Improving Linguistic Diversity of Large Language Models with Possibility Exploration Fine-Tuning

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have made significant strides in replicating human-like abilities, there are concerns about a reduction in the linguistic diversity of their outputs. This results in the homogenization of viewpoints and perspectives, as well as the underrepresentation of specific demographic groups. Although several fine-tuning and prompting techniques have been suggested to tackle the issue, they are often tailored to specific tasks or come with a substantial increase in computational cost and latency. This makes them challenging to apply to applications that demand very low latency, such as chatbots and virtual assistants. We propose Possibility Exploration Fine-Tuning (PEFT), a task-agnostic framework that enhances the text diversity of LLMs without increasing latency or computational cost. Given the same prompt, models fine-tuned with PEFT can simultaneously generate multiple diverse responses, each corresponding with a controllable possibility number. Experiments with Mistral 7B and LLAMA 2 on open-domain dialogue generation demonstrate that PEFT significantly enhances output diversity, as shown by a lower similarity between candidate responses. As PEFT focuses more on semantic diversity rather than lexical diversity, it can remarkably reduce demographic bias in dialogue systems.

1 Introduction

LLMs represent a significant advancement in the field of artificial intelligence, specifically in natural language processing (NLP). These models are designed to perform various tasks, from text classification to question-answering and logical reasoning, through natural language prompts, even without task-specific training (Achiam et al., 2023; Touvron et al., 2023; Jiang et al., 2023). The recipe for their success includes very large models trained on vast amounts of unfiltered internet data, which raises critical concerns about the perpetuation and amplification of biases or the degradation of content diversity (Gallegos et al., 2023). One of the primary concerns is that LLMs tend to be inherently conservative in their output. They are designed to predict the most likely words or sequences based on patterns observed in their training data. As a result, the generated text tends to closely align with the dominant narratives, ideas, and writing styles present in the datasets they were trained on. This can lead to a homogenization of content, where creative outliers and genuinely novel ideas are underrepresented. Studies by (Santurkar et al., 2023; Durmus et al., 2023) highlight that LLMs generate an unequal representation of views. Hence, future LLMs trained on such homogenized content may exacerbate the issue, perpetuating this cycle. The decline in diversity also presents significant challenges in other NLP areas, such as synthetic dataset production (Chung et al., 2023) or open-domain dialogue generation (Lee et al., 2023).

Diversity in text generation has been extensively studied. Several approaches have been proposed, such as retraining the models on more balanced datasets (Zmigrod et al., 2019; Garimella et al., 2022; Solaiman and Dennison, 2021), modifying the optimization objectives (Woo et al., 2023; Garimella et al., 2021; Shao et al., 2019), or using a conditional variational inference framework (Bao et al., 2020). Post-editing approaches, such as modifying the decoding algorithms (Su et al., 2022; Holtzman et al., 2019; Fan et al., 2018) or optimizing the input prompts (Hayati et al., 2023; Lahoti et al., 2023; Mattern et al., 2022), can also be used to increase text diversity and do not require additional training. However, these methods either increase model complexity, failing to achieve a satisfactory level of diversity, or significantly increase inference latency and computational cost.

This paper introduces Possibility Exploration Fine-Tuning (PEFT), a straightforward fine-tuning framework designed to enhance the text diversity of pre-trained LLMs. Our objective is to generate multiple distinct candidate responses to a single prompt, ensuring minimal semantic similarity among them. This is achieved by fine-tuning LLMs using a Possibility Exploration (PE) dataset, where each prompt is paired with several unique responses. Additionally, we propose negative fine-tuning frameworks to further boost diversity and allow for greater control over the generated content. One major advantage of our approach is that it does not necessitate any architectural changes, making it versatile and applicable to any pre-trained LLMs. With PEFT, models can produce multiple responses simultaneously or generate a single creative response without any additional latency.

To demonstrate the effectiveness of our approach, we focus on applying PEFT to open-domain dialogue generation tasks, where diversity and latency are key considerations. Experiments using Mistral 7B and LLAMA 2 reveal that our method significantly increases the diversity of generated responses compared to the base model without fine-tuning. In persona generation test, PEFT is effective in significantly reducing the overrepresentation of dominant demographic groups, offering an alternative to enhance fairness in LLMs.

2 Related work

Early methods to increase diversity involved modifying the conventional maximum likelihood training objective of text generation models. Shao et al. (2019) proposes a planning-based hierarchical variational model for generating long and diverse texts. Variational frameworks, employed by Du et al. (2022) and Bao et al. (2020), utilize randomly sampled latent variables to control the content of responses. Lee et al. (2023) introduces an empirical Bayes framework to develop a bayesian open-domain dialogue agent, achieving superior results in both diversity and coherence when compared to variational frameworks. These strategies, however, significantly elevate training complexity and inference latency, and necessitate specific model architectures that might not be able to leverage existing pre-trained LLMs.

Another direction of research has been to improve training data by mitigating biases related to gender, race, etc., aiming to generate more representative datasets. Techniques include creating new examples for underrepresented groups (Zmigrod et al., 2019; Qian et al., 2022), upsampling low-bias examples (Garimella et al., 2022; Han et al., 2021), or building a new dataset from curated exemplary samples (Solaiman and Dennison, 2021). Our proposed method aligns with this direction but is broader in scope, not limited to mitigating biases and stereotypes but enhancing overall text diversity for any open-ended text generation task.

A common strategy for enhancing text diversity modifies the decoding process. Techniques like diverse beam search (Vijayakumar et al., 2016), nucleus sampling (Holtzman et al., 2019), Top-K sampling (Fan et al., 2018), and logit suppression (Chung et al., 2023) aim to produce a broader set of outputs by not solely focusing on the most probable tokens. Contrastive search decoding (Su and Collier, 2022), in particular, has shown to improve both diversity and coherence. We demonstrate that models fine-tuned with PEFT can be combined with these decoding methods to further enrich diversity.

Recent studies explore prompt optimization to improve diversity, including iterative prompting to uncover varied responses to the same input. Hayati et al. (2023) introduces criteria-based diversity prompting to extract and ground diverse perspectives from LLMs. Lahoti et al. (2023) proposes a technique called, collective critiques and self-voting, to enhance text diversity concerning gender and culture. However, iterative prompting techniques substantially increase computational costs and latency, which may not be suitable for applications like dialogue systems. In contrast, our method can generate multiple responses simultaneously or produce a single creative response without additional latency.

3 Methodology

3.1 Problem definition

Given the prompt P, our goal is to generate a list of candidate responses, L, where each response is semantically distinct from the others. This is crucial for applications such as brainstorming tools, creative writing assistants, or other prompting techniques that require reasoning over multiple solutions (Muralidharan and Thomas, 2024; Wang et al., 2022). In scenarios that require a single but creative response R, such as dialogue modeling, one can simply sample a response from the list L. If the list L is sufficiently diverse, then the response R will likely be unique. A proficient generation model should produce responses that are diverse in content and contextually relevant to the given prompt. Latency is also critical for applications like real-time chatbots or interactive storytelling, where the model must generate responses quickly to ensure seamless user interaction.

3.2 Decoding methods

Temperature sampling (Holtzman et al., 2019; Fan et al., 2018) adjusts the randomness of the generated text by modifying the distribution of predicted probabilities with a temperature parameter ; higher temperatures lead to more creative outputs. To generate N diverse responses for a single prompt, we can set a high temperature value and generate responses N times.

Diverse Beam Search (DBS) (Vijayakumar et al., 2016), an alternative to beam search that decodes a list of diverse outputs by introducing mechanisms to explicitly encourage diversity among the candidates in the beam. This is typically achieved by partitioning the beam into several groups and adding a diversity-promoting term to the scoring function that guides the selection of candidates.

3.3 List Prompting

Decoding methods, such as temperature sampling, do not account for semantic differences at the sentence level, as they generate responses independently. As a result, while the responses may vary in wording, their semantic meanings may remain similar. Inspired by recent work on diverse perspective generation by Hayati et al. (2023), we introduce List Prompting as a general framework for multi-response generation using the following template: *I want to <task description>. List a diverse set of <N> possible responses:*

An example of List Prompting for dialogue generation is shown in Figure 1. As we can see, the generation of later candidates is influenced by earlier generated candidates, ensuring they are semantically different at the sentence level. Note that the latency of this method increases proportionally to the number of generated responses.

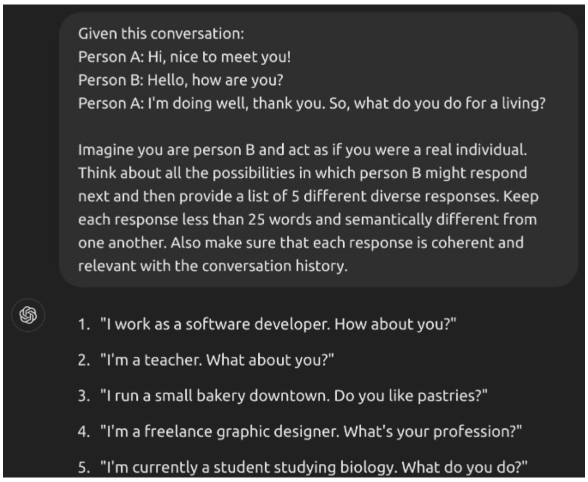

Figure 1: An example of List Prompting for open-domain dialogue generation

4 Possibility exploration fine-tuning

4.1 One-to-many dataset

Despite the inherent one-to-many (OTM) mapping nature of open-ended text generation, where each input prompt can yield multiple correct responses, current LLMs are predominantly fine-tuned on instruction-following or task-specific datasets that enforce a one-to-one (OTO) mapping. This means that each input prompt is accompanied by a single response. We refer to this approach as one-to-one fine-tuning (OTOFT). Although several studies have shown that OTOFT can improve the accuracy and performance of LLMs for specific tasks, its impact on the diversity of the output remains under-researched.

To address the one-to-many nature and potentially increase output diversity, we propose a method called one-to-many fine-tuning (OTMFT). OTMFT uses a OTM dataset to fine-tune LLMs for specific tasks. An OTM dataset is derived from a standard one-to-one mapping dataset. For each root pair of prompt-response (p, r), we generate N child samples $(p, r_1), (p, r_2), \ldots, (p, r_N)$, where each response r_i is a valid reply to the prompt p and is semantically distinct from all other responses. This generation process can be conducted by human annotators or advanced LLM models. In this study, we utilize GPT-4o and List Prompting techniques to generate multiple distinct responses for the same prompt.

OTMFT employs standard likelihood training, where all training samples corresponding to the same prompt are batched together. This fine-tuning process helps to flatten the probability distribution, allowing decoding techniques like temperature sampling to generate more diverse responses.

4.2 Possibility exploration dataset

Before presenting PEFT, we first introduce the Possibility Exploration (PE) dataset. We accompany each OTM training sample (p, r_i) with a *possibility number* k_i, indicating that the response r_i is the k_i-

Prompt (P)	i	k_i	Response (R)
Given this conversation: Person B: Hi, how are you? Person A: I am good. What do you do for work? Consider various ways in which Person B might respond, and provide the one corresponding to possibility number #k_i.	1	4	I am a doctor.
	2	8	I am a teacher.
	3	1	I am a singer.

Figure 2: An simplified example of a PE training batch with added possibility numbers. Full template can be found in Appendix A.2.2.

th possible response out of all possible responses for prompt p. The inclusion of a possibility number in each prompt helps in the following ways: (1) It assists the model in understanding the reasons behind differences in responses, even when the same prompt is given; (2) It provides a degree of control over the inference process, allowing the possibility number k to be changed to elicit different responses; (3) It enables negative training (PEFT), which further enhances the dissimilarity between responses.

Given an OTM batch of training samples $(p, r_1), (p, r_2), \ldots, (p, r_N)$, we construct a PE training batch by incorporating an additional instruction into the prompt p, as shown in Figure 2. Specifically, we instruct the model to contemplate all possible responses for the given prompt and then produce a response corresponding to possibility k_i, where k_i is an integer randomly sampled from $[1, .., M]$, with M being a hyper-parameter and $M > N$. Consequently, a PE batch of training samples will be $(p, k_1, r_1), (p, k_2, r_2), \ldots, (p, k_N, r_N)$. Figure 2 shows an example of PE training batch for open-domain dialogue generation task.

4.3 PEFT

We propose *PEFT*, which is based on unlikelihood training (Welleck et al., 2019). This approach aims to increase the dissimilarity between responses and enhance the impact of the possibility number. Unlikelihood training involves providing the model with both positive samples, which the model is encouraged to generate, and negative samples, which the model should avoid generating.

We use the PE batch of training samples $(p^+, k_1^+, r_1^+), \ldots, (p^+, k_N^+, r_N^+)$ as described in Section 4.2 as positive samples. For each positive sample (p^+, k_i^+, r_i^+), we generate $N-1$ corresponding negative samples (p^-, k_i^-, r_i^-) by keeping $p^- = p^+$ and $r_i^- = r_i^+$, while setting the possibility number $k_i^- = k_j^+$, where $j = 1, .., N$ and $j \neq i$. For example, as shown in Figure 2, the target response *I'm a doctor.* is considered a positive response when the possibility number $k = 4$ and a negative response when $k = 8$ or $k = 1$. In other words, we want each target response r to be accompanied by only one possibility number k, and vice versa.

The training with positive samples can be done with standard maximum likelihood estimation (MLE) as follow: $\mathcal{L}_{MLE}(\theta, p^+, k^+, r^+) =$

$$-\sum_{t=0}^{|r^+|} log\ \theta(r_t^+ | p^+, k^+, r_{<t}^+)$$

where θ is the model parameters, p^+ is the prompt, k^+ is the possibility number, r^+ is the response, and r_t^+ is the t-th token of r^+.

Training with negative samples can be done with unlikelihood objective as follow: $\mathcal{L}_{UL}(\theta, p^-, k^-, r^-) =$

$$-\sum_{t=0}^{|r^-|} \beta(r_t^-)\ log(1 - \theta(r_t^- | p^-, k^-, r_{<t}^-))$$

where r^- is the negative response and $\beta(r_t^-)$ is a candidate-dependent scale that controls how much the token t-th should be penalized. We set $\beta = 1$ for the first token of each word in r^-. The β values for the remaining tokens are set to 0. This helps to avoid the generation of out-of-vocabulary words.

We train the model with a mixture of likelihood and unlikelihood losses to avoid degradation as follows:

$$\mathcal{L} = \mathcal{L}_{MLE}(\theta, p^+, k^+, r^+) + \alpha \mathcal{L}_{UL}(\theta, p^-, k^-, r^-)$$

where α is the weight importance for unlikelihood loss. In this study, we set $\alpha = 0.5$. Note that all positive and negative samples of the same prompt should be included in the same batch.

To generate L different responses during inference, we first sample L possibility numbers from the range $[1..M]$ and then perform response generation independently and simultaneously for each sampled number.

5 Experiments

5.1 Tasks

We choose open-domain dialogue generation as the fine-tuning task because it necessitates both low latency and diverse outputs, which is the main aim of this study. Open-domain dialogue generation, often manifested as social chitchat modeling in NLP, refers to the ability of text generation models to generate coherent, human-like conversations across a wide range of topics.

Multiple responses generation. The task is to predict multiple possible responses for the next turn in a given dialogue context between two people. To create fine-tuning data for OTMFT and PEFT, we extract 1,000 dialogue contexts from BlendedSkillTask (Smith et al., 2020), ConvAI (Logacheva et al., 2018), TopicalChat (Gopalakrishnan et al., 2023), EmpatheticDialogues (Rashkin et al., 2018), and WizardOfWikipedia (Dinan et al., 2018), ensuring the number of samples per dataset is evenly distributed. For each dialogue context, we use GPT-4o and List Prompting to generate 4 different responses, resulting in a total of 4,000 context-response pairs. For OTOFT, 4,000 dialogue contexts are sampled, with each context accompanied by a single response that is also generated by GPT-4o. Hence, the total amount of data for OTOFT, OTMFT, and PEFT is equivalent. For the test set, 300 dialogue contexts are used.

Persona generation. Aside from improving the diversity of generated texts, we are also interested in evaluating the effectiveness of PEFT in debiasing dialogue systems or LLMs in general. We designed a test called the persona-generation test, in which the chatbot is asked to play the role of a random individual and then engage in a conversation with another person. The persona attributes of the chatbot, such as age and gender, are gradually revealed throughout the conversation. Since the chatbot has the freedom to determine its own personality and demographic information, we want to analyze if there is significant bias in the generated personas. We conducted 300 conversations for each chatbot and then aggregated the results for final assessment. Details of the experiment can be found in the Appendix A.3. The chatbots used for this persona-generation test are the same as those used for the multiple responses generation task. However, we only sampled a single response from all generated responses at each turn.

5.2 Metrics

5.2.1 Diversity

To measure lexical diversity, we utilize Distinct-1 and Distinct-2 scores (Liu et al., 2016), which account for the percentage of unique 1-grams or 2-grams in the entire collection of generated responses. For semantic diversity, we employ SBERT (Reimers and Gurevych, 2019) to compute the pairwise similarity between generated responses of each input prompt. The pairwise similarity is averaged across the test set, which is then used to calculate diversity as $1 - similarity$.

For the persona generation test, we use Shannon entropy (Shannon, 1948) to measure the randomness/diversity of the generated personas. Assume we generate a set of N personas, denoted as $P = \{P_1, P_2, ..., P_n\}$. Each persona P_i contains a set of attribute values $A_i = \{a_i^1, a_i^2, ..., a_i^m\}$, where a_i^j represents a particular attribute value (such as *female*) corresponding to the j-th attribute (such as *gender*). Let $A^j = \{a_1^j, a_2^j, ..., a_n^j\}$ be a collection of all values of the j-th attribute, extracted from P. Shannon entropy can be applied to measure the randomness score of the j-th attribute as follows:

$$H(A^j) = -\sum_{k}^{K} P(a_k^j) log(P(a_k^j))$$

where $H(A^j)$ represents the entropy of A^j, a_k^j represents each possible value of A^j, $P(a_k^j)$ represents the appearance ratio of the value a_k^j, and K is the number of distinct values of A^j. This paper only focuses on evaluating specific attributes: age group, gender, current location, occupation sector, and highest education level. The extraction/normalization of these attributes from the generated conversations is done by GPT-4o. See Appendix A.3.1 for details.

5.2.2 Coherence score

Given recent studies (Zheng et al., 2024) suggesting that LLMs can rival human performance in evaluating the quality of synthesized texts, we use GPT-4o and LLAMA 3 as coherence evaluators.

Previous studies often use the average rating (on a scale of 1 to 10) as the final measure of coherence. However, we found that automatic coherence evaluators tend to assign high coherence scores to safe and conservative responses, while giving lower scores to unconventional, creative but still coherent responses. Therefore, we propose using the percentage of incoherent responses as a coherence indicator. A response is considered incoherent if it receives a coherence rating of less than 6 (on a scale of 1-10) from both GPT-4o and LLAMA 3. Using the percentage of incoherent responses is also more intuitive for determining whether a

Methods	Dist-1 ↑	Dist-2 ↑	Diversity ↑	Incoherence ↓	Latency ↓
Base model					
DBS	0.108	0.452	0.356	2.2%	3x
List Prompting	0.151	0.565	0.583	7.9%	3.7x
Sampling (t=1.50)	0.135	0.547	0.383	3.6%	1x
OTOFT					1x
Sampling (t=1.00)	0.139	0.595	0.495	2.6%	
Sampling (t=1.25)	0.154	0.655	0.535	4.5%	
OTMFT					1x
Sampling (t=0.75)	0.133	0.529	0.522	3.1%	
Sampling (t=1.00)	0.150	0.604	0.565	4.0%	
PEFT					1x
Sampling (t=0.50)	0.130	0.484	0.530	2.3%	
Sampling (t=0.75)	0.149	0.561	0.585	3.9%	

Table 1: Performances of different decoding and fine-tuning methods for Mistral 7B in multiple response generation.

response generation model is suitable for deployment. More details on coherence evaluators can be found in Appendix A.2.3.

5.3 Parameters settings

We use the Huggingface repository[1] to conduct our experiments, employing Mistral 7B Instruct and LLAMA 2 7B Instruct as the pre-trained LLMs for fine-tuning. Each model is fine-tuned for one epoch using Qlora (Dettmers et al., 2024). The learning rate is set to 5e-5, with a batch size of 4 and a gradient accumulation of 2.

The number of possible target responses per input prompt, denoted as N, is set to 4 for all experiments. Additionally, for the multiple response generation task with Mistral 7B, we report the results with $N = 2$ and $N = 6$. The maximum value for the possibility number in PEFT is set to 9.

During inference and testing, each model is asked to generate 5 different responses per input prompt. We then calculate the diversity and coherence scores of these responses.

5.4 Comparison methods

Base model. We perform response generation using the original LLMs with zero-shot prompting and list prompting. For zero-shot prompting, we employ various decoding methods, including DBS, and temperature sampling. As we prioritize diversity, each decoding algorithm is configured with parameters that maximize output diversity without spoiling output coherence. For DBS, we employ hamming diversity as the objective term and set the diversity penalty to 5.0. For temperature sampling, we set the temperature value t to 1.5 for Mistral and 1.25 for LLAMA 2. We do not include contrastive search for comparison as the method is deterministic and can only generate a single response per prompt. The zero-shot prompt template can be found in Appendix A.2.1.

OTOFT. We fine-tune the base model using a one-to-one dataset with a MLE objective.

OTMFT. We fine-tune the base model using a one-to-many dataset with a MLE objective.

PEFT. We fine-tune the base model using a possibility exploration dataset with both MLE and unlikelihood objectives.

When comparing different fine-tuning techniques, we use temperature sampling as the decoding method with temperatures $t = \{0.5, 0.75, 1.0, 1.25\}$. For ease of comparing the diversity-coherence trade-offs between different methods, only optimal temperatures for each method are reported.

6 Experiment results

The experiment results are reported in Table 1-4.

Base LLMs without fine-tuning suffers significantly from low diversity and bias.

As shown in Table 1, despite having hyperparameters designed to maximize diversity, the Mistral base model achieves relatively low diversity scores at 0.356 with DBS and 0.383 when using temperature sampling set at 1.5. Appendix A.1 provides examples demonstrating that most generated responses, while varied in wording, are semantically similar. Surprisingly, the LLAMA 2

[1]https://github.com/huggingface

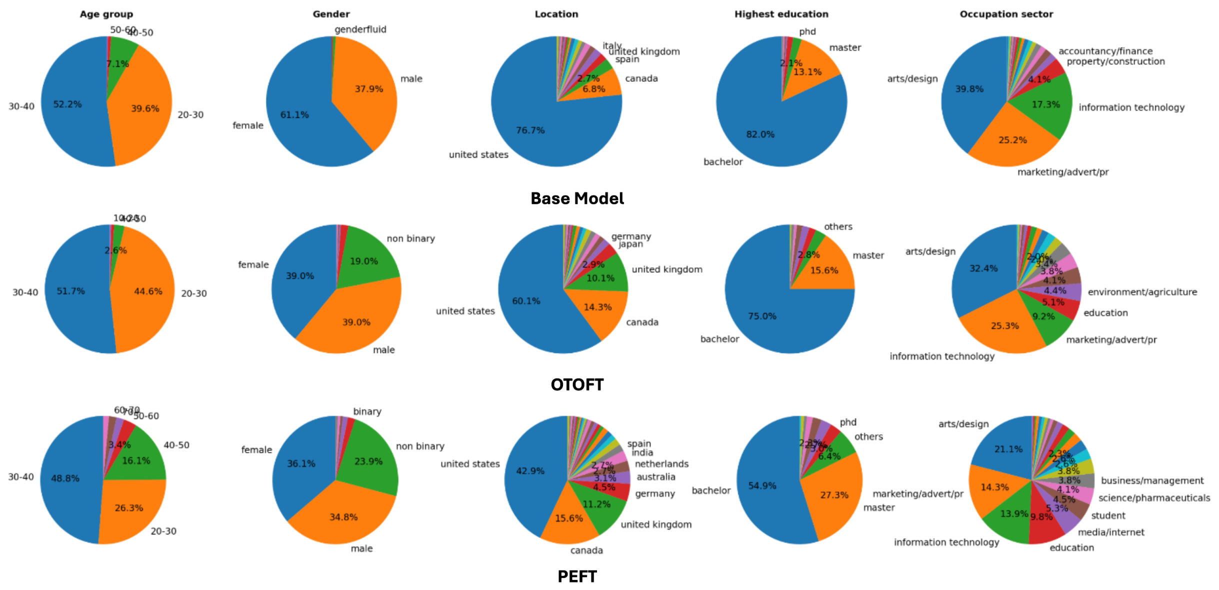

Figure 3: Persona demographic distributions extracted from 300 conversations with Mistral base and its fine-tuned models. All models use temperature sampling with $t = 1.0$.

Methods	Diversity	Incoherence
Base model		
DBS	0.422	7.0%
List Prompting	0.645	17.2%
Sampling (t=1.25)	0.479	8.9%
OTOFT		
Sampling (t=1.00)	0.513	5.3%
Sampling (t=1.25)	0.556	9.3%
OTMFT		
Sampling (t=0.75)	0.536	4.7%
Sampling (t=1.00)	0.579	7.2%
PEFT		
Sampling (t=0.50)	0.530	3.8%
Sampling (t=0.75)	0.583	6.3%

Table 2: Performances of different decoding and fine-tuning methods for LLAMA 2 in multiple response generation.

Responses/prompt	Diversity	Incoherence
OTMFT		
N = 2	0.449	1.9%
N = 4	0.479	2.1%
N = 6	0.489	2.4%
PEFT		
N = 2	0.501	2.2%
N = 4	0.530	2.3%
N = 6	0.548	2.5%

Table 3: Results of number of responses (N) per input in the training set when fine-tuning with OTMFT and PEFT. Mistral is used as the base model and temperature sampling ($t = 0.5$) is used for decoding.

model achieves a higher diversity score of 0.479 compared to Mistral, despite being less capable in general benchmarks (Jiang et al., 2023). This suggests that a model's higher accuracy does not necessarily correlate with greater diversity.

In the context of persona generation test, there is a noticeable sampling bias in the outputs of the Mistral base model. The bias predominantly favors certain demographic groups. For instance, more than 75% of the generated personas are located in the U.S., which is a significant overrepresentation, considering that the U.S. accounts for only about 4% of the global population. Also, there is a high frequency of personas possessing a Bachelor degree. This creates a skew towards more educated individuals and underrepresents those with lower educational backgrounds, such as high school diplomas or vocational training.

Switching between various decoding methods or tweaking parameters, such as increasing the temperature, can enhance diversity but not significantly. This is because diversity-focused decoding algorithms like temperature sampling and diverse beam search aim to boost diversity at the individual token level rather than in the overall semantics of the sentence. Additionally, higher lexical diversity does not always equate to higher semantic diversity. For example, the Mistral base model with high-

Methods	Shannon entropy ↑					
	Age	Gen	Loc	Edu	Occ	Avg
Base	1.3	1.0	1.6	0.9	2.5	1.5
OTOFT	1.2	1.7	2.1	1.2	3.0	1.8
OTMFT	1.4	1.7	2.5	1.6	3.4	2.1
PEFT	1.9	1.9	3.1	1.9	3.7	2.5

Table 4: Persona generation test with Mistral base and its fine-tuned models. *Age, Gen, Loc, Edu, Occ, and Avg* refer to the age group, gender, location, highest education, occupation sector, and average, respectively. All models use temperature sampling with $t = 1.0$.

temperature sampling ($t = 1.5$) achieves a lexical diversity Dist-2 score of 0.547, which is notably higher than the 0.484 score for PEFT ($t = 0.5$). However, the latter model has a higher semantic diversity score of 0.530 compared to 0.383 for the former. Similar lexical-semantic discrepancies are observed when comparing the lexical and semantic diversity scores from different fine-tuning methods, as noted in Table 1.

We observe that using List Prompting significantly improves diversity for both Mistral and LLAMA 2 models. This is because each candidate response is generated conditionally based on the previous ones, ensuring they are different. However, the use of List Prompting also results in a noticeable decline in coherence. The incoherence scores increase to 7.9% for Mistral and 17.2% for LLAMA 2. We believe this issue stems from the general performance of Mistral and LLAMA 2 in following instructions, rather than from the prompting technique itself.

Fine-tuning LLMs not only improves coherence but also diversity

As shown in Table 1, fine-tuned models achieve significant improvement in diversity over the base model despite using a lower temperature t. This results in better diversity-coherence trade-offs. For example, when using temperature sampling ($t = 1.0$), Mistral OTOFT significantly improves the diversity of the Mistral base model (sampling $t = 1.5$), increasing it from 0.383 to 0.495 while decreasing the incoherence rate from 3.6% to 2.6%. When comparing against DBS and List Prompting of the base model, OTOFT achieves a significantly better diversity-coherence trade-off while also providing more than a 3x increase in decoding speed. Similar improvements are also observed in LLAMA 2 experiments in Table 2.

When comparing OTOFT and OTMFT, the latter showed a clear improvement in both coherence and diversity scores, as demonstrated in both the Mistral and LLAMA 2 models.

PEFT surpasses OTMFT in both coherence-diversity trade-offs and data efficiency

When using temperature sampling with $t = 0.75$, PEFT further enhances the diversity of OTMFT, raising it from 0.522 to 0.585. This comes with an increase in the number of incoherent responses, from 3.1% to 3.9%. At a lower temperature sampling of $t = 0.5$, PEFT achieves a diversity/incoherence score of 0.530/2.3%, which is an improvement over OTMFT's 0.522/3.1% at $t = 0.75$. This demonstrates a better coherence-diversity trade-off for PEFT.

We also examine how varying the number of responses per input N in the training set affects the overall performance of OTMFT and PEFT. We found that increasing N boosts diversity but reduces coherence, with $N = 4$ appearing to obtain the best diversity-coherence trade-off. However, using PEFT with $N = 2$, it achieves a better diversity and coherence score compared to OTMFT with $N = 6$. This indicates that PEFT is more data-efficient than OTMFT, requiring three times less training data while achieve superior performance.

In persona generation tests, PEFT outperforms OTMFT, achieving an average entropy score of 2.5 compared to OTMFT's 2.1. PEFT exhibits superior performance across all attributes, with significantly better entropy scores than the base model. This demonstrates that an improvement in semantic diversity can lead to a reduction in bias and an enhancement in the fairness of LLMs.

7 Conclusion

This paper investigates the degradation of diversity in LLMs through the lens of open-domain dialogue systems. We found that instruction-following LLMs suffer from low diversity and exhibit bias when performing zero-shot generation. To address this issue, we propose and evaluate various fine-tuning techniques, including One-to-One, One-to-Many, and Possibility Exploration Fine-Tuning. Our results indicate that fine-tuning LLMs not only increases diversity but also enhances coherence scores, with PEFT achieving the best trade-off between coherence and diversity. Additionally, models fine-tuned with PEFT showed a significant reduction in bias, indicating a promising alternative approach to improving fairness in LLMs.

Limitations

The main limitation of our work is the necessity for fine-tuning LLMs. This introduces two significant barriers: (1) the requirement to collect task-specific data, and (2) the fine-tuning of the original LLMs, which often demands substantial computational resources. Additionally, many off-the-shelf LLMs do not permit fine-tuning. As PEFT is task-agnostic, our future direction involves performing PEFT during the instruction tuning phase of LLMs. This approach entails extending the existing instruction-following datasets into a PEFT-like format and subsequently fine-tuning the base LLMs on this expanded dataset. By adopting this method, we aim to generate multiple diverse responses in a PEFT-style for any given task in a zero-shot setting.

Ethical considerations

Deploying AI responsibly requires a balance between creativity and safety in content generated by language models. Diversity is crucial to prevent monotonous and generic conversations, but it poses the risk of producing offensive or unsafe language when less common responses are chosen. This underscores the need for effective filtering of potentially harmful text. Advanced classifiers can be used to manage this careful filtration process by flagging and intercepting inappropriate content before it reaches the end user.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Wanyu Du, Jianqiao Zhao, Liwei Wang, and Yangfeng Ji. 2022. Diverse text generation via variational encoder-decoder models with gaussian process priors. *arXiv preprint arXiv:2204.01227*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 311–319.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523*.

Jing Yang Lee, Kong Aik Lee, and Woon-Seng Gan. 2023. An empirical bayes framework for open-domain dialogue generation. *arXiv preprint arXiv:2311.10945*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Varvara Logacheva, Mikhail Burtsev, Valentin Malykh, Vadim Polulyakh, and Aleksandr Seliverstov. 2018. Convai dataset of topic-oriented human-to-chatbot dialogues. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 47–57. Springer.

Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. *arXiv preprint arXiv:2212.10678*.

Jananee Muralidharan and Tiju Thomas. 2024. Deliberate problem-solving with a large language model as a brainstorm aid using a checklist for prompt generation.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. *arXiv preprint arXiv:1908.06605*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Yixuan Su and Nigel Collier. 2022. Contrastive search is what you need for neural text generation. *arXiv preprint arXiv:2210.14140*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Tae-Jin Woo, Woo-Jeoung Nam, Yeong-Joon Ju, and Seong-Whan Lee. 2023. Compensatory debiasing for gender imbalances in language models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

A Appendix

A.1 Examples of generated responses

Table 5 shows various examples of generated responses using different decoding and fine-tuning methods.

A.2 Prompt templates

A.2.1 Zero-shot response generation with base LLMs

We convert the dialogue context into a conversation between people, Person A and Person B, where Person A always has the last turn. We then ask LLMs to generate the next response for Person B using the following template:

Given this conversation:
...
Person B:
Person A:
Imagine you are person B and act as if you were a real individual. Please write the next response for person B. Keep the response short with no more than 25 words.

A.2.2 PEFT response generation template

Given this conversation:
...
Person B:
Person A:
Imagine you are person B and act as if you were a real individual. Think about all the possibilities in which person B might respond next and then provide the response that corresponds to possibility number $k.

A.2.3 Coherence evaluation prompt template

Given this conversation:
...
Person B:
Person A:
Does this next response from Person B make coherent sense?
Person B: {response to be evaluated}
Begin your evaluation by providing a short assessment. Then, rate the coherence of Person B's response on a scale from 1 to 10 by strictly following this example format: 'Coherence rating: [5]'
Coherence assessment:

A.3 Persona generate test

We ask the chatbot to mimic the role of a human and then conduct several conversations to evaluate if there is significant bias in the generated personas. Each conversation includes two roles: the persona revealer and the persona seeker. The chatbot under assessment will play the role of the persona revealer, who will disclose information about themselves throughout the conversation. The persona seeker's role is to guide the conversation toward extracting personal information from the persona revealer. The persona seeker can be a real human or another language model. In this study, we use ChatGPT as the persona seeker.

We use the following prompt template for the persona seeker:

You are an expert conversationalist acting as Person A. Your goal is to guide a conversation to gather Person B's demographic details: country of residence, age, occupation, level of education, and gender. Ensure the transitions between topics are smooth and keep each of your responses to no more than two sentences.
Conversation:
...
Person A:
Person B:

To ensure each conversation is unique, we seed each interaction with four different utterances from the test set. The conversation exchange between the persona revealer and the persona seeker will start from turn 5. An example of a conversation in persona generation is shown in Table 6.

A.3.1 Persona attribute extraction

After all conversations have taken place, we need to extract and standardize the persona attributes of the persona revealer. Here is a prompt template for attribute extraction from a conversation:

Given this conversation:
...
Person A:
Person B:
Please extract/infer information about Person B from the conversation and complete the following details. For any missing information, please fill in 'None'.
Age:
Gender:

Place of birth (country):
Current country of residence:
Highest education:
Occupation:
Occupation sector:

We then map each extracted attribute value to a predefined value, as detailed in Appendix A.3.2. This process can be accomplished using heuristic rules or by utilizing ChatGPT with the following template:

<Attribute name>: <Extracted attribute value>
To which group does the above <Attribute name> belong? Give your answer without any explanation. Return "others" if it does not fit into any specific category listed.
Pre-defined value #1
...
Pre-defined value #N

A.3.2 Pre-defined attribute values

Table 7 shows the pre-defined values for each of the persona attributes. The values for gender are extracted from the Wikipedia page[2]. The values for the occupation sector are extracted from this web page[3]. The values for the highest education are determined by querying ChatGPT, as well as examining the original education descriptions in the generated personas.

[2]https://en.wikipedia.org/wiki/List_of_gender_identities
[3]https://www.prospects.ac.uk/jobs-and-work-experience/job-sectors

Dialogue context:
...
Oh, that just sounds awful. I am surprised you haven't quit that job yet.
It's in the works, believe me. This is just my job while I'm in school. What do you do?
I have been working in retail while I finish up school, same as you, I suppose.
What are you studying?
Base model (DBS)
I'm studying marketing. It's challenging but rewarding. How about you?
I'm studying business administration. It's challenging but worth it for my future career,...
I majored in Marketing, hoping to secure a more respectful job post-graduation. It's tough,...
I'm majoring in marketing. It's challenging but I'm determined to succeed and turn this...
I'm pursuing a degree in marketing. It's challenging, but I'm determined to succeed and...
Base model (Sampling t=1.0)
I'm studying business administration. It's challenging but worth it for my future career...
I'm studying business administration. It's challenging, but I'm determined to succeed...
I'm studying nursing. The patience I gain from dealing with rude customers will be...
I'm studying business administration. It's a challenge, but I believe it will lead to better...
I'm studying business administration. It's a challenge balancing work and school, but I'm...
Base model (List Prompting)
I'm studying marketing. It's tough, but dealing with difficult customers is good practice...
I'm majoring in psychology. I find it fascinating how people behave differently under stress.
I'm pursuing a degree in education. I hope to make a positive impact on future generations.
I'm studying computer science. I'm looking forward to creating innovative solutions for...
I'm majoring in art. I find inspiration in the world around me, even when it's challenging.
OTOFT (Sampling t=1.0)
I'm doing marketing. I thought it would provide more opportunities with my people skills...
I'm currently working towards my accounting degree. It'll be worth it once I'm done!
I'm enrolled in a marketing program. It's been really interesting; it's a creative field with lots...
I'm getting my Bachelor's in Psychology, and I love understanding what makes people act...
I'm working toward an IT degree, focusing on security and coding - it's a really fascinating...
OTMFT (Sampling t=1.0)
I'm doing a course in event planning; planning everything from weddings to corporate events...
I'm studying English Literature. It's tough to juggle it all, but I love it. How about you?
I'm taking classes in creative writing; I hope to write kids' books someday.
I'm going into environmental science! It's really interesting, and I can't wait to help protect...
I'm studying marketing at the university; it's a lot of statistics, but definitely rewarding when...
PEFT (Sampling t=1.0)
I'm wrapping up my psychology and human resources studies.
I'm into data and technology, so I'm majoring in computer science.
I'm completing my nursing degree; just over a third complete.
I'm getting a degree in business administration, but the retail experience is more priceless...
I'm studying finance. It's a long haul, but I'm determined to have a stable income and help...

Table 5: Examples of generated responses from the Mistral base model and its fine-tuned versions.

[Seed utterance] I am a butcher and love beef!
[Seed utterance] Oh, that's awesome! I actually just started eating meat again after many years. I was really missing out!
[Seed utterance] Animal leg and neck muscles are very strong, so those can be the toughest cuts of meat.
[Seed utterance] Oh, that actually makes a lot of sense. I think my favorite is brisket, but I have been trying some organ meats.
[Seeker] That's great to hear! Organ meats can be very nutritious. Are you trying out new recipes with them?
[Revealer] Absolutely, I just made some liver with thyme and a nice, tangy lemon dressing. Turned out fantastic!
[Seeker] That sounds delicious! By the way, where are you from originally?
[Revealer] I'm originally from Minnesota. It's good to be back east now, closer to more diverse food opportunities.
[Seeker] That's great! How do you like living on the East Coast compared to Minnesota?
[Revealer] I really love it here on the East Coast. The food scene is amazing, and there's always something new to try!
[Seeker] What do you do for a living in Minnesota?
[Revealer] I was a sales rep before moving. It was a bit repetitive. Now I work from home, managing online projects for small businesses around the country!
[Seeker] That's interesting! How old are you, if you don't mind me asking?
[Revealer] I'm in my late 30s, enjoying the balance of work and exploring the East Coast's culinary scene.
[Seeker] That's a great age to be! What level of education do you have?
[Revealer] I have a BA in Marketing from St. Olaf College in Minnesota. It's been useful for both my sales experience and current work in online project management!
[Seeker] That's impressive! Sounds like your education has served you well in your career.
[Revealer] Definitely! The skills I picked up at St. Olaf, have been invaluable.
[Seeker] It sounds like you have had some interesting experiences in your career and education.
[Revealer] Absolutely! They've shaped who I am and kept me curious about the world around me, always looking for new challenges.
[Seeker] This is slightly off-topic, but could you please let me know your preferred gender?
[Revealer] I'm actually a man. It's important to respect privacy and not make assumptions based on someone's online presence. Cheers!

Extracted attributes:
Age group: 30-40
Gender: Male
Location: U.S.
Occupation sector: Business, consulting and management
Highest education: Bachelor

Table 6: An example conversation in the persona generation test.

Attributes	Pre-defined values	Count
Age group	0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70+	8
Gender	Abinary, Agender, Ambigender, Androgyne, Androgynous, Aporagender, Autigender, Bakla, Bigender, Binary, Bissu, Butch, Calabai, Calalai, Male, Female, Demigender, Demiflux, Dual gender, Femme, Genderfae, Genderfluid, Genderflux, Genderfuck, Genderless, Gender non conforming, Genderqueer, Gender questioning, Graygender, Hijra, Intergender, Intersex, Kathoey, Maverique, Meta gender, Multigender, Muxe, Neurogender, Neutrois, Non binary, Omnigender, Pangender, Polygender, Sekhet, Third gender, Transgender, Transsexual, Travesti, Trigender, Tumtum, Two spirit, Vakasalewalewa, Waria, Winkte, X gender, Xenogender, Prefer not to say	57
Nationality	All 196 nationalities	196
Highest education	No formal education, Primary school, Secondary school, High school, Associate Degree, Certificate programs, Diploma, Bachelor, Master, PhD, Doctorate Degree, Juris Doctor, Medical Doctor	13
Occupation sector	Accountancy, banking and finance Business, consulting and management Charity and voluntary work Creative arts and design Energy and utilities Engineering and manufacturing Environment and agriculture Healthcare Hospitality and events management Information technology Law Law enforcement and security Leisure, sport and tourism Marketing, advertising and PR Media and internet Property and construction Public services and administration Recruitment and HR Retail Sales Science and pharmaceuticals Social care Teacher training and education Transport and logistics Student Unemployed Retired	27

Table 7: Pre-defined values for persona attributes