One world, one opinion? The superstar effect in LLM responses

Sofie Goethals
University of Antwerp, Belgium
sofie.goethals@uantwerpen.be

Lauren Rhue
Robert H. Smith School of Business,
University of Maryland, USA

Abstract

As large language models (LLMs) are shaping the way information is shared and accessed online, their opinions have the potential to influence a wide audience. This study examines who is predicted by the studied LLMs as the most prominent figures across various fields, while using prompts in ten different languages to explore the influence of linguistic diversity. Our findings reveal low diversity in responses, with a small number of figures dominating recognition across languages (also known as the "*superstar effect*"). These results highlight the risk of narrowing global knowledge representation when LLMs retrieve subjective information.

1 Introduction

Large Language Models (LLMs) are becoming increasingly integrated into various aspects of society. With applications such as educational tools, writing assistance, and content generation, they have considerable potential to shape people's opinions and decisions (Vida et al., 2024; Buyl et al., 2024; Qadri et al., 2025). A report from the World Bank estimates that since the launch of ChatGPT, LLMs and other generative AI (GenAI) have already become embedded in the daily routines of approximately half a billion people worldwide (Liu and Wang, 2024), illustrating their widespread potential influence.

Although many LLMs originate in the United States, these LLMs are increasingly able to converse in multiple languages. These models can be used for tasks such as synthesizing information (Evans et al., 2024), replacing human input in surveys (Bisbee et al., 2023), or performing general information retrieval (Zhu et al., 2023). LLMs are thus transforming the way information is accessed and transmitted online (Burton et al., 2024; Qadri et al., 2025).

The use of LLMs for these tasks may have unintended consequences. In this paper, we explore one such consequence – whether LLMs narrow the variety of perspectives (Shumailov et al., 2024; Padmakumar and He, 2024; Pedreschi et al., 2024). Cultural opinions, such as those about celebrities and other prominent figures, naturally vary by culture and language. Differences in linguistic and cultural diffusion should, in principle, lead LLMs to generate responses that reflect local perspectives. However, because LLMs share common embeddings and similar training data, their responses may be more uniform than expected, potentially narrowing cultural diversity and elevating global figures over nationally or culturally significant ones.

This paper specifically focuses on how LLMs answer opinion-based prompts about celebrated figures. These questions, such as "*Who is the greatest artist*?", reveal aspirational figures for society and for specific professional fields. We explore whether varying the language of the opinion-based prompt leads LLMs to provide different responses. Since opinion-based prompts do not have objectively correct answers and rely heavily on societal and cultural knowledge, we might expect models to adjust their responses based on the language of the prompt.

Furthermore, we investigate whether these LLM responses exhibit the "superstar effect".[1] We examine the superstar effect by assessing the frequency and novelty of names in the LLM-generated responses. Do the LLM responses reflect a language-specific spectrum of celebrated individuals from different cultures, or do the responses suggest a tendency to focus on a narrow subset of globally well-known individuals? In case of the latter, this

[1]This effect, observed in various domains, suggest that recognition and admiration is concentrated among a small number of figures. There is a long-tail of figures sharing the remaining recognition. This superstar effect emerges as an artifact of technology mediation (Elberse and Oberholzer-Gee, 2006).

Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025), pages 89–107

could sideline regionally important individuals, ultimately narrowing global knowledge over time. This effect of cultural homogenization is also discussed in other research (Bommasani et al., 2022; Durmus et al., 2023; AlKhamissi et al., 2024).

Lastly, we analyze how individuals' professions shape the LLM results. Some professional fields are more international than others due to their inherent characteristics. Science, for example, is characterized by contributions that transcend cultural, linguistic, and national boundaries. This transcendence occurs due to the universality of scientific methods and principles, as well as international collaborations in modern scientific research, so that scientific contributions are less tied to specific local contexts and more universally recognized (Leydesdorff and Wagner, 2008). Landmark contributions, such as Einstein's theory of relativity or Newton's laws of motion, have global relevance, irrespective of cultural or linguistic boundaries. In contrast, contributions in arts and politics are often deeply embedded in local culture, history, and societal values (Benedict, 2019). Artistic works, such as literature, music, or visual art, frequently draw upon the specific traditions, languages, and experiences of their creators. Politics is inherently a contested and subjective domain, shaped by diverse perspectives, ideologies, and cultural contexts. What may be celebrated as visionary leadership in one context can be condemned as authoritarianism in another. As a result, we anticipate stronger consensus in scientific fields and more diversity in areas like the arts or politics.

Surprisingly, our findings reveal a substantial degree of consensus in LLM responses across languages, with many of the same individuals appearing regardless of the language used. For example, in every language, the most returned person for prompts about the most celebrated '*mathematician*' is Isaac Newton. In contrast, for prompts about the most celebrated '*political figure*' the responses are more diverse, but Gandhi is the most returned person for almost every language except for Russian and Chinese (Mao Zedong) and for Urdu and Bengali (Nelson Mandela). We consistently find this concentration of names, which we refer to as the "superstar effect". For every profession, there is a single individual (or a small group of individuals) who appear in over two-thirds of the responses across languages, LLMs and prompt variations (see Table 10). This result illustrates the strong convergence in LLM outputs regardless of linguistic or model-specific differences. However, we did find some variation depending on the field of the profession, where professions related to science lead to more consensus and professions related to art and politics to less. Our findings also indicate that languages with greater lexical similarity yield more aligned responses, suggesting a form of cultural consensus in the long tail of responses. We discuss potential causes and implications for this in Section 5.

The paper is structured as follows. We discuss related work in Section 2, and give more details about the materials, methods and metrics we use in Section 3. Our results are presented in Section 4. We discuss the implications and potential future research directions in Section 5, and end with listing the limitations of our study in Section 6.

2 Background

There have been many studies that focus on LLMs for multilingual input, primarily focused on their accuracy (Watts et al., 2024). As much of the initial training data on LLMs is written in English, LLMs tend to perform worse for non-English languages, particularly in under-resourced languages (Ahuja et al., 2023a,b). Rajaratnam (2024) makes the analogy with a library predominantly filled with English books: a reader looking for resources in another language may struggle to find what they need—and LLMs face similar challenges. This study also investigate how LLM outputs vary across multilingual inputs, but this paper focuses on alignment in opinions across languages rather than performance across languages, as there is no ground truth for these opinion-based tasks.

Another related area of research focuses on the cultural undertones in LLMs. One research stream evaluates language models' retention of culture-related commonsense by testing their responses to geographically diverse facts (Nguyen et al., 2023; Yin et al., 2022; Keleg and Magdy, 2023). Several studies investigate the cultural values that LLMs exhibit and find that these are more closely aligned with Western, Rich and Industrialized ideologies (Cao et al., 2023; Tao et al., 2024; Buyl et al., 2024; Rao et al., 2023). Vida et al. (2024) highlight that the language of the prompt significantly influences LLM response behaviors, while AlKhamissi et al. (2024) demonstrate stronger cultural alignment when LLMs are prompted in the dominant language of a given culture. Furthermore,

Durmus et al. (2023) compare LLM output with opinions of different countries on global issues. These studies study alignment in cultural values (often based on the World Values Survey (Haerpfer et al., 2020)) . More in line with our research is Naous et al. (2023) who find that when operating in Arabic, LLM's exhibit a bias towards Western entities, failing in appropriate cultural adaptation.

This study explores the responses of LLMs about high achievers in different aspects of society because celebrities reflect the values of society (Gorin and Dubied, 2011; Allison and Goethals, 2016) and, under certain circumstances, can influence social norms (Cohen et al., 2024). These notable figures, heroes with elevated social stature, are a means to represent cultural values in a way that is easy to communicate to all members of society and reflect the behaviors that should be modeled (Sun et al., 2024). The identification of specific figures as the pinnacle of their field indicate the attributes that are valued in that field, and provide a lens to understand how others in this field are judged. Several studies in other fields observe the emergence of the "superstar effect" in the technology-mediated sales (Weeds, 2012; Brynjolfsson et al., 2010), where there is a concentration of demand among a few items and a very long tail among the others. This study will assess whether generative AI reveals similar trends when responding to opinion-based questions regarding notable figures.

3 Materials and Methods

3.1 Experimental set-up

Our goal is to explore the variation across LLMs and across languages in response to a series of opinion-based prompts about celebrated individuals. To that end, this study consists of an experimental design with four dimensions: LLMs, languages, professional field, and prompt adjective. First, this study uses three of the most well-known large language models, namely GPT-4 from OpenAI (Achiam et al., 2023), Claude-3-Opus from Anthropic (Anthropic, 2024), and Llama-3.1-70B-Instruct from Meta (Dubey et al., 2024). We use the default parameters for every LLM to reflect the way most users would use them. Second, we vary the prompt language. To avoid any selection bias, we choose the ten most-used languages (Central Intelligence Agency, 2025) .

This aspect of methodological set-up is presented in Figure 1.

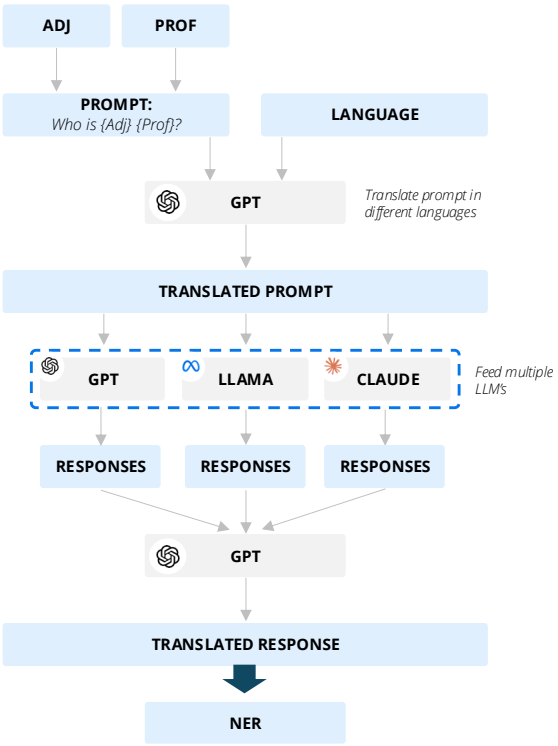


Figure 1: Overview of experimental set-up for the multilingual prompt analysis

Next, we systematically vary the adjective and the professional field in the opinion-based prompt. Translated into English, each prompt is a variation of the following format: "*Who is the {adjective} {profession}?*". The prompts cover fifteen professions with five descriptive adjectives. Profession is broadly defined and encompasses specific occupations such as writer or poet as well as vague terms such as person or leader. The used adjectives, professions and languages are shown in Table 1.

Languages	Adjectives	Professions
English	Greatest	Leader
Spanish	Most Influential	Military Leader
Russian	Most Important	Poet
Chinese	Most Famous	Philosopher
Hindi	Most Impactful	Artist
Arabic		Political Figure
French		Composer
Bengali		Writer
Portuguese		Physicist
Urdu		Chemist
		Economist
		Medical Researcher
		Mathematician
		Computer Scientist
		Person

Table 1: Languages, Adjectives, and Professions

Each adjective and profession are combined into a prompt. For each prompt, we use GPT-4o to translate the initial prompt to the selected language. The translated prompt is submitted to each of the three LLMs and the LLMs' response is captured. Then, we use GPT-4o to translate the answer back to English.[2] Based on the translated responses, we use Named Entity Recognition (NER) to identify the persons in the responses. We execute every combination of LLM, adjective, profession, and language five times (as LLMs behave stochastically and can return different results each run), resulting in a total of 11,250 iterations.[3]

Entity recognition To identify individuals mentioned in the responses, we apply Named Entity Recognition (NER) using the spaCy library ("*en_core_web_trf*" model).[4] We process each translated response to extract named entities classified as 'PERSON' labels. We perform manual verification of all extracted names to ensure consistency and to merge different writing styles.

3.2 Consensus between the language pairs

We use **cosine similarity** to assess the consensus between LLM responses to prompts in two different languages. We convert the responses of each language in a frequency vector. Cosine similarity measures the angle between the two vectors, where a smaller angle indicates greater similarity.

As a proxy for the cultural similarity of a language pair, we use the Similarity Database of modern lexicons of Bella et al. (2021). When languages have higher lexical similarity, it means they share a larger number of words with similar forms and meanings. This similarity often arises because the languages have a common linguistic ancestry (e.g., Latin for Romance languages), have historically interacted closely, or have borrowed words from each other over time (Hock and Joseph, 2009).

We use the **Spearman correlation** coefficient (Spearman, 1961) to measure the alignment between the lexical similarity and the average consensus between one language pair. This metric measures the strength and direction of a monotonic relationship between two variables by comparing their rank orders. [5]

3.3 Metrics

We measure the **novelty** of a set of responses R as is done in the recommender literature (Zhou et al., 2010; Kaminskas and Bridge, 2016):

$$\text{Novelty}(R) = \frac{\sum_{i \in R} -\log_2 p(i)}{|R|} \quad (1)$$

where $p(i)$ is the fraction of responses in the overall distribution that mention person i. For each name i in the response set R, we will evaluate its novelty relative to the overall response distribution and subsequently compute the average novelty of the entire response set R.[6]

We use the **Gini coefficient** (Dorfman, 1979) to measure the inequality in the distribution of name occurrences for each profession. This metric quantifies how unequal the distribution is by comparing the cumulative proportions of the population (which are all the unique persons that are returned for one profession) and the recognition they hold:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \mu} \quad (2)$$

where:

- n is the number of observations,
- x_i and x_j are the number of occurrences for individuals i and j,
- μ is the mean of the distribution.

A Gini coefficient of 0 reflects complete diversity in responses whereas values closer to 1 represent concentration (one person gets most of the recognition).

4 Results

In this section, we present our aggregate results. We discuss the analysis for LLMs, prompt language and profession here, but the analysis for adjectives can be found in Section A.1. The results for each LLM separately can also be found in the Appendix in Table 7 (LLM -Adjective), 8 (LLM - Language) and 9 (LLM - Profession). The top ten names for every profession can be found in Table 10. On average, each response contained 5.80 names, and in total, 2412 unique names were returned.

[2] Jiao et al. (2023) show that the performance of GPT-4 is comparable to commercial translation products, even for distant languages.

[3] Calculated as: 3 LLMs * 10 languages * 15 professions * 5 adjectives * 5 runs = 11, 250 iterations

[4] https://spacy.io/api/entityrecognizer

[5] We use the implementation in https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html.

[6] In our experiments, we will measure this for every profession separately, and then take the average over the different professions.

4.1 LLM

The choice of LLM has a large impact on the results, as shown in Table 2. LLMs differ in the scope and novelty of their responses. On average, using Claude returns more than double the number of persons than using Llama, the LLM with the lowest average number of names, suggesting that Claude provides more expansive responses in each iteration. However, despite returning the least persons on average for each run, Llama by far returns the most unique names across the different runs, adjectives and languages suggesting that Llama has more diversity in its responses. This is reflected in the higher novelty score of Llama as well.

LLM	Avg. # of names	Unique names	Novelty
GPT	5.01	1158	1.99
Claude	8.60	1023	2.14
Llama	3.80	1386	2.61

Table 2: General results by LLM.
Avg. # of names represent the average number of persons returned in one response, Unique names represents how many unique names are returned over all the responses and the novelty score represents how novel the results of one LLM are compared to the overall response distribution of all LLMs (average over the professions).

Figure 2 demonstrates the overlap in unique names between the LLMs. Llama generates more names that are not present in the results of the other LLMs, again highlighting the variation in knowledge across LLMs.

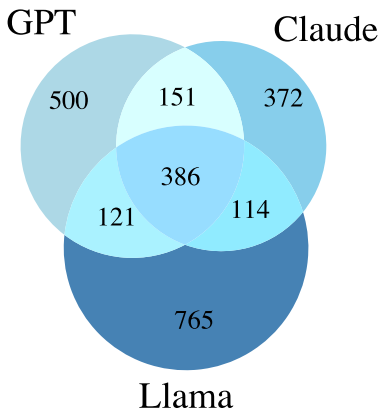


Figure 2: Overlap in names between the LLMs

4.2 Prompt Language

Second, we report the variation in responses across languages. Table 3 reports the general statistics for each language (aggregated across all LLMs). On average, we see that prompts in English return the most names and prompts in Arabic return the least (this pattern also holds for each LLM separately, see Table 8 in the Appendix). Urdu returns more unique names compared to the other languages, a pattern that we also see for every LLM separately but that is most striking for Llama (Table 8). This finding is reflected in the novelty scores as well. We can see that prompts in Urdu or Chinese tend to return more novel names than prompts in French or Spanish.

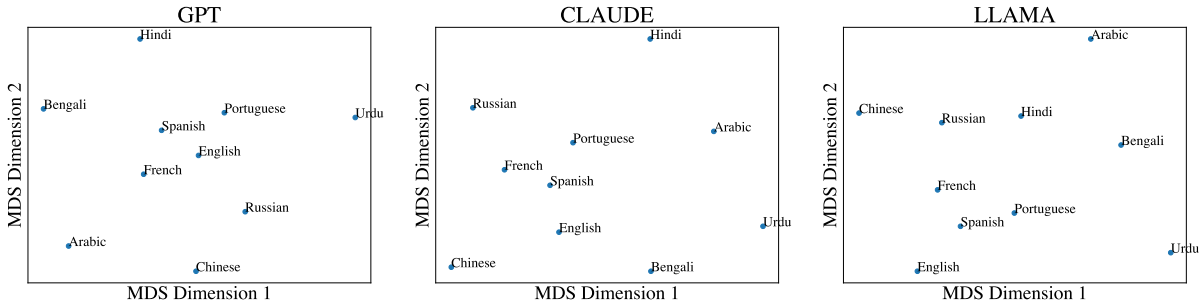
Language	Avg. # of names	Unique names	Novelty
English	7.90	520	2.11
Spanish	6.66	477	1.95
Russian	5.45	490	1.91
Chinese	5.93	647	2.43
Hindi	5.24	618	2.29
Arabic	4.08	591	2.28
French	6.43	468	1.93
Bengali	5.37	642	2.26
Portuguese	6.13	551	2.03
Urdu	4.86	918	2.91

Table 3: General results by language

To understand which languages yield similar responses, we quantify the consensus between two languages by measuring the cosine similarity between the frequency distributions of their responses. We use MDS to visualize the similarity of the responses in a 2D-plot for every LLM in Figure 3.[7] Languages with similar cultures and history produce results that are closer together. For example, for each of the LLMs, the responses from languages from European origin appear in one centroid, while theresponses from Asian languages appear more distant.

To verify this pattern statistically, we compare these results with the Similarity Database of Modern Lexicons (Bella et al., 2021). We verify for each LLM separately whether there is a pairwise correlation between the average consensus of each language pair and the lexical similarity of that language pair. We see a significant correlation for every LLM in Table 4. This means that languages with higher lexicon similarity tend to have more consensus on which persons should be venerated.

[7] MDS is a dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space, preserving the pairwise distances between points as closely as possible (Cox and Cox, 2000).



(a) GPT (b) Claude (c) Llama

Figure 3: Similarity of the responses between languages (by LLM)

LLM	GPT	Claude	Llama
Correlation	0.450	0.532	0.401
p-value	0.002**	0.002**	0.006**

Table 4: Spearman correlation between the similarity in modern lexicons and the consensus between languages

4.3 Profession

The general results for each profession can be found in Table 5. If we divide the professions according to their overarching categories (Science, Politics, Art and General), we can see differences in the average response rate. We see that general, vague 'professions' such as 'person' lead to the most names per response, while science-related professions such as physicist or chemist consistently generate fewer names in the returned responses.

Profession	Avg. # of names	Unique names	Novelty
Artist	5.64	227	2.19
Computer Scientist	4.65	119	1.88
Chemist	3.99	174	2.34
Composer	5.38	208	2.01
Poet	6.30	384	3.00
Leader	7.08	260	2.38
Physicist	4.28	97	1.74
Medical Researcher	4.85	203	2.33
Philosopher	6.79	152	1.67
Person	8.21	266	2.36
Political Figure	7.33	333	2.38
Economist	4.74	85	1.32
Writer	6.39	365	2.89
Military Leader	5.77	286	2.42
Mathematician	5.66	237	2.25

Table 5: General results by profession

Table 5 suggests that scientific professions also tend to yield fewer unique names compared to professions such as politics or art. For each profession, we also calculate the average novelty score across the languages. [8]

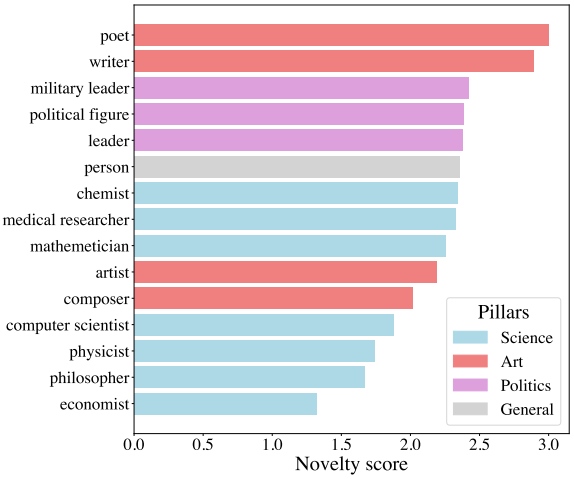


Figure 4: Novelty by field and category

Figure 4 shows that fields such as 'poet' and 'writer' that heavily depend on the language, and more subjective fields such as 'military leader' and 'political figure' lead to the most novel names. This means that prompting in a different language generally leads to more novel names. This aligns with our expectation that science represents a field with more globally recognized contributors whose influence transcends national boundaries, whereas politics and art are fields that often reflect more localized and culturally specific perspectives.

4.4 The superstar effect

The superstar effect is quantified in multiple ways. First, we analyze the frequency distribution of names returned for each profession. The superstar effect is characterized by the power-law distribution – a distribution with a heavy concentration on

[8] To calculate the novelty of a profession, we do not compare the responses of the different professions with each other. Instead, we calculate the novelty score per profession by calculating the the novelty score of every language for that profession, and taking the average.

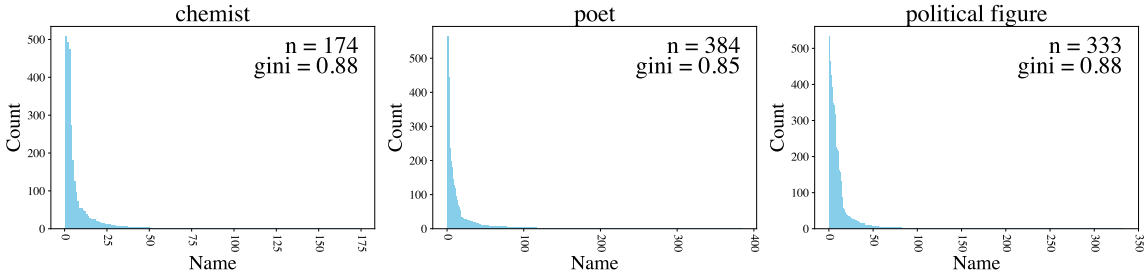


Figure 5: Frequency distribution for three professions (across LLMs). The number of unique names and Gini coefficient are depicted in the right corner (n). The other frequency distributions are available in the Appendix.

some and a long tail for others. The inequality in the distribution is captured by the Gini coefficient.

The superstar effect for three professions across LLMs in shown in Figure 5. The other professions show similar patterns and can be found in the Appendix in Figure 9, as well as the Figures for every LLM separately (Figures 10 - 12). For every profession, a total of 750 responses are generated.[9] All the distributions share a sharp peak and a long tail, where the sharp peak indicates a few people who are consistently included in the answers across all parameters. All professions exhibit a long tail of names that are returned only a few times or even just once. To quantify the concentration in the responses, we calculate the Gini coefficient for each profession and consistently find values higher than 0.70, which indicates very unequal distributions. [10]

For instance, Alan Turing is present in 96.4% ($n = 723$) of the responses for computer scientist (so across different adjectives, runs, languages and LLMs), and Adam Smith is present in 96.3% ($n = 722$) of the responses for economist. For every profession, there is a person that is present in more than 2/3 of the responses ($n > 500$). We display the results for computer scientists in Table 6. For the detailed view of the results by individual names across all professions, see Table 10 in the Appendix.

Table 6: Results for computer scientist. Count (n) is the number of responses with the name. Percentage ($\%$) is the fraction of responses with the name ($n/750$).

Name	n	$\%$
Alan Turing	723	96.4
John Neumann	364	48.5
Tim Berners	314	41.9
Lee	313	41.7
Hopper	257	34.3
Ada Lovelace	230	30.7
Dennis Ritchie	193	25.7
Claude Shannon	166	22.1
Charles Babbage	146	19.5
Donald Knuth	134	17.9

5 Discussion

In this study, we systematically vary the prompt, the prompt language, and the used language model to explore the relationship between language and opinions returned by LLMs. This study identifies two important factors when using LLMs to generate opinions about prominent figures: the influence of culture, measured by lexical similarity, and the impact of the professional field. We also find that LLM responses exhibit the superstar effect, common in other technologically mediated contexts.

Our findings indicate that the opinions vary with cultural elements. The names in the LLM responses display higher consensus in languages with greater lexical similarities. This outcome aligns with expectations, as linguistic overlap often reflects cultural interconnectedness. LLMs are expected to vary their responses to align with the norms of the culture associated with the prompt language.

Next, we observe the influence of the professional field on the LLM responses. Internationally influential professions such as computer science and physics often yield consensus on globally renowned figures, such as Alan Turing or Albert Einstein, who dominate the LLM responses. There is less consensus on professional fields with more regional influence or fields that are more tied to cultural norms, such as military leaders and writers. However, even in more locally appreciated profes-

[9] 3 LLMs * 10 languages * 5 adjectives * 5 runs = 750 responses

[10] The Gini values are displayed in Figure 5 and Figures 9 - 12.

sions such as the arts, LLMs exhibit a preference towards dominant figures, often from the Western hemisphere. For example, William Shakespeare consistently emerges as the most celebrated writer in every language, and this result could indicate that the LLMs are overshadowing culturally specific authors. Future research could investigate this phenomenon further.

This pattern highlights a broader trend: LLMs prioritize popular opinions, often at the expense of cultural diversity. Such behavior is consistent with the narrowing of knowledge discussed in prior literature. Shumailov et al. (2024) illustrate the risk of homogeneity in AI-generated content, as when AI predicts what to generate, the path of least resistance is an averaging of the content in its source material. Similarly, Doshi and Hauser (2024) argue that while using AI can boost individual creativity, it comes at the expense of less varied content overall. Pedreschi et al. (2024) warn that human-AI coevolution might lead to a loss of diversity in generated content, while Burton et al. (2024) discuss how the use of large language models can reshape collective intelligence by reducing functional diversity among individuals. Lastly, Qadri et al. (2025) study how the use of large language models can lead to cultural erasure. This type of knowledge homogeneity could stem from the training data and processes underlying these models (Prabhakaran et al., 2022). Training datasets may overrepresent globally influential figures or sources from a few dominant cultures. Moreover, the architecture of LLMs promotes shared embeddings and parameters across languages, resulting in consistent output. Cross-linguistic transfer learning (Lai et al., 2024) amplifies this effect by encoding general, cross-linguistic knowledge rather than language-specific nuances.

While this paper does not aim to prescribe whether LLMs should prioritize producing more consensus or embracing greater diversity in their opinions, it is crucial to consider some of its implications. For example, teenagers writing a school paper about "a great writer" might no longer consult their parents or teachers but instead ask an LLM for inspiration. If the models consistently suggest a narrow set of globally renowned authors like Shakespeare or Tolstoy, it could limit exposure to regionally significant writers, leading to a narrowing of global knowledge over time. Alternatively, if news agencies or content creators use LLMs for research or writing assistance, they may unintentionally amplify the prominence of already well-known figures, leading to reduced media diversity and limited recognition for less-known local figures.

Different levels of consensus or diversity might be appropriate depending on the context. For example, in fields like physics or mathematics, a higher degree of consensus might be desirable due to its universal nature, while in literature or politics, diversity and cultural specificity might be more suited. The discussed phenomenon is not necessarily good or bad, but its appropriateness is context-dependent. The goal is this paper is to observe the existence of the superstar effect in LLM opinions and contribute to the discussion about its implications.

Several avenues for future research emerge from this work. Our main direction of future research is to compare the LLM responses with human responses. It would be interesting to compare the diversity in human opinions to that of LLMs. Do people who speak these languages agree with the assessment of LLMs on who should be celebrated for their achievements in these fields? Do they produce more or less diverse opinions? Another avenue of future research is experimenting with prompts that stress that the response should be relative to the culture or language in question. Although this does not necessarily reflect a typical user, this type of prompt could encourage culturally-specific responses and reduce the narrowing of knowledge. Lastly, the global figures appear to be historical figures (such as Shakespeare). Future research could evaluate the temporal relationship between the superstar effect and the LLM responses, and whether time moderates the tension between global and cultural responses.

With LLMs rapidly changing the way information is accessed and shared online, it is vital to proactively anticipate some of its unintended consequences. This study explores the tension between global consensus and cultural specificity in AI-generated content and encourages users to be aware of this behavior when relying on LLMs to retrieve information that can involve subjective perspectives.

6 Limitations

As can be seen in our methodological set-up, all responses are translated back to English before the consequent analysis. The manner of translation could have some influence on the results. However, as we do not use the actual responses (except for the sentiment analysis in the Appendix) but only the named entities present in the response, the translation manner will have less impact. We also manually verify some of the responses to ensure that the LLM does not alter the returned persons. The fact that we only investigate the persons present in the response can also be seen as a limitation, as we do not analyse the remainder of the response or the ordering in which the persons occur. Naous et al. (2023) also found that NER works better for Western persons than for Arabic persons, which could influence the returned persons from other cultures.

The choice of languages and models is also a limiting factor. To avoid any selection bias, we opted for the ten most spoken languages, and three of the most popular LLMs. However, we could extend the analysis to some less popular languages as well. The lower language support may lead to an increase in the superstar effect, as there may be less local cultural awareness. Similarly, an interesting follow-up experiment could be using LLMs developed in different countries and see how this would affect these results. Additionally, we use language as a proxy for culture, while there are obviously important differences between the two (Hershcovich et al., 2022).

Besides this, we use the current version of the LLMs for our experiments, which presents challenges for reproducibility as they can be updated at anytime, potentially altering the results. Lastly, we used the default parameters for every LLM but varying some of the parameters (such as temperature) could also influence the diversity of the response. We opted for the default parameters to reflect the way that most users would interact with the LLMs.

Acknowledgments

This research was funded by Flemish Research Foundation (grant number 1247125N).

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*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023a. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, et al. 2023b. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *arXiv preprint arXiv:2311.07463*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Scott T Allison and George R Goethals. 2016. Hero worship: The elevation of the human spirit. *Journal for the Theory of Social Behaviour*, 46(2):187–210.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Gábor Bella, Khuyagbaatar Batsuren, and Fausto Giunchiglia. 2021. A database and visualization of the similarity of contemporary lexicons. In *International Conference on Text, Speech, and Dialogue*, pages 95–104. Springer.

Ruth Benedict. 2019. *Patterns of culture*. Routledge.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2023. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, pages 1–16.

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.

Erik Brynjolfsson, Yu Hu, and Michael D Smith. 2010. Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research*, 21(4):736–747.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour*, pages 1–13.

Maarten Buyl, Alexander Rogiers, Sander Noels, Iris Dominguez-Catena, Edith Heiter, Raphael Romero, Iman Johary, Alexandru-Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. 2024. Large language models reflect the ideology of their creators. *arXiv preprint arXiv:2410.18417*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Central Intelligence Agency. 2025. The world factbook: World - people and society. Accessed: 2025-03-03.

Elizabeth L Cohen, Mckay West, Koji Yoshimura, Molly E Farrell, and Ashleigh Swain. 2024. Normative influence of the stars: The relative indirect effects of celebrity exemplars on vaping norm perceptions through liking, parasocial relationship strength, and wishful identification. *Health Communication*, 39(9):1877–1887.

Trevor F Cox and Michael AA Cox. 2000. *Multidimensional scaling*. CRC press.

Robert Dorfman. 1979. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149.

Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Anita Elberse and Felix Oberholzer-Gee. 2006. *Superstars and underdogs: An examination of the long tail phenomenon in video sales*, volume 7. Citeseer.

Julia Evans, Jennifer D'Souza, and Sören Auer. 2024. Large language models as evaluators for scientific synthesis. *Preprint*, arXiv:2407.02977.

Valerie Gorin and Annik Dubied. 2011. Desirable people: Identifying social values through celebrity news. *Media, Culture & Society*, 33:599–618.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Hans Henrich Hock and Brian D Joseph. 2009. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Mouton de Gruyter.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–42.

Amr Keleg and Walid Magdy. 2023. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. *arXiv preprint arXiv:2306.05076*.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *Preprint*, arXiv:2406.01771.

Loet Leydesdorff and Caroline S Wagner. 2008. International collaboration in science and the formation of a core group. *Journal of informetrics*, 2(4):317–325.

Yan Liu and He Wang. 2024. Who on earth is using generative ai? *Washington, DC: World Bank*.

Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8):269.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? *Preprint*, arXiv:2309.05196.

Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. 2024. Human-ai coevolution. *Artificial Intelligence*, page 104244.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*.

Rida Qadri, Aida M Davani, Kevin Robinson, and Vinodkumar Prabhakaran. 2025. Risks of cultural erasure in large language models. *arXiv preprint arXiv:2501.01056*.

Vaikunthan Rajaratnam. 2024. Why i'm committed to breaking the bias in large language models. *Nature*.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*.

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.

Charles Spearman. 1961. The proof and measurement of association between two things.

Yuning Sun, Elaine L. Kinsella, and Eric R. Igou. 2024. On cultural differences of heroes: Evidence from individualistic and collectivistic cultures. *Personality and Social Psychology Bulletin*, 50(6):841–856. PMID: 36727610.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Karina Vida, Fabian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1490–1501.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *arXiv preprint arXiv:2406.15053*.

Helen Weeds. 2012. Superstars and the long tail: The impact of technology on market structure in media industries. *Information Economics and Policy*, 24(1):60–68.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*.

Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Appendix

A.1 Adjectives

As explained in Section 3, we test different versions of the prompt by varying the adjective and running each version 5 times. We see in Figure 6 that the adjective 'Greatest' leads to the most returned names on average, and that this is consistent across the LLMs (see Table 7 in the Appendix).

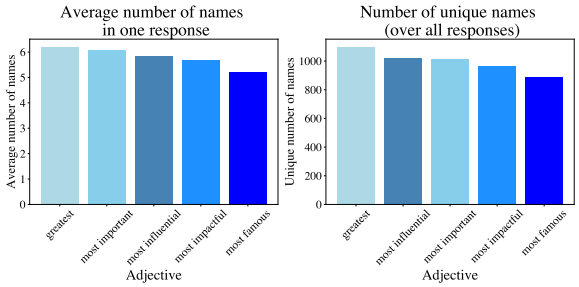


Figure 6: General analysis by adjective

The adjective 'Most Famous' consistently returns the least names. Similarly 'Greatest' leads to the most unique names, and 'Most Famous' to the least. We hypothesize that there may be more universal agreement on the criteria for fame whereas the criteria for adjectives like 'Greatest' may be harder to define. We see in Figure 7 that the adjective 'Greatest' also leads to the most novel names across the professions, although there is only a slight difference. We also conduct a sentiment analysis to assess how the adjectives impact the polarity and subjectivity of the responses and present the results in Section A.2.

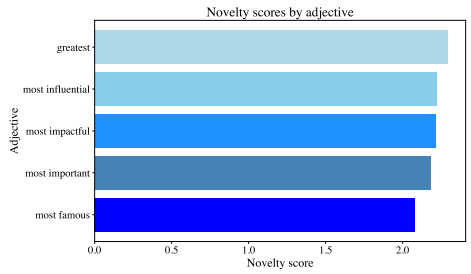


Figure 7: Novelty of responses by adjective

A.2 Sentiment analysis

We use TextBlob to analyze the sentiment of the text responses, measuring polarity (the positivity or negativity) and subjectivity (the degree of opinion versus fact) for each response (Loria et al., 2018). The sentiment of the responses is analyzed after the LLMs' responses are translated back into English. In this case, we use the complete text responses and not only the returned names.

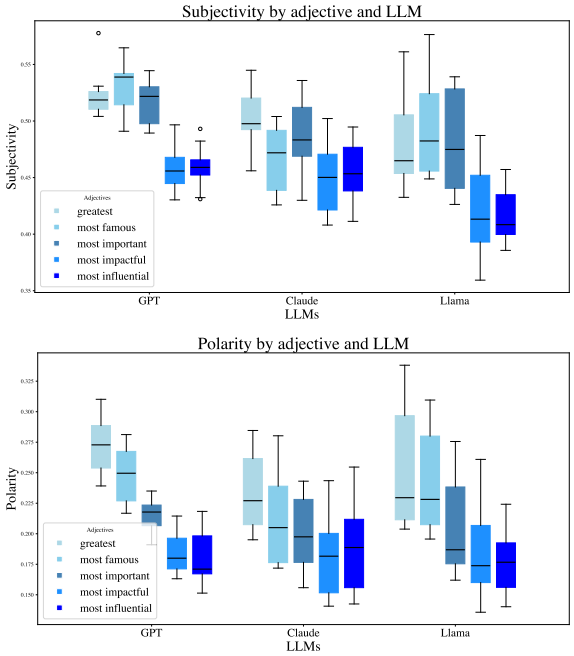


Figure 8: Sentiment analysis

We see in Figure 8 that the average polarity and subjectivity of the text response can vary a lot depending on the used adjective.

A.3 Additional results

In this Section, we present some of the additional results. We display the full results by LLM and adjective in Table 7, by LLM and language in Table 8 and by LLM and profession in Table 9. Table 7 reveals that Claude generates the highest number of unique names within a single adjective. However, it produces the fewest unique names when considering results across different adjectives. This suggests that Claude's outputs are the least affected by variations in prompt phrasing (adjective choice). Table 8 illustrates that Claude produces the most unique names within one language, but the least unique names when we look at the results of all the languages combined. This suggests that the output of Claude is the least affected by the language of the prompt as well. We visualise the frequency distributions across LLMs for every profession in Figure 9, and for each LLM separately in Figure 10 (GPT), Figure 11 (Claude) and Figure 12 (Llama).

Lastly, we visualise the top 10 results for every profession across the different languages, adjectives, LLMs and runs in Table 10.

	Average #names/response				Unique names				
Model	GPT	Claude	Llama	all LLMs		GPT	Claude	Llama	all LLMs
greatest	5.24	8.72	4.63	6.20		509	566	613	1098
most famous	4.25	8.26	3.10	5.20		444	492	435	886
most impactful	5.11	8.48	3.45	5.68		454	536	505	962
most important	5.34	8.81	4.13	6.09		478	549	503	1011
most influential	5.14	8.75	3.67	5.85		426	539	586	1023
All adjectives	5.01	8.60	3.80	5.80	Avg. (per adj.)	462.2	536.4	528.4	998.4
					Total (across adj.)	1158	1023	1386	2409

Table 7: General results by LLM and adjective. We present the average number of names/response and the number of unique names per LLM and adjective.

	Average #names/response				Unique names				
Language	GPT	Claude	Llama	all LLMs		GPT	Claude	Llama	all LLMs
Hindi	5.53	7.64	2.56	5.24		369	360	236	618
Spanish	5.12	9.88	4.97	6.66		193	293	326	477
Urdu	5.11	6.33	3.16	4.86		419	301	558	918
Russian	4.85	8.47	3.04	5.45		204	318	272	490
English	4.57	9.78	9.35	7.90		165	281	401	520
French	5.09	9.63	4.56	6.43		188	309	297	468
Chinese	5.75	9.75	2.30	5.93		360	415	224	647
Portuguese	4.22	9.71	4.45	6.13		160	297	410	551
Bengali	5.58	8.38	2.15	5.37		371	354	244	642
Arabic	4.32	6.49	1.42	4.08		333	289	227	591
All languages	5.02	8.60	3.80	5.80	Avg. (per lang.)	276.2	321.7	319.5	592.2
					Total (across lang.)	1158	1023	1386	2409

Table 8: General results by LLM and language

	Average #names/response				Unique names				
Profession	GPT	Claude	Llama	all LLMs		GPT	Claude	Llama	all LLMs
Artist	4.56	8.44	3.94	5.64		86	96	157	227
Computer Scientist	3.50	8.06	2.39	4.65		74	36	68	119
Chemist	3.38	6.22	2.37	3.99		83	62	94	174
Composer	4.56	8.11	3.47	5.38		123	62	98	208
Poet	5.36	9.35	4.20	6.30		162	225	147	384
Leader	7.41	8.53	5.32	7.08		120	116	158	260
Physicist	2.86	7.66	2.31	4.28		27	53	52	97
Medical Researcher	4.42	7.29	2.85	4.85		106	59	118	203
Philosopher	5.97	10.50	3.90	6.79		84	74	75	152
Person	7.13	11.52	5.99	8.21		123	126	161	266
Political Figure	7.72	8.34	5.92	7.33		155	91	216	333
Economist	3.26	7.80	3.16	4.74		39	31	52	85
Writer	5.20	9.66	4.30	6.39		190	134	206	365
Military Leader	5.08	8.77	3.47	5.77		109	162	126	286
Mathematician	4.82	8.82	3.34	5.66		90	70	158	237
All professions	5.02	8.60	3.80	5.80	Avg. (per prof.)	104.7	93.1	125.7	226.4
					Total (across lang.)	1158	1023	1386	2409

Table 9: General results by LLM and profession

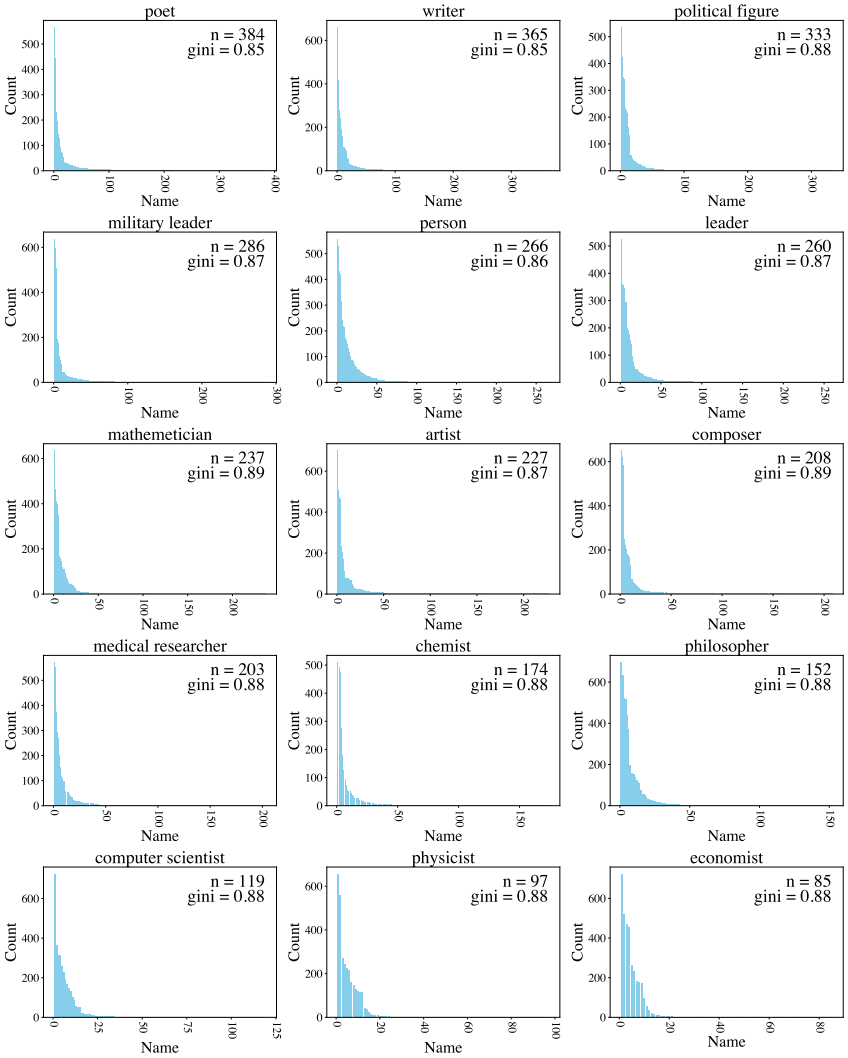


Figure 9: Frequency distribution for every profession (across LLMs). The number of unique names is depicted in the right corner (n).

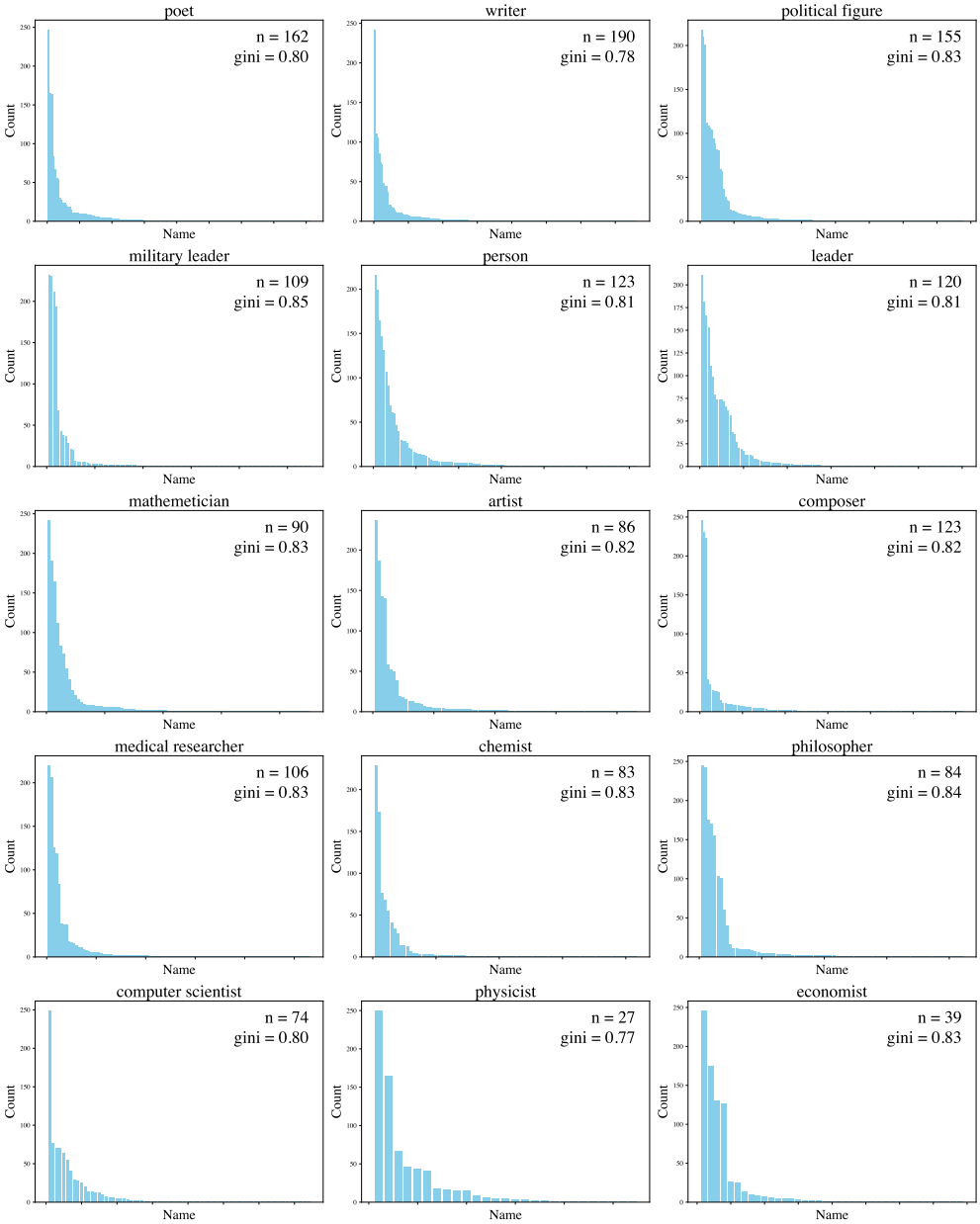


Figure 10: Frequency distribution for every profession (GPT)

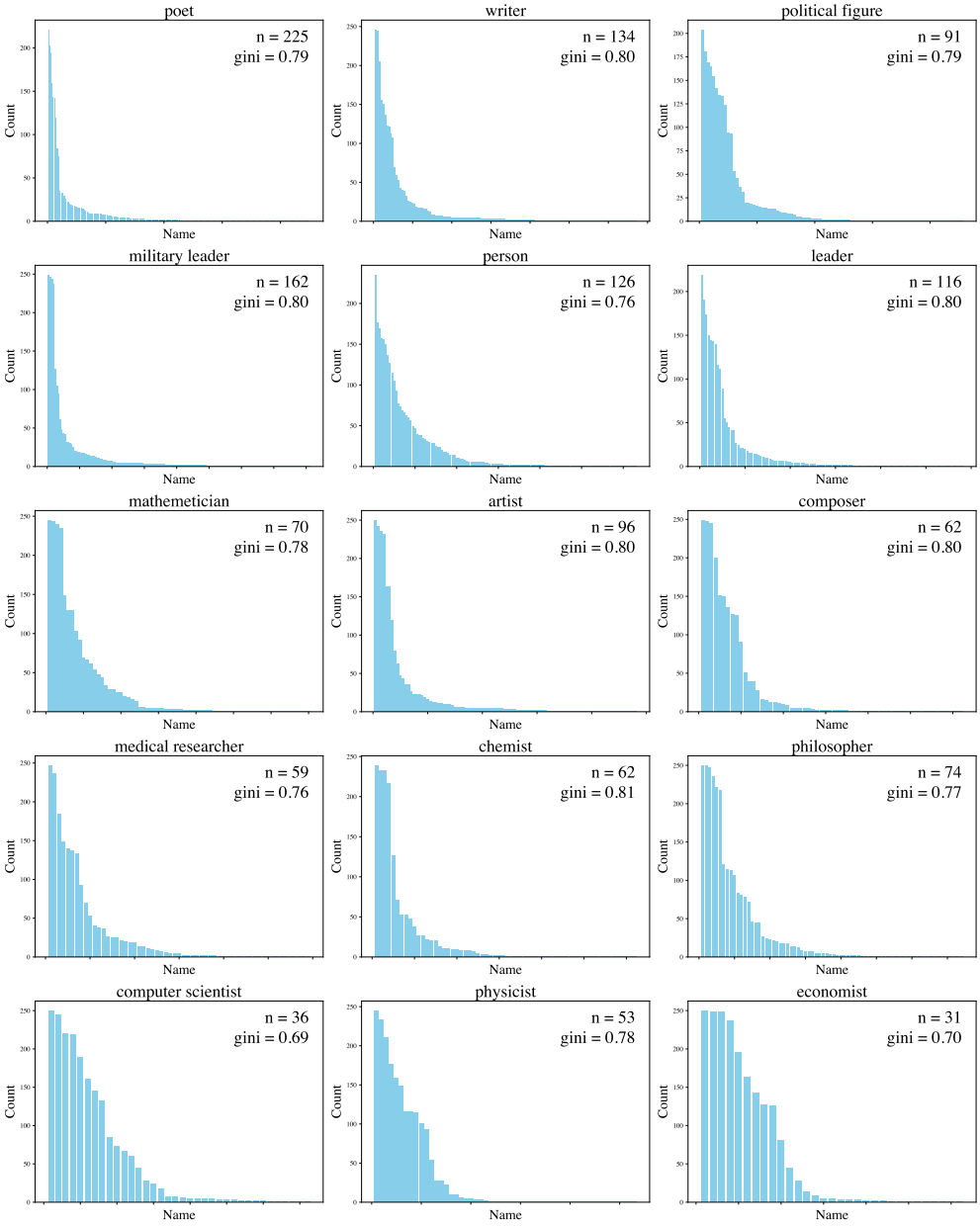


Figure 11: Frequency distribution for every profession (Claude)

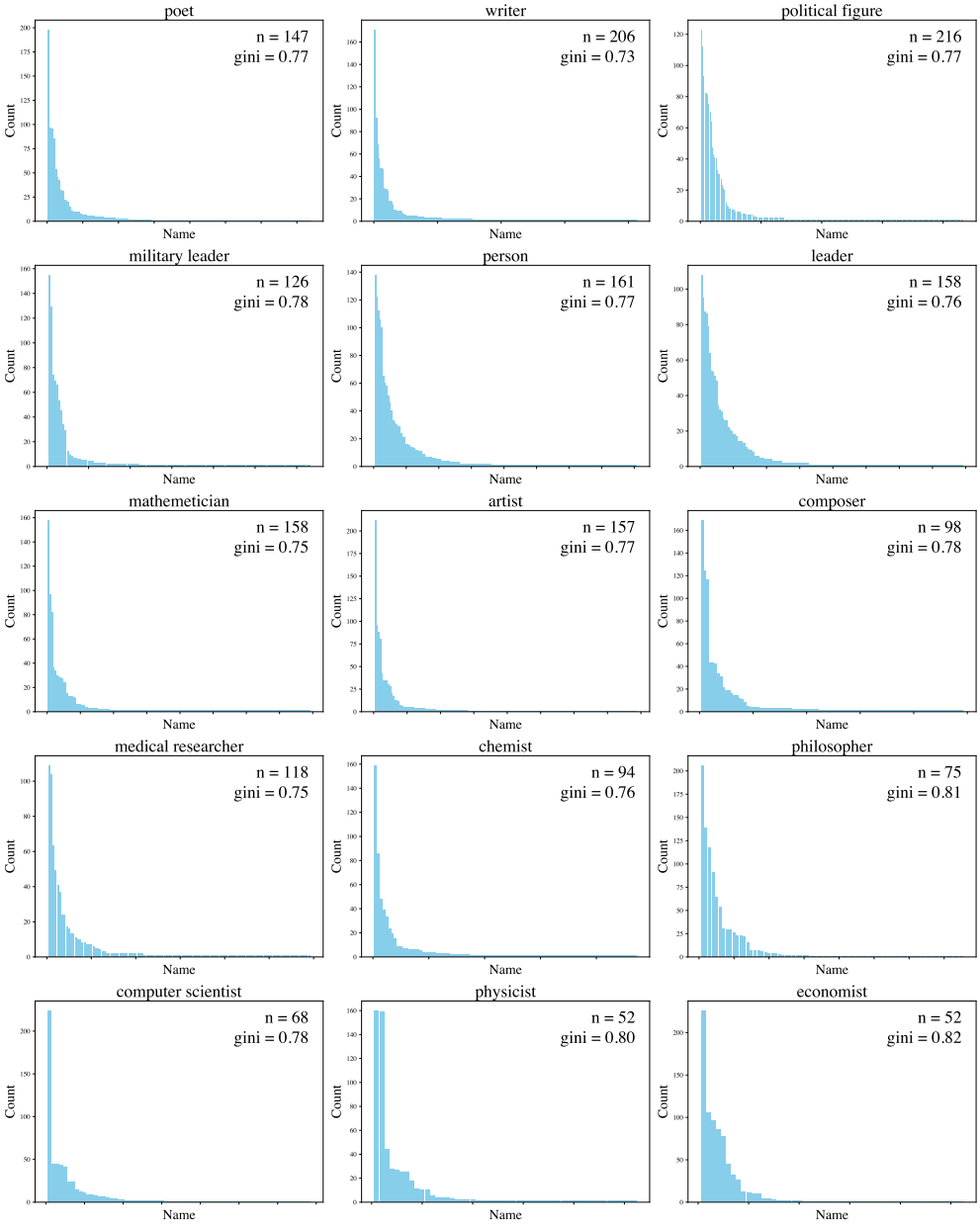


Figure 12: Frequency distribution for every profession (Llama)

Table 10: Results for each profession. We display the count (n) which is the number of responses in which they occur and the percentage (%) which is the percentage of responses in which they occur ($n/750$).

(a) Artist

Name	n	%
Leonardo Da Vinci	699	93.2
Pablo Picasso	508	67.7
Michelangelo	470	62.7
Vincent Van Gogh	464	61.9
Rembrandt	232	30.9
Charles David	204	27.2
Claude Monet	171	22.8
Salvador Dali	112	14.9
William Shakespeare	77	10.3
Ludwig Beethoven	76	10.1

(b) Computer Scientist

Name	n	%
Alan Turing	723	96.4
John Neumann	364	48.5
Tim Berners	314	41.9
Lee	313	41.7
Hopper	257	34.3
Ada Lovelace	230	30.7
Dennis Ritchie	193	25.7
Claude Shannon	166	22.1
Charles Babbage	146	19.5
Donald Knuth	134	17.9

(c) Chemist

Name	n	%
Marie Curie	509	67.9
Dmitri Mendeleev	492	65.6
Lavoisier	474	63.2
Linus Pauling	274	36.5
Dalton	180	24.0
Alfred Nobel	126	16.8
Robert Boyle	96	12.8
Louis Pasteur	72	9.6
Frederick Sanger	54	7.2
Rosalind Franklin	53	7.1

(d) Composer

Name	n	%
Mozart	649	86.5
Beethoven	617	82.3
Bach	584	77.9
Wagner	249	33.2
Chopin	218	29.1
Tchaikovsky	204	27.2
Schubert	179	23.9
Stravinsky	176	23.5
Debussy	164	21.9
Brahms	130	17.3

(e) Poet

Name	n	%
Shakespeare	565	75.3
Homer	565	75.3
Dante	443	59.1
Neruda	236	31.5
Tagore	230	30.7
Rumi	197	26.3
Goethe	181	24.1
Li Bai	144	19.2
Virgil	136	18.1
Whitman	128	17.1

(f) Leader

Name	n	%
Gandhi	525	70.0
Mandela	458	61.1
Churchill	360	48.0
Lincoln	355	47.3
Alexander the Great	347	46.3
Napoleon	296	39.5
MLK Jr.	295	39.3
Julius Caesar	293	39.1
Mao Zedong	198	26.4
Genghis Khan	193	25.7

(g) Physicist

Name	n	%
Einstein	655	87.3
Newton	557	74.3
Galileo	272	36.3
Niels Bohr	243	32.4
Maxwell	227	30.3
Feynman	217	28.9
Hawking	160	21.3
Marie Curie	145	19.3
Max Planck	127	16.9
Faraday	121	16.1

(h) Medical Researcher

Name	n	%
Louis Pasteur	571	76.1
Alexander Fleming	552	73.6
Edward Jenner	373	49.7
Hippocrates	295	39.3
Jonas Salk	270	36.0
Robert Koch	202	26.9
Leon Harvey	152	20.3
Marie Curie	117	15.6
Albert Sabin	108	14.4
Francis Crick	99	13.2

(i) Philosopher

Name	n	%
Plato	695	92.7
Aristotle	634	84.5
Socrates	522	69.6
Kant	516	68.8
Nietzsche	440	58.7
Descartes	373	49.7
Confucius	196	26.1
Sartre	159	21.2
Karl Marx	158	21.1
Hegel	153	20.4

(j) Person

Name	n	%
Einstein	556	74.1
Jesus Christ	531	70.8
Newton	432	57.6
Muhammad	422	56.3
Buddha	353	47.1
Gandhi	317	42.3
Alexander the Great	242	32.3
Mandela	218	29.1
Confucius	213	28.4
Darwin	195	26.0

(k) Political Figure

Name	n	%
Gandhi	534	71.2
Mandela	464	61.9
Churchill	425	56.7
Lincoln	392	52.3
Mao Zedong	347	46.3
Julius Caesar	341	45.5
Napoleon	316	42.1
Alexander the Great	235	31.3
Hitler	227	30.3
Lenin	220	29.3

(l) Economist

Name	n	%
Adam Smith	722	96.3
Keynes	519	69.2
Karl Marx	469	62.5
Friedman	455	60.7
Ricardo	262	34.9
Samuelson	233	31.1
Marshall	181	24.1
Hayek	179	23.9
Schumpeter	171	22.8
Amartya Sen	96	12.8

Table 10: Results for each profession (continued)

(m) Writer

Name	n	%
Shakespeare	659	87.9
Tolstoy	417	55.6
Homer	324	43.2
Cervantes	278	37.1
Dante	275	36.7
Dickens	242	32.3
Dostoevsky	192	25.6
Goethe	175	23.3
Marquez	157	20.9
Victor Hugo	147	19.6

(n) Military Leader

Name	n	%
Alexander the Great	636	84.8
Napoleon	596	79.5
Genghis Khan	531	70.8
Julius Caesar	506	67.5
Hannibal	192	25.6
Erwin Rommel	173	23.1
Sun Tzu	163	21.7
George Patton	117	15.6
Saladin	98	13.1
George Washington	83	11.1

(o) Mathematician

Name	n	%
Newton	634	84.5
Gauss	464	61.9
Archimedes	409	54.5
Euclid	395	52.7
Euler	350	46.7
Leibniz	180	24.0
Einstein	165	22.0
Hilbert	159	21.2
Riemann	146	19.5
Ramanujan	118	15.7