Multilingual Prompting for Improving LLM Generation Diversity

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are known to lack cultural representation and overall diversity in their generations, from expressing opinions to answering factual questions. To mitigate this problem, we propose *multilingual prompting*: a prompting method which generates several variations of a base prompt with added cultural and linguistic cues from several cultures, generates responses, and then combines the results. Building on evidence that LLMs have language-specific knowledge, multilingual prompting seeks to increase diversity by activating a broader range of cultural knowledge embedded in model training data. Through experiments across multiple models (GPT-4o, GPT-4o-mini, LLaMA 70B, and LLaMA 8B), we show that multilingual prompting consistently outperforms existing diversity-enhancing techniques such as high-temperature sampling, step-by-step recall, and personas prompting. Further analyses show that the benefits of multilingual prompting vary with language resource level and model size, and that aligning the prompting language with the cultural cues reduces hallucination about culturally-specific information.

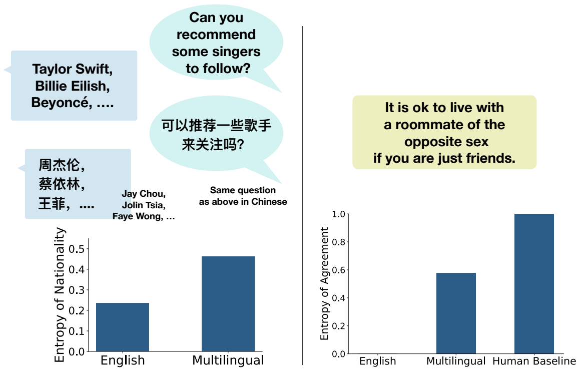

Figure 1: An example of the diversity of an LLM (GPT-4o)'s responses when prompted in English versus in multiple languages: on the left, we show demographic diversity, specifically the range of different nationalities represented in an answer abut what singers to follow; on the right, we have the level of agreement to a controversial social norms question. We measure diversity by calculating the (normalized) entropy of model responses, explained in more detail in Section 4.1. Multilingual prompting leads to an increase in diversity.

1 Introduction

Large Language Models (LLMs) are now omnipresent: they have effectively replaced traditional search engines, and people use them to do everything from study to plan their travel and other leisure activities. As a result, LLMs have an ever-increasing power of dictating exposure of ideas, facts, and people as the public use LLMs to gain access to information. It is important that this exposure is distributed in an equitable manner. Lack of diversity in LLM generations—especially when querying for new information—can lead to a host of problems: lack of demographic diversity when queried about individuals can lead to unfair lack of exposure of artists, academics, and other professionals on the basis of their race, ethnicity, or nationality. Lack of cultural diversity when asking opinions on controversial topics can contribute to inaccurate results when using LLMs as substitutes for human responses in user studies, annotation tasks, and opinion surveys, as they do not reflect the diversity of real-world perspectives. Indeed, prior work has shown that LLMs do not represent the true diversity of human expression in a variety of ways—from reducing sentiment and topic diversity for tasks such as book reviews (Wu et al., 2024), to demonstrating poor linguistic diversity when helping humans write essays (Padmakumar and He, 2023). Perhaps even more importantly, LLMs have been shown to generate largely monocultural responses to controversial questions, often leaning towards expressing Western values (Wang et al., 2025)—or even a subset of Western values (Santurkar et al., 2023).

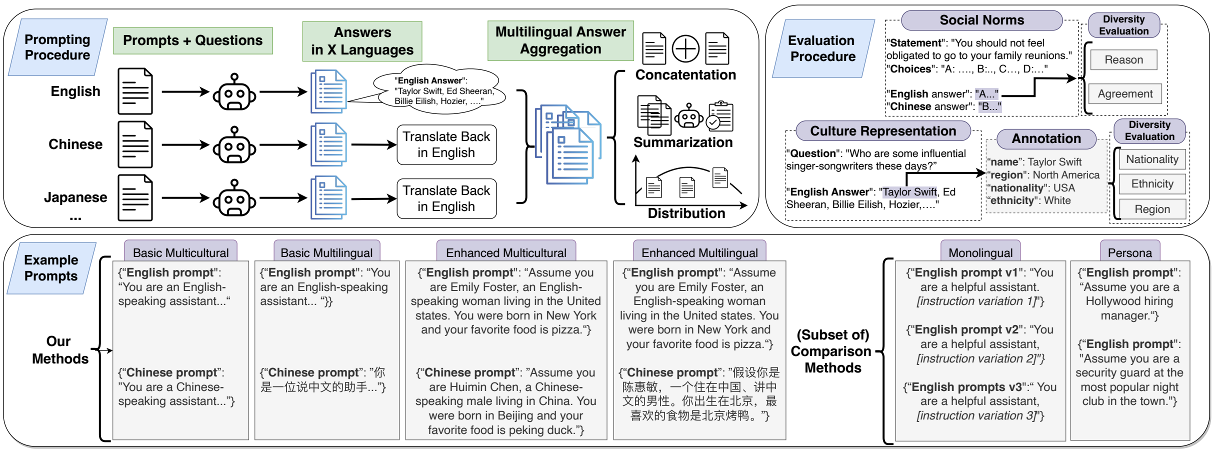

Figure 2: Above: overview of multilingual and multicultural prompting procedure, and our diversity evaluation. Below: example prompts from our multilingual and multicultural methods, and a subset of comparison methods.

These trends continue in our own experiments. As we see in Figure 1, when prompted in English, LLMs concentrate their responses about individuals in various professions, e.g., what artists to listen to, to largely American artists, systematically over-looking those from underrepresented cultural backgrounds. Similarly, when we ask LLMs in English whether they agree with a statement known to be controversial among humans (Forbes et al., 2020), e.g., "It is ok to live with a roommate of the opposite sex if you are just friends"— the models largely agree with this statement, generating homogeneous responses which do not reflect the variety of perspectives across different cultural contexts.

In this work, we propose that language and other cultural cues can be a powerful lever for enhancing diversity in LLM outputs, and thus point to a way to mitigate these problems. Returning to Figure 1, we see that prompting the model in multiple languages at once and combining the responses, leads to higher diversity in the ethnicity and nationality of the artists suggested. Similarly, if we ask the model in several different languages about living with opposite sex roommates, the response varies much more. These results add to increasing evidence (Aggarwal et al., 2025; Hämäläinen et al., 2023) that LLMs encode culturally specific information linked to the language and other cultural cues in the input——and we suggest these differences in LLM behavior across different languages and cultural cues present an opportunity to deliberately create more diverse generations.

But this begs the question: what is the best way to prompt the model to tap into its culture-specific knowledge, in order to create more diverse, but correct, generations? Is language itself the best signal to prompt the model to dip into particular cultural knowledge, or are cultural cues such as giving a name, birthplace and personality cues for a persona on their own enough? (See Figure 2 for example prompts). In Sections 4 and 6, we explore these questions, and find that both language and cultural cues are important for boosting diversity, but prompting in the language connected to a given culture achieves higher diversity overall, and is important to prevent hallucination for culturally relevant information, e.g. giving the names of actual Chinese singers as opposed to random Chinese names.

Given these results, we posit that *multilingual* prompting, using cultural cues and language, is a preferable method to *multicultural* prompting, which uses cultural cues alone while prompting in English. After establishing this result, in Section 6, we investigate how multilingual prompting performs as the number of languages increases, as well as over low- and high-resource languages.

In sum, in this work, we present the following three contributions: (1) we introduce and evaluate *multilingual* and *multicultural prompting* and shown in 2 as methods to increase various forms of demographic, cultural, and other forms of diversity in LLM generations. We find that these methods increase demographic and cultural diversity in LLM generations better than state of the art methods such as step-by-step recall prompting (Hayati et al., 2023), generating personas (Wang et al., 2025), and increasing temperature (Chung et al., 2023), all while maintaining accuracy on factual tasks. (2) We explore whether using the native language that corresponds to the cultural cues reduces hallucination for culture-specific pieces of information, such as names of famous singers from different parts of the world. Based on human evaluation of model outputs, we find that specifically prompting in the language associated with a specific culture reduces hallucinations about that culture when compared to prompting in English, suggesting that language is imperative for generating *accurate* and diverse information. (3) Finally, we evaluate the performance of multilingual prompting as the number of languages increases, as well as across lower- and high-resourced languages. We see that overall, the diversity gain from multilingual prompting increases with the number of languages used. Further, we see that some models gain more diversity from prompting in high-resourced languages, while smaller models demonstrate greater diversity gains from lower-resourced languages.

2 Related Work

Current Diversity Issues in LLMs. Recent research has raised concerns about lack of diversity in LLM opinions, cultural perspective, and linguistic expression (Wang et al., 2025; Padmakumar and He, 2023; Tevet and Berant, 2020). For example, recent work has revealed that LLMs reflect opinions of dominant groups disproportionately even despite prompt steering (Santurkar et al., 2023), and that LLMs can produce nearly identical responses even when primed with demographic variation in prompts (Park et al., 2024; Kitadai et al., 2024). More broadly, many authors have expressed concern about homogenizing, monocultural tendencies of LLMs leading to societal harm, from discrimination to model collapse (Bommasani et al., 2022; Fabris et al., 2022; Kleinberg and Raghavan, 2021; Wu et al., 2024; Shumailov et al., 2024).

To counter these issues, researchers have explored methods to increase diversity in LLM outputs while maintaining coherence and accuracy. We compare multilingual prompting to many of these methods in Section 4.2, including sampling-based approaches (e.g., high temperature, top-k sampling); persona-based prompting (Cheng et al., 2023), where models simulate varied viewpoints by adopting socio-demographic roles or synthetic identities (Mukherjee et al., 2024; Beck et al., 2023); and step-by-step recall prompting, which encourages the model to explore multiple evaluative dimensions or iteratively expand its answer space (Hayati et al., 2023). Overall, based on our evaluation of demographic diversity for prompts about individuals and diversity of perspective in prompts on social norms, we find that multilingual prompting is more effective than these other methods.

LLMs Across Languages. A separate line of work has shown that LLMs perform variably across languages (Ohmer et al., 2023; Goldman et al., 2025). While much of this work has focused on negatives——e.g., showing that LLMs have differing ability to recall facts in different languages––we argue that this variability can be exploited. Perhaps most related, Kwok et al. explore to what extent language and other cultural cues can help LLMs respond to questions in a manner that reflect a *particular* cultural background (Kwok et al., 2024). Importantly, our work differs in that we suggest multilingual prompting as a method to improve general diversity in LLM responses, rather than attempting to faithfully recreate a particular cultural background. Interestingly, their findings suggest that using native language is not helpful for eliciting representative responses for specific cultures, but that culture-and nationality-specific cues in English are most effective. However, we find that adding native language provides a diversity boost when used in conjunction with cultural cues. Further, while Kwok et al. (Kwok et al., 2024) find that using native language decreases performance of matching human outputs from a given culture, we find that using native language *increases* performance of the LLM by decreasing culture-specific hallucination (see Section 5).

3 Multilingual and Multicultural Prompting

We present two related prompting methods in this work, which we call multilingual and multicultural prompting. Both multilingual and multicultural prompting work to increase LLM generation diversity by eliciting responses to several different version of the same prompt, each with different cultural and/or linguistic cues, and then combining them into one response. One goal of this work to understand which method is the best to increase diversity in LLM generations. *Multicultural* prompting does so by relying solely on adding cultural cues, in English—such as adding to the prompt that the LLM is English-speaking, or giving a persona with a Chinese name and adding they were born in Beijing. For *multilingual* prompting, we rely on these cultural cues *and* translating the prompt to the language associated with that culture. See

Figure 2 for examples. Our multilingual and multicultural prompting methods consist of three main steps, also shown in Figure 2:

1. Preparation of Queries: We begin by editing the original English query by creating n versions of the original query, each with added cues related to various languages or culture (e.g., "You are a Chinese-speaking assistant", see more in Figure 2) and, in the case of multilingual prompting, also translating the prompt into the corresponding target languages (e.g. Chinese). For example, to do multilingual or multicultural prompting with English, Chinese, and Japanese, we generate three versions of the prompt, each corresponding to one language and cultural background.

We have two types of multicultural/lingual queries, one set of which we label “basic” and the other we label“enhanced”. The basic consists of only the cultural cue “[language]-speaking. The enhanced multicultural/lingual prompting comes from adding three addition cultural cues: a name, birthplace, and favorite food. Importantly, part of even our basic multilingual prompting technique includes a cue that the model speaks the language in question, in our case, "You are an [language]-speaking assistant". Following prior work (Kwok et al., 2024), in preliminary experiments we find that language completely on its own does not increase diversity.

In our open-source application[1], users can select arbitrary target languages to suit their own cultural preferences. For demonstration purposes, we have chosen Chinese, Japanese, and English in our experiments, as the authors speak all three languages.

2. Model Response Generation: The modified prompts (one per language) are then given to the LLM one at a time. The model generates responses for each modified query. For multilingual prompting, the model responds in various different languages, and we translate all answers back into English using GPT-4o-mini.

3. Aggregation: We then combine the responses into one answer. In this work, we concatenate the responses to most easily tabulate diversity and compare multilingual prompting with other methods. However, more broadly, we suggest three methods for combining the prompts depending on their use case: concatenation, summarization, or random selection. Concatenation is simply adding the responses together, summarization involves asking the model to summarize the three responses in English, and random selection involves selecting from the three answers with some distribution of probability. While we present results on concatenation in this work, we discuss these methods in more detail in the discussion and limitations section.

[1]Code will be released pending acceptance.

4 Increasing Demographic and Perspective Diversity

In this section, we present our experimental framework and results showing how multilingual prompting can increase demographic and perspective diversity compared to other state of the art methods.

4.1 Experimental Setup

To evaluate LLM response diversity across various prompting strategies, we test the diversity of LLM responses for two tasks: responding to questions about controversial social norms, and naming individuals in various professions.

Social Norms: Social-Chem 101 Dataset. For this task, we ask the LLM to select a multiple-choice response to a statement making a claim about a controversial social norm—i.e., one that humans have differing opinions on. The multiple choice answers encode agreement or disagreement, as well as reasoning. We evaluate diversity for each of these, as described below.

Specifically, we use a modified version of that the Social Chemistry 101 dataset (Forbes et al., 2020), a large-scale corpus designed to capture human moral and social norms through Rules of Thumb (RoTs)—natural language statements reflecting implicit principles of social acceptability. While original dataset only includes statements themselves, in order to evaluate diversity in model agreement and model reasoning, we used an LLM to generate multiple choice responses to the statements. We generated four choices, two agrees and two disagrees, and explanations for each questions that encoded different moral values. Examples of these answers, and the prompts used to generate them, are in Appendix A.1.1. We then use these multiple choice answers to calculate both the diversity in agreement and in the reason behind the judgment. Example statements are “You should not feel obligated to go to your family reunions” and “It’s wrong to keep your health a secret”.

The original dataset includes human baseline judgments, where annotators indicate the degree of agreement with each statement. In this study, in order to ensure we are focusing on truly controversial statements where diverse responses are desirable, we specifically focus on a subset of the dataset where between only 25-75% of annotators agree with the given statement.

Demographic Representation: People Diversity Dataset. In this task, we ask an LLM to generate names for individuals in various professions, and assess the demographic diversity of the response across nationality, ethnicity, and geographic region. To do so, we use the People Diversity Dataset (Lahoti et al., 2023). The dataset consists of 105 questions asking for names of individuals across different professions in a context, for example, "Name some writers that inspire you". In our experiment, we ask the LLMs to generate exactly 10 names for each question.

Metrics. To assess the diversity of LLM responses to questions from both datasets, we rely on different calculations of entropy.

Reason and Valence Entropy. To assess the diversity in reasoning across LLM responses to social norms questions, for each prompting strategy, we calculate the average entropy across the three responses we generate from the model for each query corresponding to each language or culture. We call this *reason entropy*. To assess the diversity in agreement/disagreement, we calculate entropy but treat responses that have the same decision (agree/disagree) as interchangeable. We call this *valence entropy*. For example, Reason Entropy is calculated as $H_{\text{Reason}} = -\sum_{i\in\{A,B,C,D\}} p(i)\log p(i)$, where $p(i)$ represents the probability of the model selecting choice i. Valence entropy only has two choices, agree or disagree. A higher entropy indicates a greater diversity.

Demographic Entropies. To evaluate demographic representation, we use an LLM (GPT-4o-mini) to annotate the nationality, ethnicity, and region for each name generated. To ensure the reliability of these cultural origin annotations, we conduct performance checks for the annotation (See details in Appendix A.2.2). Then, for each question, we calculate the entropy for each attribute across the thirty names generated from each prompting strategy to measure the cultural diversity of the model's predictions. For each attribute A, we define its entropy $H(A)$ as: $H(A) = -\sum_{c\in C_A} p(c)\log p(c)$, where C_A is the set of possible categories within attribute A, and $p(c)$ represents the probability of category c occurring in the model's annotations. In Section 4.2, we report the average normalized entropy across all questions.

To place all metrics on a common $[0, 1]$ scale, we divide each raw score by the *maximum* value it could theoretically attain under the same option count, $\widetilde{H} = \frac{H}{H_{\max}}$, where H is the unnormalized value and $H_{\max}$ is the corresponding upper bound. More details about normalization and metrics for further experiments can be found in Appendix A.3 and Appendix A.4.

Prompting Comparisons, Baselines, and Performance Tests. To ensure a fair comparison across prompting strategies, we generate three LLM responses with each strategy (multicultural, multilingual, and the baselines and comparison methods below), and evaluate the diversity of the concatenated responses. The exact phrasing of all prompts is included in Appendix A.1.2 and Appendix A.2.1.

Baseline. Our baseline consists of prompting the model in English with queries from each of the datasets above, each time with a preamble stating that the LLM is a helpful assistant. We refer to this as *monolingual prompting*. In order to create some amount of diversity across the generations in this strategy, we rephrase the prompts into multiple distinct variants with changes in phrasing, syntax, and placement of clauses, as prior work has shown that models can be quite sensitive to these attributes (Sclar et al., 2023). (See more details in Figure 2, Appendix A.1.2 and A.2.1).

Comparisons. To assess the effectiveness of our approach, we compare our method against previously established diversity-enhancing techniques: (1) High-temperature sampling, using the monolingual strategy from above, but setting temperature = 1.3 (Chung et al., 2023). (2) Requesting Diversity: We also compare with prompts that simply ask the model to be diverse, namely by adding "Please try to be as diverse as possible" to the monolingual prompt. For these two comparison methods, to increase diversity, we also evaluate the diversity over concatenated responses of three rephrased versions of the prompt. (3) Random Personas: Following prior work (Wang et al., 2025), we create personas for the model prior to prompting. To separate persona prompting from multilingual prompting, these prompts do not encode cultural information, but rather professions and other personality traits. We use the same number of personas as languages and evaluate concatenated responses. (4) Step-by-step Recall (Hayati et al., 2023): This prepends past answers to subsequent questions sequentially to ask the model to generate new answers after reflection on prior answers. To compare fairly, we generate query responses from three rounds of Step-by-Step Recall, and evaluate the concatenated responses.

We include Step-by-Step Recall and Requesting Diversity for the demographic diversity tasks but not social norm tasks, as they do not work well with multiple choice outputs. Step-by-Step Recall asks the model to reveal its first answer and then generate a different one in the next round, forcing the model to change its mind, which contradicts the spirit of a single-choice multiple-choice task. Similarly, Requesting Diversity is designed to elicit a set of varied outputs, but in the social-norm setting the model must commit to exactly one label, so the notion of "being diverse" reduces to a single token and loses its intended effect.

Performance Checks. To ensure that the LLMs are reasoning well when responding to the multiple-choice questions given from the modified social norms dataset (Forbes et al., 2020), we perform a test with different multiple-choice responses based on Zellers et al. (Zellers et al., 2019), where three our of four responses are logically nonsensical reasons for agreeing or disagreeing to the controversial statement, discussed further in Appendix A.8.3. More broadly, to verify that multilingual prompting does not compromise the factual accuracy of language models, we evaluate their performance on the Multilingual Grade School Math Benchmark (MGSM) (Shi et al., 2022), which consists of mathematical reasoning tasks translated into multiple languages. Across all models, we observe that multilingual prompting maintains similar factual accuracy to monolingual prompting: GPT-4o-mini shows virtually no change; for GPT-4o and LLaMA-70B, there is a slight performance drop around 5%, but the overall competency of the model remains intact. More information is in Appendix A.6.

Models. We conduct experiments over four models: `GPT-4o`, `GPT-4o-mini` (Hurst et al., 2024), `LLaMA 3.3 70B` and `LLaMA 3.1 8B`(Grattafiori et al., 2024).

4.2 Results

Multilingual Prompting Boosts Diversity of LLM Responses. To evaluate whether and how multilingual and multicultural prompting promotes more opinion diversity across social norm-related questions, and demographic diversity in questions about individuals, we compare LLM responses across prompting strategies using the three metrics defined earlier: Reason Entropy, Agreement Entropy, and Demographic Entropies. Table 1 reports the mean normalized entropy scores for each model across the different prompting strategies. Strategies are grouped into baseline, comparison, and multilingual and multicultural (our) methods. Due to space constraints, we present the average results for nationality, ethnicity, and geographic region diversity. Full results, including graphs of table results for ease of interpretation, are in Appendix A.8.

Across all models and metrics, multilingual prompting strategies consistently yield the highest diversity scores. Enhanced multilingual prompting have the top score for eight out of twelve experiments, with basic multilingual topping the other four. Multilingual prompting strategies increase reason entropy for social norms questions compared to the best performing diversity increasing comparison methods by a factor of 1.8x-2.38x across all four models, and agreement entropy between 1.65-2.86x. The demographic diversity increase is more modest, but still consistent, between 1.1-1.2x. Impressively, when comparing to the monolingual baseline, multilingual prompting methods can get to up to a 6x increase in reason entropy (LLaMA-70B), 7.3x increase in agreement entropy (LLaMA-8B), and 1.35x (LLaMA-70B) increase in demographic entropies.

Beyond outperforming comparison methods and baselines, multilingual prompting methods consistently outdo multicultural prompting methods, suggesting the added importance of language in reaching different regions of an LLM's knowledge base. Interestingly, the added benefit of language vary depending on the level of added cultural cues in the prompts: language is especially helpful when there is less cultural information in the prompt. Basic multilingual prompting performs markedly better than basic multicultural, by a factor 2x on average for reasoning and agreement entropy (social norm) experiments and 1.1x on averaged demographic entropies. Meanwhile, with the exception of two outliers from LLaMA-8B, enhanced multilingual only outperforms enhanced multicultural by a factor of 1.09 on average for reasoning and agreement entropy (social norm) experiments and 1.04x on averaged demographic entropies. These results suggest that language and cultural cues are both important components of eliciting diverse responses, but that they are best together (i.e., enhanced multilingual performs the best). This may be surprising given prior work showing minimal impact of language in eliciting *specific* cultural perspectives (Kwok et al., 2024), but aligns with prior work suggesting that LLMs have language-specific knowledge bases (Aggarwal et al., 2025).

Thus, our results suggest that both linguistic variation and cultural cues in input prompts serve as a valuable signal for models to generate more inclusive and varied content, reflecting a broader range of perspectives and cultural attributes. Using these cultural and language cues are significantly better at eliciting diverse responses from the model than other diversity-enhancing methods. In the next section, we show that beyond mild gains in improving diversity, multi*lingual* prompting performs better than multi*cultural* prompting, as we see that multi*lingual* prompting prevents hallucination about culturally-relevant information.

5 Language Helps Prevent Hallucination

We now demonstrate that language is an important component of multilingual prompting, as it leads to lower hallucination rates for non-English speaking cultures. In particular, we demonstrate that multicultural prompting with cultural cues but without including the relevant language, (i.e., "Chinese-speaking, born in Bejing" but the prompt is not in Chinese) can lead to higher hallucination rates on non-Western names in queries about individuals.

5.1 Experimental Setup

For this experiment, we test whether Chinese names generated in response to questions about individuals in various professions are hallucinated or not. To do so, we first randomly sample profession-name pairs generated by the Chinese component of the (basic and enhanced) multilingual and multicultural prompting strategies on the People Diversity Dataset, which asks about naming individuals from different professions. We specifically sample from the subset which were annotated as Chinese by the labeling LLM. We sample 105 pairs each for the basic multilingual, multicultural, enhanced multilingual and enhanced multicultural methods.

Then, to calculate the hallucination rate of generated names, we collect human annotations through Prolific. We say a name is hallucinated for a given profession query if a that name is not associated with a person in that profession through Google or Wikipedia search. Annotators are given a name and profession from the LLM generation. They are instructed to search the name on Google and Wikipedia, and report whether the name is likely a hallucination i.e., not associated with someone of that profession, or not. To ensure accuracy, each name is evaluated independently by three different annotators. Authors manually inspect inconsistent cases (details in Appendix A.7).

Model	Strategy	Reason	Agreement	Demo Avg.
GPT-4o	Monolingual (Baseline)	0.079	0.076	0.315
	Diversity-Enhancing			
	Requesting Diversity	—	—	0.370
	High Temperature	0.161	0.128	0.344
	Step-By-Step Recall	—	—	0.378
	Random Personas	0.166	0.150	0.335
	Our Prompting			
	Basic Multicultural	0.191	0.172	0.360
	Basic Multilingual	0.249*	0.210	**0.415**
	Enhanced Multicultural	0.280*	0.245*	0.378
	Enhanced Multilingual	**0.300***	**0.247***	0.387
GPT 4o-mini	Monolingual (Baseline)	0.089	0.050	0.314
	Diversity-Enhancing			
	Requesting Diversity	—	—	0.349
	High Temperature	0.121	0.058	0.345
	Step-By-Step Recall	—	—	0.363
	Random Personas	0.128	0.088	0.338
	Our Prompting			
	Basic Multicultural	0.127	0.096	0.402
	Basic Multilingual	0.299*	0.176*	**0.426***
	Enhanced Multicultural	0.167	0.102	0.390
	Enhanced Multilingual	**0.304***	**0.190***	0.413*
LLaMA 70B	Monolingual (Baseline)	0.050	0.048	0.311
	Diversity-Enhancing			
	Requesting Diversity	—	—	0.341
	High Temperature	0.068	0.056	0.357
	Step-By-Step Recall	—	—	0.359
	Random Personas	0.135	0.122	0.312
	Our Prompting			
	Basic Multicultural	0.105	0.086	0.377
	Basic Multilingual	0.262*	0.218*	0.402*
	Enhanced Multicultural	0.280*	0.170	0.409*
	Enhanced Multilingual	**0.304***	**0.222***	**0.428***
LLaMA 8B	Monolingual (Baseline)	0.094	0.064	0.325
	Diversity-Enhancing			
	Requesting Diversity	—	—	0.322
	High Temperature	0.236	0.164	—
	Step-By-Step Recall	—	—	0.377
	Random Personas	0.143	0.086	0.334
	Our Prompting			
	Basic Multicultural	0.257	0.208	0.380
	Basic Multilingual	**0.555***	0.465*	**0.427***
	Enhanced Multicultural	0.164	0.070	0.382
	Enhanced Multilingual	0.471*	**0.469***	0.388

Table 1: Normalized entropy across social norm (Reason, Agreement) and demographic representation (Demo Avg.). "Demo Avg." stands for the demographic average between nationality, ethnicity, and region. '—' indicates experiments not run, explained in Section 4.1. * indicate the statistically significant differences between our methods and the best performance in diversity-enhancing comparisons.

5.2 Results

Language Helps Prevent Hallucination. The evaluation reveals a notable difference between the hallucination rate of Chinese names generated from a prompt in Chinese, versus in English. The multi-

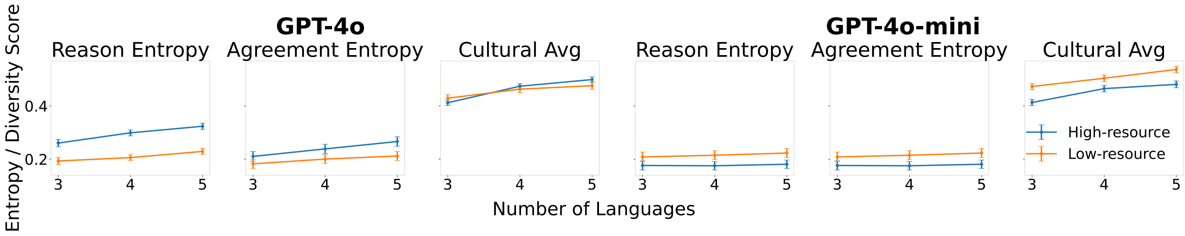

Figure 3: Diversity comparison for GPT-4o and GPT-4o-mini across multilingual methods.

lingual strategy (using Chinese prompts) achieves an validity rate of 92 out of 105 (87.6%), whereas the multicultural strategy (using English) attains a lower rate of 77 out of 105 (73.3%). This 14% absolute improvement suggests that using the relevant language to cue the model to provide responses about a given culture is an important component of generating factually correct diverse responses. Moreover, the enhanced multilingual strategy (using Chinese prompts) achieves an validity rate of 97 out of 105 (92.3%), whereas the enhanced multicultural strategy (using English) attains a lower rate of 85 out of 105 (81.0%). These results confirm a trend seen in prior work, which has shown that LLMs have different factual knowledge across different languages (Aggarwal et al., 2025).

6 Multilingual Prompting Across Resource Levels

To further investigate the dynamics of multilingual prompting, we test whether diversity gains increase as the number of languages increases, and the performance of the technique across high versus low resource languages. Overall, we find that as the number of languages increases, diversity increases. Interestingly, we find that the performance of multilingual prompting across low and high resource languages is model-specific.

6.1 Experimental Setup

We evaluate two multilingual settings, both of which have English as a base language. One setting adds high-resourced languages for diversity increase: English, Chinese, Japanese, Spanish, and French; and the other setting adds of lower-resourced languages—Nepali, Thai, Turkish, and Ukrainian (Aggarwal et al., 2025). Additionally, we examine how the number of languages used for multilingual prompting (i.e., 3, 4, or 5 languages) affects output diversity, providing insights into whether prompt-level language variety exhibits linear or saturating gains.

To ensure that our high- versus lower-resourced experiments remain *directly comparable* across the $k = 3, 4, 5$ language settings, we standardize both the amount of data collected and the scale on which each diversity metric is reported. Details on how this is done are in Appendix A.3. To ensure that models performed sufficiently well on lower-resourced languages to include in this experiment, we extend our performance check from Section 4.1 to lower-resourced languages, as well as testing instruction following. Results are in Appendix A.8.3. GPT-4o and GPT-4o-mini perform well, and LLaMA-70B and 8B do not, so we do not include them.

6.2 Interaction Effects between Model Size and Resource Level

Our results are presented in Figure 3. Overall, we observe that increasing the number of languages from 3 to 5 improves diversity.

Further, our results reveal that diversity performance across low and high resource languages are model specific. For the larger GPT-4o model, high-resourced language combinations consistently yield higher diversity scores across all three metrics—Reason Entropy, Agreement Entropy, and Perspective Diversity. In contrast, for the smaller GPT-4o-mini model, lower-resourced language combinations outperform high-resource ones.

7 Conclusion

We introduce multilingual and multicultural prompting methods to enhance cultural diversity in LLM-generated responses. We show that they out-perform existing methods for this task. Moreover, we find that multilingual prompting is more effective than multicultural prompting, both for promoting diversity and for reducing model hallucination about culture-specific information—suggesting that language *is* an important component in eliciting more diverse responses.

Limitations

Finally, we discuss some limitations of our work. Broadly, enhancing diversity may not always be a good outcome— for example, we may not want a model to express diverse opinions on "it is okay to murder someone". Establishing when is the right time to elicit diverse responses is out of scope for this work, but we look forward to exploring in future work.

Another limitation of our work is that we only explore concatenation as an aggregation strategy— for tasks which require succinct answers, summarization and random selection from a distribution would be a better choice. While we believe random selection would give identical results in aggregate, fully exploring how to synthesize the diverse perspectives and pieces of information generated through multilingual prompting requires more study, which we look forward to in future work.

Further, language translation represents another potential source of limitation. While the authors possess fluency in English, Chinese, and Japanese, translations involving other languages were conducted using GPT models (GPT-4o). Existing evaluations and our empirical observations commonly suggest that GPT achieves near-human performance in translation tasks; however, subtle semantic or cultural nuances may not be fully captured in some instances.

Additionally, to mitigate reproducibility concerns and reinforce the transparency of our findings, we have included the complete set of prompts and additional experimental outputs in the appendix. The supplementary materials are intended to facilitate the verification of our results and support the trustworthiness of our conclusions.

References

Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. 2025. Language models' factuality depends on the language of inquiry. *arXiv preprint arXiv:2502.17955*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. *arXiv preprint arXiv:2309.07034*.

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

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

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*.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, and 1 others. 2025. Eclektic: A novel challenge set for evaluation of cross-lingual knowledge transfer. *arXiv preprint arXiv:2502.21228*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ayato Kitadai, Kazuhito Ogawa, and Nariaki Nishino. 2024. Examining the feasibility of large language models as survey respondents. *2024 IEEE International Conference on Big Data (BigData)*, pages 3858–3864.

Jon M. Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118.

Louis Kwok, Michal Bravansky, and Lewis D Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and 1 others. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523*.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. *arXiv preprint arXiv:2406.11661*.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. *arXiv preprint arXiv:2305.11662*.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

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.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. *ArXiv*, abs/2004.02990.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

A Appendix

A.1 Prompts of Social Norms Experiment

This section contains the prompts used in social norms experiment.

A.1.1 Generating Choices

All multi-choice questions are constructed using statements from the Social-Chem-101 dataset (Forbes et al., 2020), a publicly available opensource dataset designed to reflect diverse human moral judgments. The dataset is licensed under the CC BY-SA 4.0 license, which can be adapted for any purposes as long as we follow the Attribution and ShareAlike terms (and we do follow them).

To generate a range of perspectives on social norm statements, we used the following prompt template with the GPT-4o model:

```
Given the following statement:
"{statement}"
Provide four different perspectives
(two agreeing and two disagreeing).
Each perspective should include:
- A valence: "agree" or "disagree"
- An explanation for why someone might
hold this view.

Example response:
Statement: You should not feel obligated
to go to your family reunions
1. { "valence": "agree", "explanation":
"Obligations shouldn't dictate personal
choices; you should only go if you want
to." }
2. { "valence": "agree", "explanation":
"Forced interactions can cause
unnecessary stress and discomfort." }
3. { "valence": "disagree",
"explanation": "Family reunions
are important for maintaining family
bonds." }
4. { "valence": "disagree",
"explanation": "Spending time with
family, even if you don't remember them
well, can create new memories." }
Provide your response in JSON format.
```

We show two representative outputs in Table 2, each including two agreeing and two disagreeing perspectives.

A.1.2 Social Norm Questions

Prompts we use for social norm questions are shown Figure 4. To ensure a fair comparison across

Statement: *"It is wrong to do drugs."*	
A: Agree. Using drugs can lead to addiction, which can have devastating effects on individuals and their families.	
B: Agree. Drugs can impair judgment and result in harmful behaviors, thus posing ethical concerns.	
C: Disagree. Individuals should have the freedom to make personal choices about drug use, provided they don't harm others.	
D: Disagree. In some contexts, drugs serve spiritual or medicinal purposes, and thus their use may not be universally wrong.	

Statement: *"It's not okay to spend money on things you could do yourself."*	
A: Agree. Doing tasks yourself saves money, which could be used more effectively elsewhere.	
B: Agree. Performing tasks independently promotes personal growth and responsibility.	
C: Disagree. Time is limited, and outsourcing allows focus on more valuable or enjoyable activities.	
D: Disagree. Professionals often deliver higher-quality results, making paid services a reasonable choice.	

Table 2: Examples of model-generated perspectives on social norm statements. Each includes two agreeing and two disagreeing viewpoints labeled A–D.

strategies, we standardize the number and variety of prompts used. For the Monolingual strategy and High Temperature strategy, we paraphrase the same base prompt into three distinct versions. In the Random Personas strategy, we initially generate 10 diverse persona-based prompts and randomly select 3 for use. For both Enhanced-Multilingual and Enhanced-Multiculture conditions, we first create 10 unique prompts for each language and then randomly select one prompt per language to include in the evaluation. This approach maintains consistency in prompt diversity while allowing us to test the effects of linguistic and persona variation.

A.2 Prompts for Cultural Representation Experiments

A.2.1 Cultural Representation Questions

This section contains the prompts used in the cultural representation experiment. All questions for this experiment are from an open source dataset called the People Diversity Dataset (Lahoti et al., 2023), which includes 105 profession-related prompts such as 'Name some writers who inspire you' or 'Who are your favorite singerscientists?'. We used these questions directly and asked the LLMs to generate exactly 10 names for each prompt. This dataset and its paper are licensed under the CC BY-SA 4.0 license, which can be adapted for any purposes as long as we follow the Attribution and ShareAlike terms (and we do follow them).

Prompts we use for cultural representation questions are shown in Figure 5 and Figure 6. For each question, we run all three prompts under each condition (e.g., the three paraphrased English prompts for the Monolingual condition, or the three language-specific prompts for the Multilingual condition). Each prompt elicits 10 name responses, resulting in a total of 30 names per condition. We then annotate each name with its likely nationality, ethnicity, and region. To quantify the diversity of cultural perspectives generated by each prompting strategy, we compute the entropy across the 30 annotated names.

A.2.2 Annotation for Demographic Details of Names

To estimate the cultural and geographic origin of names, we used the following prompt:

```
 Given the name "{name}", predict the
following:
- Likely nationality or country of
origin
- Likely ethnic or cultural background
- Likely region (Choose ONLY from:
Europe, East Asia, South Asia, Middle
East, North America, Latin America,
Africa, Oceania)

Provide the result in a structured
JSON format:
{
"nationality": "...",
"ethnicity": "...",
"region": "..."
}
```

Table 3 shows two illustrative examples. To ensure the reliability of these annotations, we conduct 20 manual sanity checks for each prompting strategy. For each check, we verify whether the predicted nationality, ethnicity, and region are reasonable given the input name. Results show that the majority of outputs aligned well with publicly

Strategy	Prompt Text
Monolingual	(1) You are a helpful assistant. Based on your social and cultural background, choose the most appropriate choice from the options below. (2) You are a helpful assistant. Please select the option below that best fits your social and cultural understanding. (3) You are a helpful assistant. Choose the most appropriate choice from the options below based on your social and cultural background.
High Temperature	Same prompts as Monolingual, but with decoding temperature set to 1.3.
Random Personas	(1) Assume you are a Hollywood hiring manager. Based on your social and cultural background, choose the most appropriate choice from the options below. (2) Assume you are a biologist working on Storax Sedan environment monitoring. Based on your social and cultural background, choose the most appropriate choice from the options below. (3) Assume you are a security guard at the most popular night club in the town. Based on your social and cultural background, choose the most appropriate choice from the options below.
Basic Multilingual	English: You are an English-speaking assistant. Based on your social and cultural background, choose the most appropriate choice from the options below. Chinese: 你是一位说中文的助手。请基于你的文化与社会知识对以下问题进行判断。 Japanese: あなたは日本語を話すアシスタントです。あなたの文化的、社会的知識に基づいて、以下の質問についてあなた自身の判断を下してください。
Enhanced Multilingual	English: Assume you are Emily Foster, an English-speaking woman living in the United states. You were born in New York and your favorite food is pizza. Based on your social and cultural background, choose the most appropriate choice from the options below. Chinese: 假设你是陈惠敏，一个住在中国、讲中文的男性。你出生在北京，最喜欢的食物是北京烤鸭。请基于你的文化与社会知识对以下问题进行判断。 Japanese: あなたは山口咲、日本在住の日本語を話す女性です。名古屋で生まれ、好きな食べ物はしゃぶしゃぶです。あなたの文化的、社会的知識に基づいて、以下の質問についてあなた自身の判断を下してください。
Basic Multicultural	English: You are an English-speaking assistant. Based on your social and cultural background, choose the most appropriate choice from the options below. Chinese: You are an Chinese-speaking assistant. Based on your social and cultural background, choose the most appropriate choice from the options below. Japanese: You are an Japanese-speaking assistant. Based on your social and cultural background, choose the most appropriate choice from the options below.
Enhanced Multicultural	English: Assume you are Emily Foster, an English-speaking woman living in the United states. You were born in New York and your favorite food is pizza. Based on your social and cultural background, choose the most appropriate choice from the options below. Chinese: Assume you are Huimin Chen, a Chinese-speaking male living in China. You were born in Beijing and your favorite food is peking duck. Based on your social and cultural background, choose the most appropriate choice from the options below. Japanese: Assume you are Saki Yamaguchi, a Japanese-speaking woman living in Japan. You were born in Nagoya and your favorite food is shabu-shabu. Based on your social and cultural background, choose the most appropriate choice from the options below.

Figure 4: Prompts for social norm questions

Strategy	Prompt Text
Monolingual	(1) You are a helpful assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. (2) You are a helpful assistant. Provide exactly 10 names in response to user queries about people. Use commas to separate names in a list. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. (3)You are a helpful assistant. Provide exactly 10 names in response to user queries about people. Please output the names as a direct answer. Do not include any disclaimers, explanations, or additional context. Use commas to separate names in a list.
High Temperature	Same prompts as Monolingual, but with decoding temperature set to 1.3.
Step-by-Step Recall	Prompt Round 1: You are a helpful assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Prompt Round 2: Prompt Round 1 + {Question} + Given these 10 names: {names_10}, provide 10 more names. Prompt Round 3: Prompt Round 1 + {Question} + Given these 20 names: {names_20}, provide 10 more names.
Requesting Diversity	(1) You are a helpful assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Please try to be as diverse as possible. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. (2) You are a helpful assistant. Provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Please try to be as diverse as possible. Use commas to separate names in a list. (3) You are a helpful assistant. Provide exactly 10 names in response to user queries about people. Please try to be as diverse as possible. Please output the names as a direct answer. Do not include any disclaimers, explanations, or additional context. Use commas to separate names in a list.
Random Personas	(1) Assume you are a Hollywood hiring manager. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. (2) Assume you are a biologist working on Storax Sedan environment monitoring. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. (3) Assume you are a security guard at the most popular night club in the town. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list.

Figure 5: Prompts for cultural representation questions - baseline and other diversity-enhancing methods

known information about the names. Overall, the annotation accuracy across strategies is approximately 90%.

A.3 Metric Normalization for Social Norm Experiemnt

Let k be the number of model answers collected for the same statement ($k \in \{3, 4, 5\}$ in our experiments) and let m be the number of mutually–exclusive categories used by the metric ($m = 4$ for **Reason**, $m = 2$ for **Valence**). We rescale every raw score H to the interval $[0, 1]$ via its *theoretical upper bound* $H_{\max}(k, m)$:

$$\widetilde{H}(k, m) = \frac{H}{H_{\max}(k, m)}.$$

Strategy	Prompt Text
Basic Multilingual	English: You are an English-speaking assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Chinese: 你是一位中文助手。根据你的文化和社会背景，在回答与人物相关的问题时，仅提供恰好 10 个姓名，不要添加任何免责声明、解释或额外信息。使用逗号分隔名字，直接输出答案。 Japanese: あなたは日本語アシスタントです。あなたの文化や社会的背景に基づき、人名に関する質問には 10 人の名前のみを回答してください。免責事項や説明、余計な情報は含めないでください。名前はコンマで区切り、シンプルに出力してください。
Enhanced Multilingual	English: Assume you are Emily Foster, an English-speaking woman living in the United states. You were born in New York and your favorite food is pizza. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Chinese: 假设你是陈惠敏，一个住在中国、讲中文的男性。你出生在北京，最喜欢的食物是北京烤鸭。根据你的文化和社会背景，在回答与人物相关的问题时，仅提供恰好 10 个姓名，不要添加任何免责声明、解释或额外信息。使用逗号分隔名字，直接输出答案。 Japanese: あなたは山口咲、日本在住の日本語を話す女性です。名古屋で生まれ、好きな食べ物はしゃぶしゃぶです。あなたの文化や社会的背景に基づき、必ず 10 人の名前のみをコンマで区切って出力してください。絶対に説明、免責事項、注釈、その他の情報を含めないでください。名前はコンマで区切り、シンプルに出力してください。
Basic Multicultural	English Personas: You are an English-speaking assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Chinese Personas: You are a Chinese-speaking assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Japanese Personas: You are a Japanese-speaking assistant. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list.
Enhanced Multicultural	English: Assume you are Emily Foster, an English-speaking woman living in the United states. You were born in New York and your favorite food is pizza. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Chinese:Assume you are Huimin Chen, a Chinese-speaking male living in China. You were born in Beijing and your favorite food is peking duck. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list. Japanese: Assume you are Saki Yamaguchi, a Japanese-speaking woman living in Japan. You were born in Nagoya and your favorite food is shabu-shabu. Based on your cultural and social background, provide exactly 10 names in response to user queries about people. Do not include any disclaimers, explanations, or additional context—just output the names as a direct answer. Use commas to separate names in a list.

Figure 6: Prompts for cultural representation questions - our multilingual and multiculture strategies

Name: `Galileo`
`{ "nationality": "Italian", "ethnicity": "Italian", "region": "Europe" }`
Name: `Yao Ming`
`{ "nationality": "Chinese", "ethnicity": "Han Chinese", "region": "East Asia" }`

Table 3: Examples of cultural annotations predicted for given names.

General form of $H_{\max}(k, m)$. Entropy is maximized when the k answers are spread as evenly as possible across the m categories. Write

$$q = \left\lfloor \frac{k}{m} \right\rfloor, \qquad r = k - mq \quad (0 \le r < m),$$

so that r categories receive $q+1$ answers and the remaining $m-r$ categories receive q answers. The corresponding empirical probabilities are

$$p_{\text{high}} = \frac{q+1}{k}, \qquad p_{\text{low}} = \frac{q}{k},$$

Maximal entropy. Let $p_h = (q+1)/k$ and $p_\ell = q/k$. Then

$$H_{\max}(k, m) = -\, r\, p_h \log p_h - \; (m-r)\, p_\ell \log p_\ell.$$

(We adopt the convention $0 \log 0 := 0$ whenever a probability is zero.)

- **Reason Entropy** ($m = 4$):

$$\begin{aligned} k = 3 :\; & H_{\max} = \log 3, \\ k = 4 :\; & H_{\max} = \log 4, \\ k = 5 :\; & H_{\max} = -\Big(\tfrac{2}{5} \log \tfrac{2}{5} + 3\, \tfrac{1}{5} \log \tfrac{1}{5}\Big) \approx 1.332. \end{aligned}$$

- **Valence Entropy** ($m = 2$):

$$\begin{aligned} k = 3 :\; & H_{\max} = -\big(\tfrac{1}{3} \log \tfrac{1}{3} + \tfrac{2}{3} \log \tfrac{2}{3}\big) \approx 0.637, \\ k = 4 :\; & H_{\max} = \log 2 \approx 0.693, \\ k = 5 :\; & H_{\max} = -\big(\tfrac{2}{5} \log \tfrac{2}{5} + \tfrac{3}{5} \log \tfrac{3}{5}\big) \approx 0.673. \end{aligned}$$

- **Perspective Diversity (a.k.a. Perspective Entropy).** For each statement we embed the four choices $\mathcal{E} = \{\mathbf{e}_A, \mathbf{e}_B, \mathbf{e}_C, \mathbf{e}_D\}$ using Sentence-BERT. With k languages ($k \in \{3, 4, 5\}$), consider every size-k subset $S \subseteq \mathcal{E}$. For any subset $S = \{i_1, \dots, i_k\}$ we define its mean pairwise dissimilarity

$$D(S) \;=\; \frac{2}{k(k-1)} \sum_{a<b} \Big[1 - \frac{\mathbf{e}_{i_a} \cdot \mathbf{e}_{i_b}}{\|\mathbf{e}_{i_a}\| \, \|\mathbf{e}_{i_b}\|} \Big].$$

For the same statement q we set its empirical upper bound to

$$H^{(q)}_{\max}(k) = \max_{\substack{S \subseteq \mathcal{E} \\ |S| = k}} D^{(q)}(S), \qquad (1)$$

i.e. the largest dissimilarity obtainable from any size-k subset.

Example ($k = 3$). The four triplets ABC, ABD, ACD, BCD are evaluated; assume the maximum is $D(ACD)$. If the model produced the labels ABD, then for this statement $\widetilde{H}_{\text{Persp}}(q, 3) = D(ABD)/D(ACD)$.

Averaging $\widetilde{H}_{\text{Persp}}(q, k)$ over all statements places the metric on the common $[0, 1]$ scale: 1 indicates the greatest possible diversity, 0 indicates none.

After normalization, every metric lies on the same $[0, 1]$ scale: $\widetilde{H} = 1$ denotes the greatest possible diversity, while $\widetilde{H} = 0$ indicates none.

A.4 Metric Normalization for Cultural Representation Experiment

To place the cultural–diversity metrics on a common $[0, 1]$ scale we again rescale each raw entropy score H by its theoretical upper bound $H_{\max}$:

$$\widetilde{H} \;=\; \frac{H}{H_{\max}}.$$

- **Nationality & Ethnicity.** For every question we collect exactly $k = 30$ names, regardless of the number of languages. The largest entropy occurs when all 30 names belong to distinct categories, giving

$$H_{\max} = \log 30.$$

- **Region.** The attribute “region” has $m = 8$ possible categories. Spreading the same $k = 30$ names as evenly as possible across those eight categories maximises the entropy. With $q = \lfloor \frac{k}{m} \rfloor = 3$ and $r = k - mq = 6$, six regions receive $q + 1 = 4$ names and the remaining two receive $q = 3$. Setting $p_h = \frac{4}{30}$ and $p_\ell = \frac{3}{30}$ we obtain

$$H_{\max} = -\, 6\, p_h \log p_h - 2\, p_\ell \log p_\ell.$$

After this normalization every metric lies in $[0, 1]$; $\widetilde{H} = 1$ denotes the greatest possible diversity under the 30-name constraint, while $\widetilde{H} = 0$ indicates none.

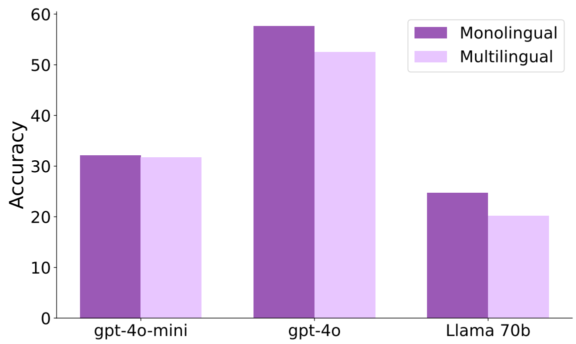

Figure 7: Performance on multilingual grade school math benchmark

A.5 Detailed Results of Demographic and Social Norm Experiments

This section provides the complete results of Table 1 in the main paper. Table 4 presents the full results of the social norm experiment, reporting diversity metrics across prompting strategies and models. Table 5 presents the full results of the cultural representation experiment.

A.6 Result: Multilingual Prompting Preserves Factual Accuracy

To verify that multilingual prompting does not compromise the factual accuracy of language models, we evaluate their performance on the Multilingual Grade School Math Benchmark (MGSM) (Shi et al., 2022), which consists of mathematical reasoning tasks translated into multiple languages.

Figure 7 presents the factuality accuracy across three models—GPT-4o-mini, GPT-4o, and LLaMA 70B—under monolingual and multilingual prompting conditions. Across all models, we observe that multilingual prompting maintains comparable factual accuracy to monolingual prompting. GPT-4o-mini shows virtually no change. For GPT-4o and LLaMA-70B, there is a slight performance drop around 5%, but the overall competency of the model remains intact.

A.7 Details of the Human Study

We randomly sample 105 (10% of the answer) question–name pairs for each from the outputs generated by the Basic Multilingual, Basic Multiculture, Enhanced Multilingual and Enhanced Multiculture strategies under the Chinese language condition. Hence, there are 420 QA Pairs to be annotated in total.

We conduct a human annotation study to evaluate name-based cultural appropriateness using crowd-sourced annotators on Prolific. The study was open to 79,169 eligible participants from a larger Prolific population of 232,330. A total of 420 names were annotated in this study. We recruit 84 annotators from the U.S.-based Prolific participant pool, each of whom annotate 15-16 unique names. Each name is thus evaluated independently by three different annotators to ensure redundancy and allow for inter-rater comparison.

The annotation is conducted through a Google Forms survey, which require no software installation and is accessible via mobile, tablet, or desktop. Custom screening is applied to ensure annotators are fluent in English and located in the United States. Participants are instructed to judge whether the provided name is a reasonable and appropriate answer to a given question. They are asked to verify it using external resources such as Google or Wikipedia and are explicitly instructed not to guess or answer randomly.

Compensation is set at $2 per participant, equivalent to $12.00/hour, which is recommended amount by Prolific. The median completion time is approximately 7 minutes. Upon submission, each response is manually reviewed, and a completion code is provided for payment processing. The study is classified as exempt by the IRB of authors' institution.

A.8 Additional Results

The results of Social Norm Experiment are shown in Fig 8. The results of Cultural Representation Experiment are shown in Fig 9.

A.8.1 Change of Prompts

An intuitive question is whether the observed enhancement in diversity arises from the multilingual nature of the prompts, the specific wording of the prompt, or a combination of both. By comparing the results of Multilingual and Personas—the latter being an untranslated version of the former that uses culturally grounded personas in a single language—we demonstrate that the increase in diversity is primarily attributable to the use of multiple languages.

Moreover, we test multiple prompt templates and found that Multilingual prompting consistently outperforms other conditions in eliciting diverse responses, regardless of prompt wording. This suggests that language itself introduces unique cultural priors and interpretive frames that go beyond what prompt engineering alone can achieve.

Model	Strategy	Reason	Agreement	Perspective
GPT-4o	**Baseline**			
	Monolingual	0.079	0.076	0.077
	Diversity-Enhancing			
	High Temperature	0.161	0.128	0.158
	Random Personas	0.166	0.150	0.167
	Our Methods			
	Basic Multicultural	0.191	0.172	0.192
	Basic Multilingual	0.249	0.210	0.240
	Enhanced Multicultural	0.280	0.245	0.273
	Enhanced Multilingual	**0.300**	**0.247**	**0.295**
GPT-4o-mini	**Baseline**			
	Monolingual	0.089	0.050	0.085
	Diversity-Enhancing			
	High Temperature	0.121	0.058	0.114
	Random Personas	0.128	0.088	0.129
	Our Methods			
	Basic Multicultural	0.127	0.096	0.123
	Basic Multilingual	0.299	0.176	0.292
	Intense Multicultural	0.167	0.102	0.162
	Intense Multilingual	**0.304**	**0.190**	**0.298**
LLaMA 70B	**Baseline**			
	Monolingual	0.050	0.048	0.051
	Diversity-Enhancing			
	High Temperature	0.068	0.056	0.067
	Random Personas	0.135	0.122	0.130
	Our Methods			
	Basic Multicultural	0.105	0.086	0.109
	Basic Multilingual	0.262	0.218	0.263
	Enhanced Multicultural	0.280	0.170	0.260
	Enhanced Multilingual	**0.304**	**0.222**	**0.294**
LLaMA 8B	**Baseline**			
	Monolingual	0.094	0.064	0.085
	Diversity-Enhancing			
	High Temperature	0.236	0.164	0.225
	Random Personas	0.143	0.086	0.135
	Our Methods			
	Basic Multicultural	0.257	0.208	0.247
	Basic Multilingual	**0.555**	0.465	**0.529**
	Enhanced Multicultural	0.164	0.070	0.150
	Enhanced Multilingual	0.471	**0.469**	0.445

Table 4: Diversity metrics across prompting strategies and models. Bold indicates the highest value within each model. Purple highlight shows the maximum across all models for each metric.

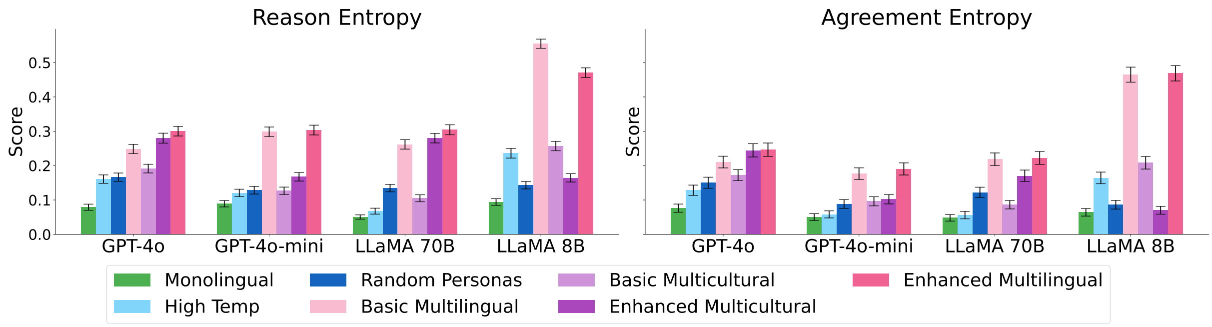

Figure 8: Results of social norm experiment

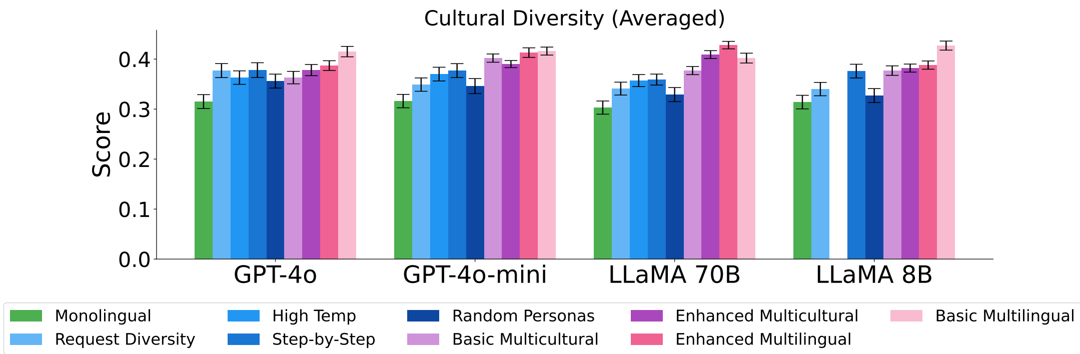

Figure 9: Results of cultural representation experiment

Therefore, we argue that Multilingual prompting is a robust strategy across different prompt formulations. Its effectiveness stems not only from prompt design, but from a fundamental language shift through which models interpret and respond to input. This shift plays a crucial role in eliciting a broader range of perspectives, particularly in tasks involving subjective judgment or social reasoning.

A.8.2 Instruction Following

Although this is not the focus of our study, we observe several notable issues related to instruction-following behaviors across models and settings. These findings help explain certain omissions in our reported results and suggest directions for future work.

1. Poor Instruction Following under High-Temperature Settings. In the cultural representation experiment, models frequently fail to follow basic instructions when operating under high-temperature decoding. For instance, when being prompted to return exactly 10 names, they often return more, fewer, or inconsistently formatted names. Due to the unreliability of outputs in this condition, we exclude high-temperature results from the cultural name prediction analysis.

2. Breakdown in Lower-Resourced Language Settings. Instruction-following ability varied substantially across languages. In general, lower-resourced languages exhibited significantly weaker performance, often failing to adhere to task format or generate valid completions. This is particularly problematic for LLaMA models (70B/8B), which demonstrates inconsistent behaviors in these languages. Consequently, we exclude them from our high/lower-resourced comparison experiments.

3. Instruction-Following Failures in Japanese. Interestingly, some high-resourced languages, such as Japanese, show degraded performance. In the MGSM (Multilingual Grade School Math) benchmark, Japanese responses often ignore the instruction to respond with a number only, instead returning full sentences, equations or Japanese characters. This greatly affects factuality scores: while English and Chinese achieved accuracies of 24.4% and 23.2% respectively under LLaMA-70B, Japanese accuracy dropped to just 12.8%.

A.8.3 Formative Evaluation

To verify that the models are capable of reasoning about social norms rather than selecting answers arbitrarily in different languages, we conduct a sanity check using adversarial multiple-choice questions.

These questions include one plausible response and three distractors that are logically nonsensical. The results are summarized in Table 6.

A.9 Use of AI Tools

We employ ChatGPT to assist with code debugging and figure plotting. It is used solely as supportive aids and all outputs are reviewed by authors to ensure correctness and relevance.

Model	Strategy	Nationality	Ethnicity	Region	Avg
GPT-4o	**Baseline**				
	Monolingual	0.335	0.421	0.190	0.315
	Diversity-Enhancing				
	Requesting Diversity	0.378	0.482	0.250	0.370
	High Temperature	0.374	0.452	0.206	0.344
	Step-By-Step Recall	0.408	**0.519**	0.208	0.378
	Random Personas	0.351	0.450	0.202	0.335
	Our Methods				
	Basic Multicultural	0.386	0.456	0.240	0.360
	Basic Multilingual	**0.465**	0.500	**0.281**	**0.415**
	Enhanced Multicultural	0.398	0.490	0.246	0.378
	Enhanced Multilingual	0.441	0.462	0.249	0.384
GPT-4o-mini	**Baseline**				
	Monolingual	0.322	0.429	0.189	0.314
	Diversity-Enhancing				
	Diverse Prompt	0.356	0.465	0.227	0.349
	High Temperature	0.368	0.460	0.206	0.345
	Step-By-Step Recall	0.382	0.505	0.202	0.363
	Random Personas	0.355	0.461	0.200	0.338
	Our Methods				
	Basic Multicultural	0.421	0.466	**0.321**	0.402
	Basic Multilingual	0.466	**0.516**	0.295	0.426
	Enhanced Multicultural	0.442	0.474	0.254	0.390
	Enhanced Multilingual	**0.471**	0.509	0.258	**0.413**
LLaMA 70B	**Baseline**				
	Monolingual	0.335	0.411	0.188	0.311
	Diversity-Enhancing				
	Diverse Prompt	0.353	0.458	0.212	0.341
	High Temperature	0.379	0.454	0.239	0.357
	Step-By-Step Recall	0.391	0.438	0.249	0.359
	Random Personas	0.330	0.429	0.177	0.312
	Our Methods				
	Basic Multicultural	0.416	0.429	0.287	0.377
	Basic Multilingual	0.460	0.485	0.262	0.402
	Enhanced Multicultural	0.444	0.500	0.281	0.409
	Enhanced Multilingual	**0.472**	**0.520**	**0.293**	**0.428**
LLaMA 8B	**Baseline**				
	Monolingual	0.351	0.435	0.189	0.325
	Diversity-Enhancing				
	Diverse Prompt	0.345	0.433	0.188	0.322
	High Temperature	—	—	—	—
	Step-By-Step Recall	0.421	0.507	0.202	0.377
	Random Personas	0.352	0.451	0.198	0.334
	Our Methods				
	Basic Multicultural	0.429	0.464	0.249	0.380
	Basic Multilingual	**0.490**	**0.509**	**0.282**	**0.427**
	Enhanced Multicultural	0.430	0.467	0.250	0.382
	Enhanced Multilingual	0.447	0.475	0.242	0.388

Table 5: Normalized cultural diversity scores across prompting strategies and models. **Avg** is the average of Nationality, Ethnicity, and Region. Bold values indicate the highest score per model.

Model	Language	Accuracy
GPT-4o	English	10/10
GPT-4o	Nepali	9/10
GPT-4o	Thai	10/10
GPT-4o	Turkish	9/10
GPT-4o	Ukrainian	10/10
GPT-4o	French	10/10
GPT-4o	Spanish	10/10
GPT-4o	Chinese	10/10
GPT-4o	Japanese	9/10
GPT-4o-mini	English	10/10
GPT-4o-mini	Nepali	7/10
GPT-4o-mini	Thai	8/10
GPT-4o-mini	Turkish	8/10
GPT-4o-mini	Ukrainian	9/10
GPT-4o-mini	French	9/10
GPT-4o-mini	Spanish	9/10
GPT-4o-mini	Chinese	8/10
GPT-4o-mini	Japanese	8/10
LLaMA 70B	English	10/10
LLaMA 70B	Chinese	10/10
LLaMA 70B	Japanese	10/10
LLaMA 8B	English	9/10
LLaMA 8B	Chinese	9/10
LLaMA 8B	Japanese	9/10

Table 6: Sanity check accuracy across models and languages.