(P \parallel Q_j).$$

Each individual KL divergence $D_{\mathrm{KL}}(P \parallel Q_j)$ is calculated as:

$$D_{\mathrm{KL}}(P \parallel Q_j) = \sum_i P(i) \log \frac{P(i)}{Q_j(i)}.$$

To avoid issues with zero values, a small epsilon is added to P and Q_j to prevent division by zero or undefined logarithms. For a baseline prompt without any persona (e.g., Table 6, Prompt 1), $n = 1$; for human or persona baseline prompts, $n = 3$. This method allows for accurate assessment of divergence across different style distributions.

The larger the KL-divergence, the less similar the two distributions are. In our experiments, each distribution is over the following 7 styles: 'Cheerful', 'Judgmental', 'Inquiry', 'Analytical', 'Direct', 'Unenthusiastic', and 'Professional'.

4.1 Persona-Specific Writing Styles

First, we want to check whether different personas in Reddit indeed exhibit different writing styles. Our observation is that there are clear differences of writing styles across different personas in the same category. We use three example personas within each socio-demographic category to illustrate the differences. As we can see in Figure 4, different personas show drastically different styles except for the political affiliation category. For example, in the profession category, engineers' writing styles lean towards 'Inquiry' and 'Analytical,' whereas chefs are more 'Judgmental' and 'Cheerful.' In the age category, we can see that GenZs are more 'Direct' and 'Cheerful,' whereas the Millennials are more 'Judgmental' and 'Analytical.' These radar charts highlight that writing styles on Reddit are varied and non-homogeneous.

4.2 Comparison Across Persona Categories

In comparing the styles of text generated by LLMs with Reddit comments, we can examine raw probabilities and KL-divergence across various personas categorized by age, political affiliation, profession, and location. This analysis shows us how well the LLMs can recreate the communication styles of different groups of real people discussing topics online.

Location-Based Personas. Our study shows that the 'Judgmental' and 'Cheerful' styles are common across all locations. Importantly, GPT consistently gives the highest KL-divergence scores, suggesting that GPT may have its own biased writing styles

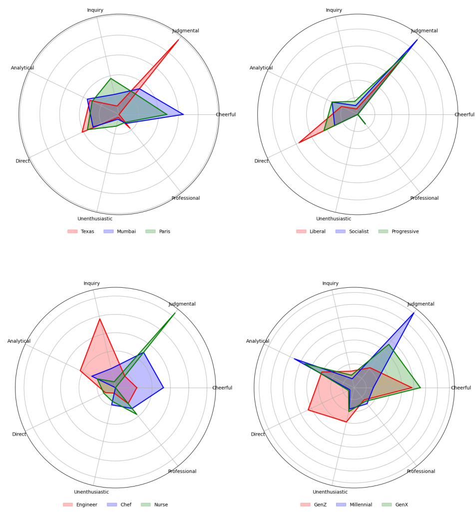

Figure 4: Different writing styles based on real Reddit comments among different socio-demographic groups. In clockwise manner showing writing styles based on locations, political affiliations, professions, and age groups.

compared to the other LLMs. For example, based on Reddit comments, people from Paris display high levels of cheerfulness (0.2415) and inquiry styles (0.1864). While Llama and Mistral show similar distributions with some variations, GPT demonstrates significantly higher cheerful styles (0.6757). KL-divergence values for location-based personas reflect how closely LLMs align with Reddit comments, with Llama and Mistral generally exhibiting lower and moderate divergence, respectively. For the Parisian persona, GPT's high divergence (e.g., 8.7819) indicates a notable departure from Reddit's style distribution, whereas Llama's low divergence values (e.g., 0.2949 for the Parisian persona) suggest a closer match to Reddit's style. Table 11 in Appendix D provides all the KL-divergences.

Profession-Based Personas. Reddit comments for different professions show various style patterns. For example, comments by Finance Managers on Reddit often have high judgmental (0.2946), inquiry (0.1776), and analytical (0.2848) styles. LLMs reflect these style distributions but with some differences. GPT has a higher analytical style (0.4335) for Finance Managers. Mistral, on the other hand, shows a different pattern with higher professional styles (0.5162) but lower judgmental (0.0089) and inquiry (0.0581) styles, differing from Reddit's balance. In terms of KL-divergence scores, Llama has a moderate divergence for Finance Manager (0.3182), meaning it closely aligns with Reddit's style patterns. However, Mistral's high divergence (3.6102) indicates significant differences. This trend is seen across other professions, with Llama (0.3182) and Mistral usually having moderate (3.8891) divergence. GPT shows higher divergences, suggesting distinct style differences from Reddit. Further results are tabulated in Table 12.

Political Affiliations Personas. Political affiliation personas on Reddit also display different style patterns as shown in Table 2. For example, Conservative comments are often highly judgmental (0.4508) and direct (0.2398). Llama reflects this with similar judgmental styles (0.4576), while GPT adds a more professional style (0.4119) compared to Reddit's 0.0742. Mistral diverges significantly, with a high professional style (0.4819) and minimal other styles, indicating a different communication style. For Conservatives, Llama's KL-divergence is moderate (1.1223), suggesting some alignment with Reddit. However, Mistral's high KL-divergence (6.3147) indicates substantial differences. This pattern is consistent across affiliations: Llama generally has lower divergence (e.g., 0.7039 for Liberals), while Mistral consistently shows higher divergence (e.g., 8.9812 for Liberals), indicating it produces text styles that significantly differ from Reddit.

Age-Based Personas. We observe big differences in how LLMs use styles when assigned with personas of people of different age groups. For example, GenZ comments on Reddit exhibit varied styles such as cheerful (0.2418), judgmental (0.1072), analytical (0.1516), and direct (0.2154). Llama and GPT reflect this diversity but with some differences in specific styles. For instance, Llama has a higher direct style (0.2980) compared to Reddit's 0.2154. Mistral, however, diverges significantly with a very high analytical style (0.6279) for GenZ, indicating a distinct style. For GenZ, Llama has a relatively low KL-divergence (0.0869), suggesting it closely matches Reddit's style distribution. In contrast, Mistral's high KL-divergence (5.5082) indicates a significant departure from Reddit's style. This pattern is consistent across other age groups, where GPT shows moderate divergence, but Mistral often presents higher divergence, especially for Baby Boomers (0.9775) and GenX (0.5707). Table 13 in Appendix D gives comprehensive data.

Poli. Affi.	Model	Writing Styles							KL
		Cheerful	Judgmental	Inquiry	Analytical	Direct	Unenthusiastic	Professional	
Conservative	Reddit	0.0000	0.4508	0.0532	0.1820	0.2398	0.0000	0.0742	-
	Llama	0.0000	0.4576	0.0000	0.0639	0.3413	0.0000	0.1372	1.1223
	Mistral	0.0000	0.4626	0.0000	0.0554	0.0000	0.0000	0.4819	6.3147
	GPT	0.0000	0.3363	0.0000	0.2280	0.0238	0.0000	0.4119	1.5868
Liberal	Reddit	0.0000	0.4113	0.0329	0.1068	0.3823	0.0000	0.0667	-
	Llama	0.0000	0.4851	0.0000	0.0320	0.3339	0.0000	0.1489	0.7039
	Mistral	0.0000	0.4955	0.0000	0.0395	0.0000	0.0000	0.4650	8.9812
	GPT	0.0000	0.3886	0.0000	0.1449	0.0748	0.0000	0.3917	1.1417
Libertarian	Reddit	0.0000	0.5225	0.0493	0.1222	0.1936	0.0000	0.1124	-
	Llama	0.0000	0.5097	0.0000	0.0374	0.0892	0.0000	0.3637	1.1626
	Mistral	0.0000	0.3306	0.0000	0.0556	0.0000	0.0000	0.6138	5.2715
	GPT	0.0000	0.2587	0.0000	0.1363	0.0352	0.0000	0.5697	1.4882
Progressive	Reddit	0.0000	0.4627	0.0781	0.1674	0.2202	0.0000	0.0717	-
	Llama	0.0000	0.5998	0.0000	0.0427	0.1455	0.0000	0.2120	1.7204
	Mistral	0.0000	0.4700	0.0000	0.0600	0.0000	0.0000	0.4700	6.3640
	GPT	0.0000	0.3516	0.0000	0.1265	0.0236	0.0000	0.4982	2.1245
Socialist	Reddit	0.0000	0.5591	0.0532	0.1665	0.1502	0.0000	0.0710	-
	Llama	0.0000	0.5603	0.0000	0.1184	0.0000	0.0000	0.3213	4.1913
	Mistral	0.0000	0.3603	0.0000	0.1986	0.0000	0.0000	0.4412	4.3297
	GPT	0.0000	0.3404	0.0000	0.2453	0.0000	0.0000	0.4143	4.3306
Anarchist	Reddit	0.0346	0.5725	0.0328	0.1052	0.1512	0.0000	0.1038	-
	Llama	0.0000	0.5325	0.0000	0.0550	0.1156	0.0000	0.2969	1.3642
	Mistral	0.0000	0.5244	0.0000	0.0678	0.0355	0.0000	0.3723	1.5061
	GPT	0.0000	0.4548	0.0000	0.1071	0.0385	0.0000	0.3995	1.5197
Centrist	Reddit	0.0000	0.5260	0.0000	0.1498	0.2727	0.0000	0.0516	-
	Llama	0.0000	0.6249	0.0000	0.1426	0.0587	0.0000	0.1737	0.2728
	Mistral	0.0000	0.4447	0.0000	0.3001	0.0000	0.0000	0.2553	5.8257
	GPT	0.0000	0.3031	0.0000	0.4594	0.0000	0.0000	0.2375	5.9673

Table 2: Comparison based on political affiliation using KL-Divergence between LLMs and Reddit's Distribution

4.3 Traits of Different LLMs' Writing Styles

Based on our observations with the three LLMs (i.e., Llama, Mistral, and GPT) as discussed above, we find that different LLMs have their own special traits that make them suitable for different situations and audiences. We examine these traits closely to understand how they can be used and how well they can copy the style of discussions on sites like Reddit.

Llama often has a style that is very similar to the informal, conversational style used on Reddit. This suggests Llama may work well for replicating the casual, discussion-based style typical on Reddit while having discussions.

Mistral consistently uses a style that is quite different across various personas. Its style is very professional and formal, contrasting with the more casual Reddit style. This distinct professional pattern might make Mistral suitable for formal communications or discussions requiring a proper style.

GPT demonstrates a balanced mix of styles, especially analytical and professional, across different personas. Its style deviates somewhat from Reddit but not as extremely as Mistral. This balance makes GPT versatile, potentially appealing to audiences that value both critical analysis and professional discourse.

4.4 Comparison with Baseline Personas

Following Gupta et al. (2024), we also use two baseline prompts to ask LLMs to generate comments that do not represent any persona. The first baseline prompt simply asks an LLM to write comments, without mentioning any persona in the prompt. The second baseline prompt uses the phrase "an average human" to replace a persona such as "a lawyer" or "a GenZ". We then compute the KL-divergence between the style distribution of each persona and those of these baseline personas. The analysis of KL divergence values across various categories — location, profession, political affiliation (Table 3), and age — reveals significant stylistic differences. The complete data is in Tables 14 to 16 within Appendix D.

For location, Texas and Canada stand out with high divergence values, particularly in the Llama and Mistral models, indicating distinct regional language styles. Among professions, chefs exhibits the most substantial divergence, especially in the Llama and Mistral models, highlighting a unique professional language. Politically, socialists and liberals show significant deviations, with socialists having the highest divergence values in the Llama model, and liberals notably divergent in the Mistral model; this suggests marked differences in political discourse. Age-wise, GenZ demonstrates the highest divergence, particularly in the GPT model, indicating a distinct generational language style.

Common narratives emerging about the models and their baseline comparisons indicate that LLMs tend to show higher KL divergence values compared to Baseline 1 (N) in several cases. This suggests that the language style of LLMs differs more from Baseline 1 (N) across most of the personas. Overall, the groups with the most divergent language usage were chefs, socialists, and GenZ. This variety highlights the importance of accounting for regional, professional, political, and generational influences when modeling human language to accurately capture how different groups communicate.

5 Related Work

Evaluation of Role-Playing in LLMs Role-playing in large language models (LLMs) is becoming an exciting research area. Giving specific roles to these models can greatly impact how well they perform and make decisions. Although there is interest in personalizing LLMs, we still don't fully understand how well these role-playing models work. Some recent studies have tested their ability to answer interview questions (Shao et al., 2023), mimic speaking styles, hold role-specific knowledge and memory (Wang et al., 2024), and even pass the Turing Test (Aher et al., 2023; Ng et al., 2024). Other research has found biases in these persona-assigned LLMs (Gupta et al., 2024). However, evaluating these role-playing LLMs is still not well-studied with few more works (Aher et al., 2023; Santurkar et al., 2023). Recent research (Zheng et al., 2023) has shown that assigning roles affects response accuracy due to factors like prompt similarity, uncertainty, and word frequency in training data. It was found that gender-neutral roles often lead to better response accuracy than gender-specific roles. Further research has focused on enhancing the decision-making and reasoning abilities of LLMs through role-playing (Shen et al., 2024). Additionally, role-play prompting methodologies have enhanced zero-shot reasoning abilities (Kong et al., 2023) by functioning as effective implicit Chain-of-Thought prompts (Wei et al., 2022), which is somewhat similar to what we're doing.

Stylometry and Its Applications Stylometry involves analyzing writing styles to determine authorship. Previous work has studied subtasks including authorship attribution, authorship verification, and authorship profiling (Neal et al., 2017). Foundational work by (Bergsma et al., 2012b) demonstrated the feasibility of this approach through quantitative analysis. (Hitschler et al., 2017) research advanced this field by utilizing convolutional neural networks, enhancing accuracy in authorship attribution.

Age and Gender Prediction from Writing Styles on Social Media Predicting people's age and gender from text has been a topic of interest in computational linguistics for many years. The rise of social media has provided a lot of data for analysis. The earliest work in this area, by Argamon et al. (2007), used traditional machine learning with style and content features. Later, researchers like (Burger et al., 2011) and Rosenthal and McKeown (2011) used data from social media platforms like Twitter, applying n-gram models and word-based features to predict users' demographics with good accuracy. Recently, attention-based models like the Transformer by Vaswani et al. (2017) have been used for predicting age and gender. Studies have also looked at writing style and demographics. (Rangel and Rosso, 2013) found that certain writing features predict age and gender well. Schwartz et al. (2013) linked language use with psychological traits related to age and gender. Dataset diversity and fairness are also important. (Waseem, 2016) pointed out biases in social media data that could reinforce stereotypes, leading to more research on ethical issues in demographic prediction (Blodgett et al., 2020). Our work is different in that we focus on analysing LLM-generated text whereas those prior works look at human-generated social media content.

Political Affiliation	KL Llama N	KL Llama B	KL Mistral N	KL Mistral B	KL GPT N	KL GPT B
Conservative	1.4579	0.9762	2.0262	2.5475	1.2259	0.7535
Liberal	0.1008	0.3035	3.6252	6.8047	0.2655	0.5641
Libertarian	1.0079	1.3814	1.6153	1.5251	0.1366	0.1054
Progressive	0.1354	0.8413	0.1728	1.4914	0.2151	0.3681
Socialist	3.3548	7.0779	1.5437	1.0820	1.1541	1.6943
Anarchist	1.7319	1.1697	0.1473	0.0508	0.1020	0.1485
Centrist	0.2129	0.8653	0.0216	0.0873	0.0611	0.0372

Table 3: KL Divergence Values by Political Affiliation, where **N** denotes values calculated between 'Persona' and 'No Persona' distribution & where **B** denotes values calculated between 'Persona' and 'Baseline' distribution

6 Conclusions

We have looked at the writing styles of LLMs when they are given different socio-demographic personas. We used three different LLMs and examined their writing styles to see if they follow the writing styles of the given personas. We found that LLMs given different personas wrote in different styles. For example, texts made for chefs, socialists, and Gen Z had very distinct styles. We also saw that where a person is from and his political views can influence his writing style. These findings show that it is important to think about different socio-demographic factors when personalizing LLMs to make them more accurate and relatable. For the three LLMs that we have tried — Llama, Mistral and GPT — we broadly characterized the styles of their texts when given different personas. This can help us use LLMs more effectively in different situations.

Limitations

The scope of this work is limited by the following challenges:

Data Source Bias. Our study has several limitations that should be considered when interpreting the results. First, the use of Reddit comments as a data source may introduce bias, as these comments may not be fully representative of the entire population of any given demographic group. Reddit users tend to represent a specific subset of internet users, often younger, more tech-savvy, and predominantly English-speaking. Consequently, the writing styles we analyzed might not capture the full linguistic diversity and nuances present within broader demographic groups. This limitation suggests that our findings might not be entirely generalizable to all individuals within those groups.

Bias Identification Limitation. Second, although our methodology can identify biases in large language models (LLMs), such as persona-assigned LLMs producing text that aligns with stereotypes, we did not deeply investigate these biases in this paper. For example, if an LLM generates text for a persona that reflects stereotypical traits, it could reinforce harmful stereotypes and perpetuate bias in AI systems. While our stylometric analysis offers valuable understanding into the writing styles of persona-assigned LLMs, we have not examined the ethical implications or potential harms of these biases in detail. This omission is another limitation of our work.

Ethics Statement

We, the authors, affirm that our work adheres to the highest ethical standards in research and publication. We acknowledge that using persona-assigned LLMs on social media could raise ethical concerns. One risk is that these models might be used to create bots that imitate real social media users, which could lead to problems like spreading false information or deceiving people. To avoid such harm, it is important to deploy persona-assigned LLMs responsibly. Our work is a step towards understanding the impact of such persona-assigned LLMs. We encourage discussions about safely adopting and deploying such technologies. We provide detailed information to facilitate the reproducibility of our results, including sharing our code, datasets and other relevant resources to enable the research community to validate and build upon our work. The claims in our paper match our experimental results. However, with large language models, some variability is expected, which we minimize by using a fixed temperature. We thoroughly describe the annotations, dataset splits, models used, and prompting methods tried to ensure the reproducibility of our work.

References

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

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. *arXiv preprint arXiv:1910.04979*.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*.

Shane Bergsma, Matt Post, and David Yarowsky. 2012a. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Shane Bergsma, Matt Post, and David Yarowsky. 2012b. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada. Association for Computational Linguistics.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Wray Buntine. 2002. Variational extensions to em and multinomial pca. In *Proceedings of the 13th European Conference on Machine Learning*.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1301–1309.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Julian Hitschler, Esther Van Den Berg, and Ines Rehbein. 2017. Authorship attribution with convolutional neural networks and pos-eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. *arXiv preprint arXiv:2105.07263*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6).

Man Tik Ng, Hui Tung Tse, Jen tse Huang, Jingjing Li, Wenxuan Wang, and Michael R. Lyu. 2024. How well can LLMs echo us? evaluating AI chatbots' role-play ability with ECHO. *Preprint*, arXiv:2404.13957.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Francisco Rangel and Paolo Rosso. 2013. Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, 177:56–66.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2).

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

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

Chenglei Shen, Guofu Xie, Xiao Zhang, and Jun Xu. 2024. On the decision-making abilities in role-playing using large language models. *arXiv preprint arXiv:2402.18807*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zekun Wang, zhongyuan peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Mingqian Zheng, Jiaxin Pei, and David Jurgens. 2023. Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*.

A Subreddits chosen for our study

A.1 Mappings from Reddit Subreddits to Socio-demographic Personas

Table 4 shows the list of subreddits we have used and how they are mapped to the socio-demographic personas.

A.2 Subreddits chosen from MUD for training LDA

For 'go_emotions', the entire dataset containing 43,227 comments is included. For MUD, which has one million comments from 40 different subreddits with an equal proportion of comments from each subreddit, we select around 10 percent of the comments. We discard one subreddit, 'Pikabu,' as it contained comments in Russian. For the remaining 39 subreddits — Table 5 gives the list — we randomly select 2,500 comments from each, giving 97,500 comments from the MUD dataset altogether. From the two datasets, we obtain about 140K comments in total.

Table 5 shows the extensive list of the subreddits which have been used for training our model.

B Persona Prompts

Table 6, 7 shows the full set of persona prompts we have used to ask the LLMs to generate Reddit-like comments.

C Human Evaluation

By comparing the topics generated through clustering with 5, 8, 10 clusters (Tables 8, 9, and 10) and 12, 16, and 20 clusters (not shown in the tables), we conclude that 8 clusters provide the most concise clustering solution, including all topic descriptions without redundancy.

D Tabular Results

Category	Subreddits
Age	r/teenagers, r/Millennials, r/GenX, r/BabyBoomers
Location	**North America:** r/AskNYC, r/LosAngeles, r/Canada, r/Chicago, r/Texas **Europe:** r/Paris, r/Berlin, r/London, r/Scotland, r/manchester **Oceania:** r/Australia **Asia:** r/askSingapore, r/Mumbai, r/korea
Occupation	r/journalism, r/architecture, r/engineering, r/finance, r/photography, r/Teachers, r/law, r/chefit, r/nursing, r/medicine
Poli. Affi.	r/Conservative, r/Liberal, r/Libertarian, r/Progressive, r/Socialism, r/Anarchism, r/centrist

Table 4: Subreddits used in our study.

Subreddits
'AmItheAsshole', 'Animemes', 'AskReddit', 'ChapoTrapHouse', 'FortNiteBR', 'Market76', 'MortalKombat', 'RoastMe', 'Showerthoughts', 'SquaredCircle', 'The_Donald', 'apexlegends', 'asoiaf', 'aww', 'dankmemes', 'freefolk', 'funny', 'gameofthrones', 'gaming', 'gonewild', 'hockey', 'leagueoflegends', 'marvelstudios', 'memes', 'movies', 'nba', 'news', 'nfl', 'pics', 'politics', 'relationship_advice', 'soccer', 'teenagers', 'todayilearned', 'trashy', 'unpopularopinion', 'videos', 'wallstreetbets', 'worldnews'

Table 5: Subreddits Selected for Training LDA from MUD Dataset

Respond to the following post on social media. I have a title and text body. Write 10 comments that are relevant to the topic. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

You are an average human. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform. Your responses should closely mirror the language style of this persona when writing social media comments. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

Adopt the identity of an average human. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform while staying in strict accordance with the language style of this identity. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

Take the role of an average human. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform. It is critical that you stay true to the language styles of this role. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

Table 6: Baseline prompt examples for data generation using LLMs

You are a person from New York City. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform. Your responses should closely mirror the language style of this persona when writing social media comments. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

Adopt the identity of a person from New York City. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform while staying in strict accordance with the language style of this identity. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

Take the role of a person from New York City. I have a title and text body. Write 10 comments that are relevant to the topic in response to the following post on a social media platform. It is critical that you stay true to the language styles of this role. Here are the details: **Title:** Millionth Cyclist on Manhattan Bridge **Text Body:** I biked into the city on Manhattan Bridge today, and as I approached the plaza with the bike counter, a group of 5 people kept screaming for me to stop. I slowed down, and they said I was the millionth Cyclist and asked for a picture. I only looked closely at 2 of them: one looked homeless and the other didn't. So I rode right past them and didn't indulge. Whadya think, cool moment I passed up on? Or headache avoided? Please write comments without any additional details and put them in a form of a list.

Table 7: Persona prompt examples for data generation using LLMs

CLUSTERING 0: 5 Topics
Topic 0: The author speaks without filler words. The author makes a statement without any politeness. The author lacks authority. The author uses no neutral tone. The author uses language that suggests a lack of certainty. The author lacks strong evidence or logical reasoning. The author uses a sparse writing style. The author leaves no room for misinterpretation or misunderstanding. The author leaves sentences unfinished. The author lacks qualifiers or hedging language. The author shows no consideration for others. The author uses a lack of work-related words. The author uses no indication of sadness. The author uses a nonchalant attitude. The author uses dismissive words.
Topic 1: The author is avoiding words that suggest poverty. The author uses language that is respectful. The author is willing to challenge the status quo. The author is separating independent clauses with a comma. The author uses non-confrontational language. The author is making a reasoned argument. The author is making a sweeping statement. The author is tolerant. The author is following the statement with a comma. The author is open-minded. The author is presenting an opinion, rather than a fact. The author is expressing their opinion in a civil manner. The author is separating independent clauses with a question mark. The author is separating two independent clauses with a period. The author is using factual and straightforward language.
Topic 2: The author shows respect. The author is trying to convey a message in a straightforward manner. The author is writing in a cheerful manner. The author is using uplifting language. The author is free of negative emotions. The author is humble. The author uses language that suggests comfort in interacting with others. The author is tolerant. The author is in good spirits. The author is avoiding words that suggest poverty. The author uses language that is respectful. The author is socially aware. The author is using positive emotion. The author uses simple and straightforward sentence structure. The author is using a positive tone.
Topic 3: The author is insensitive. The author is making a judgmental statement. The author is uncaring. The author uses dismissive words. The author is rude. The author is unable to control their anger. The author is using words indicating poverty. The author is using a resigned attitude. The author is scornful. The author is unenthusiastic. The author makes a statement without any politeness. The author is expressing resignation. The author is expressing confusion and disbelief. The author is using a negative tone. The author is avoiding words that suggest poverty.
Topic 4: The author uses no indication of sadness. The author uses language that is respectful. The author uses short and simple words. The author writes a simple conversation between two people. The author uses a relaxed writing style. The author speaks without filler words. The author uses punctuation sparingly. The author uses non-confrontational language. The author uses a calm and collected tone. The author uses no frills. The author uses a nonchalant attitude. The author uses language that suggests comfort in interacting with others. The author uses no words related to leisure. The author uses no neutral tone. The author uses understanding.

Table 8: Overview of Topics in Clustering 0

CLUSTERING 1: 8 Topics
Topic 1: The author is omitting articles. The author is omitting a verb. The author is avoiding words that suggest poverty. The author is unenthusiastic. The author is rude. The author is non-suggestive. The author makes a statement without any politeness. The author is separating independent clauses with a comma. The author is avoiding words related to motion perception. The author is avoiding words related to food or eating. The author is impolite. The author is being non-judgmental. The author is using a resigned attitude. The author uses a sparse writing style. The author is insensitive.
Topic 2: The author speaks without filler words. The author is trying to convey a message in a straightforward manner. The author is writing in a cheerful manner. The author removes unnecessary words. The author is using uplifting language. The author is writing in plain text. The author is using no agreement errors. The author is using one complete sentence. The author only uses words that are necessary. The author is using positive emotion. The author is tentative and noncommittal. The author uses punctuation sparingly. The author leaves sentences unfinished. The author is using a single word. The author uses a sparse writing style.
Topic 3: The author uses no indication of sadness. The author writes a simple conversation between two people. The author uses language that is respectful. The author uses short and simple words. The author's writing is well-written. The author uses language that suggests comfort in interacting with others. The author uses punctuation sparingly. The author uses the correct tense when writing. The author uses understanding. The author uses no frills. The author uses affirmative language. The author uses a relaxed writing style. The author uses correct capitalization. The author uses non-confrontational language. The author's grammar style is direct.
Topic 4: The author is writing in a cheerful manner. The author is attempting to create a sense of familiarity and connection with the reader. The author is easy to understand and relate to. The author is in good spirits. The author is enthusiastic and confident. The author is using a playful style. The author is using factual and straightforward language. The author is feeling content and at ease. The author is proud of their knowledge. The author draws the reader in and keeps them engaged. The author is creating an intimate atmosphere. The author uses a relaxed writing style. The author is using a lighthearted tone. The author makes the passage accessible to a wide range of readers. The author is humble.
Topic 5: The author is making a judgmental statement. The author is making a sweeping statement. The author is using words indicating poverty. The author is expressing confusion and disbelief. The author uses dismissive words. The author is avoiding words that suggest poverty. The author is willing to challenge the status quo. The author is scornful. The author is insensitive. The author is viewing the situation in extreme, black-and-white terms. The author is expressing their opinion in a civil manner. The author is making assumptions without evidence. The author is unable to control their anger. The author is expressing resignation. The author makes a statement without any politeness.
Topic 6: The author uses language that is respectful. The author is tolerant. The author shows respect. The author is respecting boundaries. The author is open-minded. The author is avoiding words that suggest poverty. The author uses language that suggests comfort in interacting with others. The author is socially aware. The author is using gender-neutral terms. The author is sensitive. The author uses non-confrontational language. The author is taking appropriate action. The author is expressing prosocial behaviors. The author is socially responsible. The author is professional and appropriate.
Topic 7: The author uses a nonchalant attitude. The author makes a statement without any politeness. The author uses no neutral tone. The author is using negative emotion. The author lacks authority. The author shows no consideration for others. The author is using language that is considered taboo. The author uses language that suggests a lack of certainty. The author is using the wrong verb form. The author uses dismissive words. The author speaks without filler words. The author is using words expressing lack. The author uses no indication of sadness. The author is unenthusiastic. The author uses a tone of exasperation.

Table 9: Overview of Topics in Clustering 1

CLUSTERING 2: 10 Topics
Topic 0: The author is omitting a verb. The author is omitting articles. The author is trying to convey a message in a straightforward manner. The author is avoiding words related to motion perception. The author uses a sparse writing style. The author is avoiding words related to food or eating. The author leaves sentences unfinished. The author removes unnecessary words. The author is direct and to the point, avoiding unnecessary words or phrases. The author uses punctuation sparingly. The author expresses no words of fulfillment. The author is tentative and noncommittal. The author is using a single word. The author speaks without filler words. The author is using a single independent clause.
Topic 1: The author is uncaring. The author is insensitive. The author is unable to control their anger. The author is unenthusiastic. The author is rude. The author is using a resigned attitude. The author is scornful. The author is using a negative tone. The author uses dismissive words. The author is expressing resignation. The author is impolite. The author shows no consideration for others. The author is using negative emotion. The author is making a judgmental statement. The author makes a statement without any politeness.
Topic 2: The author is professional and appropriate. The author is using factual and straightforward language. The author is taking appropriate action. The author is encouraging the reader. The author makes a clear suggestion. The author is benefiting others. The author is following the statement with a comma. The author is avoiding words that suggest poverty. The author is pragmatic. The author is using words indicating wealth. The author is using formal and professional language. The author is precise with number agreement. The author is goal-oriented. The author makes the passage accessible to a wide range of readers. The author uses language that is respectful.
Topic 3: The author is tolerant. The author shows respect. The author uses language that is respectful. The author uses language that suggests comfort in interacting with others. The author is implying a familial relationship. The author is using a personal perspective. The author is expressing prosocial behaviors. The author is socially aware. The author is humble. The author is using uplifting language. The author is sensitive. The author is using words indicating family. The author is respecting boundaries. The author is self-aware. The author uses a calm and collected tone.
Topic 4: The author makes a statement without any politeness. The author uses no neutral tone. The author speaks without filler words. The author uses language that suggests a lack of certainty. The author lacks authority. The author lacks qualifiers or hedging language. The author uses a nonchalant attitude. The author lacks strong evidence or logical reasoning. The author leaves no room for misinterpretation or misunderstanding. The author shows no consideration for others. The author uses dismissive words. The author uses non-confrontational language. The author uses no indication of sadness. The author uses a sparse writing style. The author uses a lack of work-related words.
Topic 5: The author is writing in a cheerful manner. The author speaks without filler words. The author is using uplifting language. The author uses a relaxed writing style. The author shows respect. The author speaks confidently. The author uses a calm and collected tone. The author uses no indication of sadness. The author uses short and simple words. The author is vivacious. The author keeps sentences short. The author uses language that is respectful. The author is using positive emotion. The author removes unnecessary words. The author is writing in plain text.
Topic 6: The author uses non-confrontational language. The author is avoiding words that suggest poverty. The author makes a statement without any politeness. The author is separating independent clauses with a comma. The author is being non-judgmental. The author uses language that is respectful. The author expresses their opinion without worry. The author is presenting an opinion, rather than a fact. The author is separating two independent clauses with a period. The author is tolerant. The author is expressing their opinion in a civil manner. The author is using gender-neutral terms. The author is avoiding any words related to self-harm. The author is open-minded. The author is unbiased.
Topic 7: The author uses no indication of sadness. The author writes a simple conversation between two people. The author uses punctuation sparingly. The author uses short and simple words. The author uses language that is respectful. The author uses no frills. The author uses non-confrontational language. The author uses no words related to leisure. The author uses minimal grammar errors. The author uses the correct tense when writing. The author uses understanding. The author uses language that suggests comfort in interacting with others. The author uses no neutral tone. The author's writing is well-written. The author's grammar style is direct.
Topic 8: The author is attempting to create a sense of familiarity and connection with the reader. The author is writing in a cheerful manner. The author is easy to understand and relate to. The author is using a playful style. The author is in good spirits. The author is enthusiastic and confident. The author draws the reader in and keeps them engaged. The author is feeling content and at ease. The author is focused on activities that can provide enjoyment, relaxation, and amusement. The author is using a lighthearted tone. The author uses a relaxed writing style. The author is using words related to leisure. The author uses colorful language. The author is engaging in a friendly conversation. The author is poetic and lyrical.
Topic 9: The author is making a sweeping statement. The author is making assumptions without evidence. The author is expressing confusion and disbelief. The author is willing to challenge the status quo. The author is making a judgmental statement. The author uses cynicism. The author is viewing the situation in a more nuanced way. The author uses a critical tone. The author is making a reasoned argument. The author is using words indicating poverty. The author is expressing skepticism. The author is viewing the situation in extreme, black-and-white terms. The author is using a pessimistic outlook. The author is avoiding words that suggest poverty. The author is thought-provoking.

Table 10: Overview of Topics in Clustering 2

Location	Model	Cheerful	Judgmental	Inquiry	Analytical	Direct	Unenthusiastic	Professional	KL
NYC	Reddit	0.2309	0.1813	0.0718	0.1599	0.0186	0.1062	0.2311	-
	Llama	0.3167	0.1481	0.0930	0.0321	0.1053	0.1542	0.1502	0.2293
	Mistral	0.3452	0.1451	0.0773	0.1112	0.0000	0.0461	0.2748	0.4038
	GPT	0.3686	0.0931	0.0317	0.2223	0.0000	0.0101	0.2739	0.5836
LA	Reddit	0.1046	0.3332	0.1005	0.2205	0.1122	0.0216	0.1070	-
	Llama	0.1094	0.3157	0.0786	0.0936	0.1233	0.1399	0.1391	0.1477
	Mistral	0.1267	0.3906	0.0306	0.1230	0.0000	0.0419	0.2869	2.3950
	GPT	0.1934	0.2766	0.0138	0.2850	0.0126	0.0000	0.2182	0.7254
Canada	Reddit	0.0344	0.3317	0.1448	0.2561	0.0687	0.0000	0.1641	-
	Llama	0.0279	0.3440	0.1045	0.1185	0.1208	0.0384	0.2456	0.1347
	Mistral	0.0147	0.2711	0.0000	0.2293	0.0000	0.0000	0.4847	4.3990
	GPT	0.0187	0.1463	0.0170	0.3561	0.0000	0.0000	0.4617	1.7463
Chicago	Reddit	0.2773	0.1984	0.0968	0.1585	0.0000	0.1461	0.1226	-
	Llama	0.3784	0.0992	0.0646	0.1317	0.1229	0.1193	0.0836	0.1963
	Mistral	0.5984	0.1068	0.0000	0.1674	0.0000	0.0000	0.1273	4.9834
	GPT	0.6061	0.0666	0.0000	0.2658	0.0000	0.0000	0.0613	5.0899
Australia	Reddit	0.0876	0.4021	0.0865	0.1762	0.1396	0.0000	0.1075	-
	Llama	0.0872	0.2933	0.0541	0.0850	0.2342	0.0902	0.1557	0.1845
	Mistral	0.1382	0.3955	0.0000	0.1565	0.0000	0.0000	0.3095	4.5974
	GPT	0.1703	0.2408	0.0204	0.2341	0.0000	0.0000	0.3342	3.0424
Texas	Reddit	0.0000	0.4822	0.0435	0.1633	0.2067	0.0153	0.0887	-
	Llama	0.0000	0.4487	0.0000	0.0000	0.2753	0.0787	0.1971	4.2112
	Mistral	0.0000	0.5206	0.0000	0.1002	0.0166	0.0000	0.3624	1.5946
	GPT	0.0195	0.3547	0.0158	0.2344	0.0515	0.0000	0.3238	0.5953
Singapore	Reddit	0.0555	0.3066	0.0394	0.1483	0.0265	0.1663	0.2570	-
	Llama	0.0722	0.2531	0.1630	0.0606	0.1299	0.1263	0.1945	0.1961
	Mistral	0.0000	0.2769	0.0151	0.2356	0.0442	0.0334	0.3945	1.2625
	GPT	0.0000	0.1413	0.0000	0.4209	0.0000	0.0000	0.4376	5.8936
Paris	Reddit	0.2415	0.1324	0.1864	0.1508	0.1770	0.0606	0.0510	-
	Llama	0.4118	0.0973	0.2361	0.1096	0.0194	0.0604	0.0650	0.2949
	Mistral	0.5012	0.0969	0.1152	0.1711	0.0300	0.0000	0.0853	1.4508
	GPT	0.6757	0.0733	0.0000	0.1366	0.0000	0.0000	0.1142	8.7819
Mumbai	Reddit	0.3244	0.1657	0.1035	0.1778	0.1458	0.0246	0.0578	-
	Llama	0.5022	0.1015	0.1187	0.1337	0.0232	0.0468	0.0735	0.2138
	Mistral	0.6070	0.0860	0.0291	0.2027	0.0000	0.0000	0.0749	3.5517
	GPT	0.6649	0.0122	0.0000	0.2264	0.0000	0.0000	0.0963	5.8292
Berlin	Reddit	0.2071	0.2649	0.1536	0.1740	0.1070	0.0000	0.0930	-
	Llama	0.3864	0.1451	0.2056	0.1448	0.0123	0.0757	0.0298	0.3547
	Mistral	0.4841	0.1315	0.1006	0.1516	0.0000	0.0000	0.1320	2.2924
	GPT	0.5578	0.0934	0.0375	0.2109	0.0000	0.0000	0.1002	2.4736
London	Reddit	0.1327	0.2917	0.1500	0.1992	0.0773	0.0541	0.0946	-
	Llama	0.1789	0.1726	0.1627	0.0848	0.2143	0.1255	0.0609	0.1885
	Mistral	0.2850	0.1863	0.1021	0.2507	0.0000	0.0470	0.1287	1.6031
	GPT	0.4371	0.1447	0.0801	0.2172	0.0000	0.0102	0.1105	1.7822
Korea	Reddit	0.2231	0.3177	0.1396	0.1592	0.0717	0.0000	0.0883	-
	Llama	0.3606	0.2025	0.1815	0.1050	0.0000	0.0950	0.0550	1.5712
	Mistral	0.4713	0.1845	0.0218	0.1896	0.0000	0.0000	0.1326	1.6652
	GPT	0.5278	0.0227	0.0000	0.3156	0.0000	0.0000	0.1336	4.9051
Scotland	Reddit	0.1139	0.3761	0.1133	0.2172	0.1352	0.0000	0.0440	-
	Llama	0.1014	0.2918	0.0663	0.0610	0.3879	0.0130	0.0783	0.2772
	Mistral	0.1959	0.2602	0.0211	0.2937	0.0000	0.0000	0.2287	2.9717
	GPT	0.2837	0.1515	0.0491	0.3538	0.0276	0.0000	0.1340	0.3922
Manchester	Reddit	0.2620	0.2419	0.1425	0.1420	0.0823	0.0894	0.0395	-
	Llama	0.2030	0.1135	0.1201	0.0900	0.2133	0.1704	0.0893	0.1705
	Mistral	0.3094	0.2017	0.0901	0.1885	0.0000	0.0664	0.1438	1.6927
	GPT	0.4287	0.1245	0.0155	0.2397	0.0267	0.0326	0.1319	0.4082

Table 11: Comparison based on locations using KL-Divergence between LLMs and Reddit's Distribution

Occupation	Model	Cheerful	Judgmental	Inquiry	Analytical	Direct	Unenthusiastic	Professional	KL
Journalist	Reddit	0.0000	0.3906	0.0667	0.1361	0.2950	0.0000	0.1116	-
	Llama	0.0000	0.4517	0.0525	0.0942	0.2284	0.0000	0.1733	0.0358
	Mistral	0.0000	0.5418	0.0422	0.0915	0.0395	0.0000	0.2850	0.4455
	GPT	0.0000	0.2945	0.0174	0.2381	0.0416	0.0000	0.4083	0.5567
Architect	Reddit	0.3294	0.1122	0.2198	0.2002	0.0386	0.0256	0.0742	-
	Llama	0.4246	0.0000	0.2672	0.2272	0.0000	0.0000	0.0810	3.4404
	Mistral	0.4676	0.0377	0.0749	0.3097	0.0000	0.0000	0.1101	1.3869
	GPT	0.6666	0.0000	0.0000	0.2588	0.0000	0.0000	0.0746	8.0421
Engineer	Reddit	0.1165	0.0810	0.3853	0.2141	0.0618	0.0307	0.1107	-
	Llama	0.1037	0.0211	0.4562	0.1477	0.0000	0.0000	0.2713	1.8878
	Mistral	0.2435	0.0000	0.1940	0.1842	0.0000	0.0000	0.3783	3.5854
	GPT	0.2522	0.0000	0.0932	0.1612	0.0000	0.0000	0.4934	3.8629
Finance Manager	Reddit	0.0000	0.2946	0.1776	0.2848	0.1545	0.0000	0.0887	-
	Llama	0.0000	0.1199	0.1639	0.3865	0.0342	0.0000	0.2956	0.3182
	Mistral	0.0000	0.0235	0.0798	0.4335	0.0000	0.0000	0.4633	3.8891
	GPT	0.0000	0.0089	0.0581	0.4168	0.0000	0.0000	0.5162	4.2339
Photographer	Reddit	0.1109	0.1730	0.1975	0.1802	0.0645	0.1073	0.1666	-
	Llama	0.2613	0.1424	0.1960	0.0983	0.0294	0.0572	0.2154	0.1247
	Mistral	0.2690	0.2207	0.0000	0.1552	0.0000	0.0000	0.3551	7.5270
	GPT	0.3274	0.0580	0.0000	0.2151	0.0000	0.0000	0.3995	7.6579
Teacher	Reddit	0.0000	0.4604	0.0249	0.1386	0.0984	0.0713	0.2064	-
	Llama	0.0000	0.4717	0.0000	0.0391	0.0832	0.0836	0.3224	0.5586
	Mistral	0.0000	0.3644	0.0000	0.1645	0.0000	0.0274	0.4437	2.5135
	GPT	0.0000	0.1438	0.0000	0.2597	0.0000	0.0000	0.5965	4.2033
Lawyer	Reddit	0.0000	0.2931	0.0499	0.0836	0.4987	0.0000	0.0746	-
	Llama	0.0000	0.2901	0.0529	0.0435	0.3043	0.0000	0.3093	0.1952
	Mistral	0.0000	0.2360	0.0089	0.1204	0.1286	0.0000	0.5061	0.6523
	GPT	0.0000	0.1348	0.0000	0.1629	0.1039	0.0000	0.5984	1.7991
Chef	Reddit	0.2620	0.2463	0.1067	0.1440	0.0000	0.0956	0.1455	-
	Llama	0.4764	0.0000	0.2317	0.0000	0.0000	0.0841	0.2078	8.0842
	Mistral	0.2800	0.0000	0.0000	0.0000	0.0000	0.0000	0.7200	12.3071
	GPT	0.5156	0.0000	0.0000	0.0000	0.0000	0.0000	0.4844	12.2048
Nurse	Reddit	0.0000	0.5215	0.0331	0.1111	0.0707	0.0766	0.1869	-
	Llama	0.0000	0.4396	0.0163	0.0532	0.0230	0.0431	0.4248	0.1645
	Mistral	0.0000	0.2453	0.0000	0.1515	0.0000	0.0000	0.6031	3.7981
	GPT	0.0000	0.1387	0.0000	0.1673	0.0000	0.0000	0.6939	4.0580
Doctor	Reddit	0.0000	0.4412	0.0578	0.1651	0.0178	0.0222	0.2955	-
	Llama	0.0000	0.3548	0.0243	0.1312	0.0000	0.0000	0.4896	0.8029
	Mistral	0.0000	0.1682	0.0000	0.1544	0.0000	0.0000	0.6774	2.1266
	GPT	0.0000	0.1018	0.0000	0.1720	0.0000	0.0000	0.7261	2.3096

Table 12: Comparison based on occupations using KL-Divergence between LLMs and Reddit's Distribution

Age	Model	Cheerful	Judgmental	Inquiry	Analytical	Direct	Unenthusiastic	Professional	KL
GenZ	Reddit	0.2418	0.1072	0.0708	0.1516	0.2154	0.1475	0.0658	-
	Llama	0.1682	0.0676	0.1116	0.0727	0.2980	0.1960	0.0858	0.0869
	Mistral	0.0652	0.0000	0.0976	0.6279	0.1452	0.0000	0.0641	5.5082
	GPT	0.2057	0.0000	0.1251	0.2654	0.1555	0.0762	0.1720	2.2477
Millennial	Reddit	0.0802	0.4024	0.0384	0.2798	0.0203	0.0924	0.0864	-
	Llama	0.0865	0.2668	0.1061	0.0925	0.1706	0..2569	0.0206	0.4159
	Mistral	0.0731	0.3415	0.0795	0.2598	0.0290	0.1533	0.0637	0.0829
	GPT	0.2039	0.1052	0.0379	0.4967	0.0000	0.1154	0.0631	0.5053
GenX	Reddit	0.2778	0.2330	0.0538	0.2318	0.0268	0.1032	0.0736	-
	Llama	0.3006	0.1030	0.1029	0.1448	0.1687	0.1799	0.0000	1.6381
	Mistral	0.3826	0.1186	0.0428	0.3058	0.0000	0.1037	0.0465	0.5707
	GPT	0.3778	0.1052	0.0286	0.3755	0.0178	0.0481	0.0470	0.1449
BabyBoomer	Reddit	0.1958	0.3527	0.0917	0.2310	0.0206	0.0000	0.1082	-
	Llama	0.3541	0.2022	0.0748	0.2057	0.0000	0.0977	0.0655	0.5748
	Mistral	0.3769	0.1255	0.0149	0.3860	0.0000	0.0200	0.0767	0.7162
	GPT	0.4099	0.0557	0.0116	0.4478	0.0000	0.0000	0.0750	0.9775

Table 13: Comparison based on age using KL-Divergence between LLMs and Reddit's Distribution

Location	KL Llama N	KL Llama B	KL Mistral N	KL Mistral B	KL GPT N	KL GPT B
New York City	0.1717	0.1972	0.1200	0.0405	0.1484	0.0921
Los Angeles	0.0542	0.0896	0.1090	0.0484	0.0693	0.6416
Canada	0.1428	0.0788	2.1314	3.2063	0.0769	0.7426
Chicago	0.4357	0.3015	1.5677	1.5092	0.1330	0.0858
Australia	0.1998	0.0948	0.1256	1.5676	0.0989	0.8331
Texas	4.1788	2.3185	1.7326	0.8995	0.1824	0.0520
Singapore	0.1228	0.0346	0.2456	0.0612	0.0333	1.0925
Paris	0.2064	0.1163	0.2098	0.0998	1.5325	1.5741
Mumbai	0.1102	0.0890	0.1870	0.1110	0.0366	1.2635
Berlin	0.0750	0.0850	0.0646	0.1015	0.1434	0.0866
London	0.3007	0.0744	0.2078	0.0854	0.1045	0.0912
Korea	0.2335	0.6666	0.0484	0.0453	0.0470	0.0080
Scotland	0.4773	0.2779	0.4670	0.2791	0.1913	0.0878
Manchester	0.2148	0.1321	0.1514	0.0794	0.4996	0.3008

Table 14: KL Divergence Values by Location, where **N** denotes values calculated between 'Persona' and 'No Persona' distribution & where **B** denotes values calculated between 'Persona' and 'Baseline' distribution

Profession	KL Llama N	KL Llama B	KL Mistral N	KL Mistral B	KL GPT N	KL GPT B
Journalist	0.0771	0.8426	0.0898	0.0071	0.1491	0.0542
Architect	1.2447	0.7645	0.2359	0.3258	2.4924	2.0776
Engineer	0.1185	1.8575	0.0925	0.1438	0.1075	0.2100
Finance Manager	0.6938	1.3492	0.4283	1.6099	0.4351	1.7502
Photographer	0.1616	0.2844	2.3460	4.4483	0.1835	5.2143
Teacher	0.0695	0.2116	0.0412	1.3125	0.0311	1.1945
Lawyer	0.9301	0.6466	0.4965	0.9090	0.2213	2.5835
Chef	3.8673	4.6765	4.7182	6.2615	4.9812	5.3624
Nurse	0.3088	0.4383	0.0304	0.1379	0.0217	0.0929
Doctor	0.1156	2.8285	0.0092	0.0494	0.0493	0.4718

Table 15: KL Divergence Values by Profession, where **N** denotes values calculated between 'Persona' and 'No Persona' distribution & where **B** denotes values calculated between 'Persona' and 'Baseline' distribution

Age	KL Llama N	KL Llama B	KL Mistral N	KL Mistral B	KL GPT N	KL GPT B
GenZ	0.4778	0.2229	0.4987	1.8484	4.1742	1.2467
Millennial	0.2152	0.0361	0.2992	0.0824	0.2545	0.0447
BabyBoomer	0.0186	0.0860	0.0996	0.1717	0.1147	0.1110
GenX	1.6863	0.2701	0.0458	0.0362	0.1308	0.0525

Table 16: KL Divergence Values by Age, where **N** denotes values calculated between 'Persona' and 'No Persona' distribution & where **B** denotes values calculated between 'Persona' and 'Baseline' distribution