Title: Growing a Tail: Increasing Output Diversity in Large Language Models

Authors: Michal Shur-Ofry[1], Bar Horowitz-Amsalem[1]†, Adir Rahamim[2], Yonatan Belinkov[2]*

Affiliations:

[1]Law Faculty, Hebrew University of Jerusalem; Jerusalem, Israel.

[2]Faculty of Computer Science, Technion – Israel Institute of Technology; Haifa, Israel.

[3]For large groups, use the name of the group or consortium and include a full list of the authors and affiliations at the end of the main manuscript or in the Supplementary Materials.

*Corresponding author. Email: belinkov@technion.ac.il.

Abstract: How diverse are the outputs of large language models when diversity is desired? We examine the diversity of responses of various models to questions with multiple possible answers, comparing them with human responses. Our findings suggest that models' outputs are highly concentrated, reflecting a narrow, mainstream 'worldview', in comparison to humans, whose responses exhibit a much longer-tail. We examine three ways to increase models' output diversity: 1) increasing generation randomness via temperature sampling; 2) prompting models to answer from diverse perspectives; 3) aggregating outputs from several models. A combination of these measures significantly increases models' output diversity, reaching that of humans. We discuss implications of these findings for AI policy that wishes to preserve cultural diversity, an essential building block of a democratic social fabric.

Main Text:

A large body of literature recognized that the mere exposure to diverse cultural and historical contents is a constructive social factor, crucial for promoting tolerance and democratic values. Conversely, a lack of diversity can result in extremism and exclusion (e.g., *1*, *2*). Despite the importance of diversity, research in recent decades identified a trend toward uniformity and standardization in cultural markets, attributed in part to technological progress, mass media, and globalization (e.g., *3*).

How will cultural diversity fare when artificial intelligence (AI) large language models (LLMs) become a prominent source from which people derive information about the world? Recent studies in computer science discussed the notion of "algorithmic monoculture", whereby identical algorithmic outputs may lead decision makers to adopt wrong or biased decisions (4, 5). Our study's focus is not on algorithmic decisions that affect users, but rather on the exposure of users to diversity of contents ("exposure diversity") as a social end in and of itself. The significance of such diversity was recently underscored in the European Union's Artificial Intelligence Act (6). Yet, given that the current technological paradigm underlying LLMs relies on statistical frequencies (*7*), one can expect that LLMs' default outputs will display a short tail and concentrate around the popular and mainstream items. In this research we quantitatively test this hypothesis and evaluate LLMs' diversity of outputs in cases where diversity is a virtue and there are multiple possible outputs. We compare LLMs' outputs to outputs generated by humans. Relying on this baseline evaluation, we design and test several ways to increase LLMs' diversity of outputs.

Results

To obtain a broad outlook, we used 8 different LLMs: GPT4, Gemini, Claude V1.3, Claude-instant 1.1, J2-Mid, J2-Ultra, MPT-Chat, and GPT3.5 (*8*).[i] The first two models were closed models, which allow user interaction only through their user interface ("the UI models"). The additional six models have accessible APIs ("the API models") that allowed us to adapt their "temperature" – a hyperparameter that changes the randomness of the responses generated by the models (*9*).

Our study comprised two stages. At the **first, baseline, stage**, we set the initial temperature of the API models to 0.3.[ii] We asked each of the 8 models the following three questions:

> Q1: Can you list 3 influential persons from the nineteenth century?
> Q2: Can you list three good television series?
> Q3: Can you list three cities worth visiting?

The questions we selected have multiple possible answers rather than a single correct answer, and relate to different aspects of diversity (historical figures, cultural products, and geographical locations). We intentionally selected questions without an immediate political context or implications. Because our focus is on diversity as-such, rather than on bias or algorithmic decision-making, our questions also do not have an expected "correct" distribution of answers. We assume that when people present LLMs with such seemingly mundane questions, they will likely rely on their outputs without reservations (*10*).

We reiterated each question 10 times per model, so that the overall number of 'votes' (e.g., 19th century figures, tv series, or cities) comprising the outputs was 30 votes per model.[iii] We then examined how many *different* items appear in this output. We reiterated this process for each of

the treatments described below. Altogether our dataset comprised approximately 3780 'votes', which we manually reviewed and analyzed.

Measuring output diversity: There is no uniform accepted metrics for cultural diversity (e.g., *11*). We therefore used **entropy,** which measures the level of information or surprise in possible outcomes of a random variable, and provides a good indication of how distributed (diverse) the models' responses are (*12*).[iv]

In addition, we used several internal, content-related measures, specifically tailored for each question ("the Internal Measures"). For Q1 (19th century figures) we measured the percentage of non-western figures, the percentage of females, and the percentage of figures who were *not* political leaders, that appeared in the responses. For Q2 (tv series), we measured the percentage of non-English series, and examined the earliest production year that appeared in the votes. For Q3 (cities) we measured the number of different countries in the outputs.

Comparison to humans: To get a sense of how the models outputs compare to humans, we posted the same questions on the 'X' account of one of the authors, and asked people to respond without using any LLMs or search engines. We considered the first 10 replies received for each question, which similarly comprised 3 votes per response. After aggregating 30 human votes from 10 responders, we similarly counted the different items included in the aggregation of votes.[v]

In the second stage, we employed several measures to increase output diversity, while repeating 10 reiterations per model under each treatment. **First**, we raised the **temperature** of the six API models from 0.3 to 1. **Second**, we adjusted the **prompts** by adding the words: "answer from diverse perspectives", before the original prompt in one treatment (e.g., "Answer from diverse perspectives: Can you list three good television series?") (the "diversity ante" prompt), and after the original prompt in a second treatment (e.g., "Can you list three good television series? Answer from diverse perspectives") (the "diversity post" prompt). In the API models we further combined the high temperature with the diversity-inducing prompting. **Third,** we **aggregated** the outputs of the 6 API models and of the 2 UI models under each of the diversity-inducing conditions, and applied our diversity metrics to the aggregated outputs.

Research indicates that LLMs possess the capability to simulate different worldviews, or "personas" (*13*, *14*). For instance, prompting an LLM to answer as a professor may increase its truthfulness (*15*), and prompting it to consider the perspective of a certain country makes responses similar to that country's population (*16*). We therefore hypothesize that explicit prompting may increase output diversity. Furthermore, as different LLMs are trained on somewhat different datasets, they may provide different default outputs, or assume different default personas when prompted for diversity, and so we hypothesize that combining their responses may increase diversity.

Low diversity of default LLM responses

In the **baseline stage,** all the models we examined generated **highly concentrated, non-diverse, outputs,** repeating a small number of popular, mainstream, items again and again in each reiteration. For example, in Q1 (19th century personae), most of the models repeatedly generated Charles Darwin, Karl Marx, and Abraham Lincoln in their responses. Out of 30 votes, and 30 potential different responses, the average number of actual different persons was 4.3 for the six API models, and 5 for the two UI models, implying that all models displayed an extremely short-tailed distribution. Conversely, the 28 human votes on 'X' to the same question comprised 25

different figures out of a possible 30, displaying a very long tail in comparison. Very similar patterns emerged with respect to the outputs generated for Q2 (television series) and Q3 (cities): Despite the breadth of potential outputs, the actual outputs generated by the AI models to each of these questions concentrated around a small number of repeating highly-popular items, displaying a very short tail, while the human responses on 'X' were significantly more diverse and distributed, as illustrated in *Figure 1*. The complete data generated by the AI models and humans in response to each of the questions appear in the Supplementary Materials, Parts II-IV.
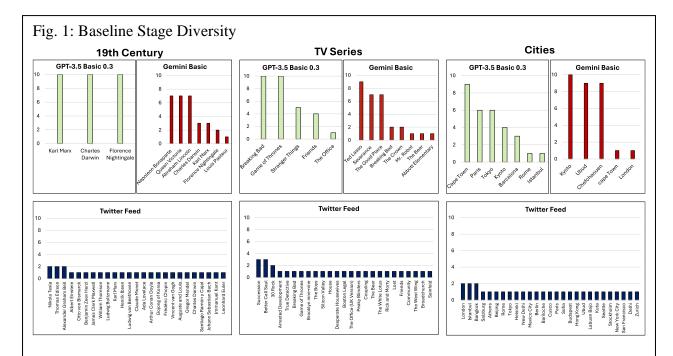
Fig. 1: Baseline Stage Diversity

Figure 1. **Distribution of responses by language models and humans.** The distribution of 30 responses received to each of the three questions we posed at the baseline stage, by one of the API models (GPT-3.5, in green), one of the UI models (Gemini, in red), and participants on 'X' (blue). The x-axis details the *different* responses received to each question, while the y-axis shows the number of 'votes' for each response, namely the number of times the specific response appeared in the results.

Accordingly, the **entropy levels** of the models under the basic treatment for all three questions were relatively low, with averages ranging between 1.8 and 2.2. The entropy of the human responses was significantly higher, ranging between 4.5 and 4.7, as summarized in **Table 1.**

	Human ['X']		6 API MODLES						2 UI MODELS		
			Base	Temp 1.0	Prompt D-ante	Prompt D-post	Prompt D-ante & Temp 1.0	Prompt D-post & Temp 1.0	Base	Prompt D-ante	Prompt D-post
19TH CENTURY	4.6	AVE	2.1	2.5	2.3	2.5	3.0	3.4	2.4	3.8	2.8
		AGG	3.0	3.1	3.3	3.6	3.7	4.2	2.7	4.2	3.8
TV SERIES	4.6	AVE	2.6	3.1	2.3	2.9	3.4	4.0	2.4	3.0	2.8
		AGG	3.2	3.9	4.0	3.9	4.8	5.1	3.1	3.9	3.8
CITIES	4.7	AVE	2.3	2.9	2.5	2.4	3.4	3.4	2.2	3.4	2.6
		AGG	2.8	3.4	3.4	3.4	4.3	4.0	2.7	4.2	3.1

Table 1. **Diversity of responses.** The entropy levels of the responses attained by the API models (in green), the UI models (red) and participants on 'X' (blue), under each of the different conditions, for each of the three questions. The columns represent the different treatments. The "AVE" line specifies the average entropy of the responses generated by each group of models, under each treatment. The "AGG" line specifies the entropy of responses generated when aggregating each group of models under each treatment. The colors are conditioned in a gradient way, so that sharper colors reflect higher entropy.

Diversity-promoting measures
How did output diversity of LLMs change when employing diversity-increasing measures? **Table 2** summarizes our main results. More detailed results appear in the Supplementary Materials, Parts II-IV.

	Human ['X']		6 API MODLES						2 UI MODELS		
			Base	Temp 1.0	Prompt D-ante	Prompt D-post	Prompt D-ante & Temp 1.0	Prompt D-post & Temp 1.0	Base	Prompt D-ante	Prompt D-post
19TH CENTURY /28	25	AVE	4.3 /30	7.8 /29	5.3 /30	6.7 /30	11.3 /29	13.3 /30	5.0 /30	9.0 /30	7.5 /30
		AGG	10.0 /180	19.0 /171	16.0 /180	21.0 /180	33.0 /176	37.0 /180	8.0 /60	16.0 /60	13.0 /60
TV SERIES /30	25	AVE	5.7 /30	11.0 /30	5.8 /30	9.2 /30	14.2 /30	17.5 /30	6.5 /30	9.5 /30	10.0 /30
		AGG	17.0 /180	37.0 /180	22.0 /180	33.0 /179	58.0 /180	66.0 /180	11.0 /60	18.0 /60	20.0 /60
CITIES	27	AVE	5.2 /30	9.8 /29	7.0 /30	7.0 /30	13.5 /30	13.6 /30	5.5 /30	8.5 /30	7.5 /30
/30		AGG	14.0 /180	30.0 /176	19.0 /180	21.0 /180	44.0 /177	41.0 /179	9.0 /60	22.0 /60	13.0 /60

Table 2. **Number of responses.** The aggregate number of responses attained by the API models (in green), the UI models (red) and participants on 'X' (blue), under each of the different conditions, for each of the three questions. The columns represent the different treatments. The "AVE" line specifies the average numbers of *different* responses generated by each group of models, under each treatment. The smaller number appearing in the lower part of the square is the average number of 'votes' generated by that group of models under the specific treatment. The "AGG" line specifies the numbers of *different* responses generated when aggregating the entire group of models, while the smaller numbers appearing in the lower part of the square are the total number of 'votes' generated by the aggregation of that group of models. The colors are conditioned in a gradient way, so that sharper colors reflect greater diversity of responses.

As is evident from Table 2, **raising the temperature** of the API models from 0.3 to 1 increased the variety of responses by roughly 80%-100%. For example, in Q1 (19th century figures), the average number of different responses increased from 4.3 to 7.8 (out of a possible 30). Increases in diversity under maximum temperature further occurred in Q2 and Q3, where, similarly, the number of different items roughly doubled.

Prompting the models to answer from diverse perspectives increased the variety of the responses by approximately 20% to 60%, relative to the baseline stage. For example, in Q1 (19th century figures), the average number of different persons in the API models increased from 4.3 under the baseline condition, to 5.33 under the diversity-ante prompt, and 6.66 under the diversity-post prompt. The average number of different persons in the UI models' outputs similarly increased from 5 under the baseline condition, to 9 under diversity-ante prompt, and 7.5 under the diversity-post prompt. Roughly similar increases in diversity under the two diversity-inducing prompts occurred in Q2 and Q3.[vi]

Combining the **diversity prompting with a maximum temperature** in the API models proved more effective than each of these measures alone, resulting in an increase of roughly 150%-200% in the variety of the responses, relative to the baseline.

Internal diversity measures: Under the diversity-inducing treatments, we have also seen limited increases in some of the internal, content-related diversity measures, which we tailored to each question. To illustrate, in Q1 (19th century figures), the average percentage of female figures in the outputs of the API models increased from 17% under the baseline condition, to 21% under the maximum-temperature condition, 25% under the diversity-post prompting, and 30% under the combined treatment of maximum-temperature plus diversity-post prompting. The percentage of non-western persons in the UI models' outputs increased from zero under the baseline condition, to 2-3% under the diversity-increasing prompting. We note, however, that increases in internal diversity measures were often slight, and did not occur with respect to all questions and parameters. For example, in Q2 (tv series) most models did not generate a greater number of older series under the diversity-increasing treatments. Additionally, even under the diversity-inducing treatments, the responses generated by the models, often reflected well-known, popular, and mainstream items (particularly in Q1 and Q2). The complete data pertaining to internal diversity measures appear in the Supplementary Materials, Part I-4.

Despite the measures described above, none of the models' outputs reached diversity levels comparable to those attained by humans responding on 'X': Even under combined treatment of diversity-prompting plus maximum temperature, the variety of human responses was roughly twofold greater than the models' averages (see Table 2), and the entropy of human responses remained substantially higher than that of the models (see Table 1). We therefore turned to examine whether aggregating several models can further improve output diversity, and analyzed the aggregations, rather than averages, of responses under each condition. We aggregated the 6 API models, and the 2 UI models, in two separate groups.

Aggregation of models: Under the baseline condition, the aggregation of models' outputs did not substantially change the diversity picture. For example, in Q1 (19th century figures), the 60 aggregated votes of the UI models generated only 8 different persons, a number identical to the average output of those models. Similarly, the 180 aggregated votes of the 6 API models generated only 10 different persons. This implies that the models' **default outputs** were **overlapping** to a large extent. In Q2 and Q3 the models' aggregated outputs (approximately 60 outputs of the `UI models and 180 outputs of the API models) yielded a somewhat greater variety of items, but still did not reach a level comparable to the 30 human responses on 'X' (see Table 2). However, combining the diversity inducing measures--namely high temperature, diversity- prompts, or both--with an aggregation of models, often increased output diversity to a level in the range of, and in some cases exceeding, the diversity of human responses on 'X'. For example, in Q2 (tv series) the aggregation of 2 UI models with a diversity-inducing prompts yielded 18 different series (under diversity-ante) or 20 different series (under diversity-post), compared to 25 different series on 'X'. The aggregation of 6 API models with the diversity-inducing treatments plus a high temperature resulted in 58 different series (under diversity-ante) or 60 different series (under diversity-post), numbers that exceed the number of human responses (see Table 2). These results imply that aggregating models while using diversity inducing treatments **decreases the overlap** between the outputs of different models. The aggregated outputs in such cases further displayed a long-tail distribution, as illustrated in **Figure 2**. The complete data pertaining to these distributions appear in the Supplementary Materials, Part I Section 5, and Parts II-IV.

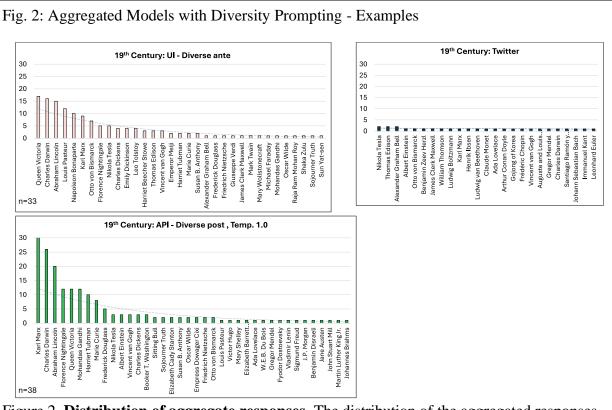Fig. 2: Aggregated Models with Diversity Prompting - Examples

Figure 2. **Distribution of aggregate responses.** The distribution of the aggregated responses to Q1 (19th century figures) received from the LLMs under 'combined treatments' conditions, relative to the distribution of responses to that question by participants on 'X'. The upper-left image (red) depicts the distribution of approximately 60 aggregated 'votes' received from 2 UI models, under a diversity-ante prompting. The bottom image (green) depicts the distribution of approximately 180 aggregated 'votes' received from 6 API models, under a diversity-ante prompting. The upper-right image (blue) depicts the distribution of 30 aggregated votes received from participants on 'X'.

Notably, even under the "aggregated" condition, the ratio between the number of different items generated by the models and the overall number of votes remains substantially low relative to the ratio in the human replies. Likewise, in the vast majority of cases, the **entropy** of the human responses comprising a total of 30 votes still exceeds the entropy of the aggregated models' responses, comprising a total of 180 (for API models) or 60 (for UI models) votes. Therefore, the aggregated results do not attest to the diversity of the models in-and-of themselves. Rather, they highlight that each models reflects a somewhat different 'worldview', and attest to the significance of combining several models.

Finally, we also monitored the numbers of errors generated under each condition (examples for errors are, e.g., naming William Shakespeare who lived in the 16th century in response to Q1, or naming a film rather than a television series in response to Q2). One concern with promoting diversity is the generation of more erroneous answers, given previous findings that accuracy and diversity of language model generations can be incongruent [*17*, *18*]. Interestingly, we did not find that the diversity-inducing treatments substantially increased the percentage of errors in the

models' outputs. The relevant errors-data appear in the Supplementary Materials, Part I, Section 6.

Discussion

Our findings yield several primary insights. **First**, the default outputs of LLMs, in cases where diversity is warranted, reflect a narrow and extremely concentrated worldview, rather than a multiplicity of options. The models' outputs in all our baseline treatments displayed a very short tail, converging around a small number of popular and well-known items. Due to significant overlaps among the outputs of different models, aggregating default outputs from several models was insufficient to substantially increase output diversity. Human responses displayed a much broader distribution and higher levels of entropy in comparison to all the models we examined. This result is not entirely surprising, given that the main technological paradigm underlying LLMs relies on statistical probability: the frequency of an item's occurrences in the underlying training datasets crucially influences the generated outputs. Therefore, even when a multiplicity of responses is desirable, as is often the case in cultural, historical, and geographical contexts, LLMs outputs are likely to be geared toward the mean, standard, and popular.

      Second, despite the concentrated outputs in the models' default states, a combination of fairly simple measures, namely temperature increase and diversity-inducing prompts, can substantially increase output diversity. Interestingly, combining the two diversity inducing measures (temperature plus prompting) produced 'a whole that is greater than the sum of its parts', namely the difference between the combined measures and the default state is larger than the sum of the differences between each individual measure and the default state.

      Third, even after combining maximum temperature with diversity prompting, none of the models alone generated a long tail of outputs that was comparable to, or exceeded, the long tail of outputs generated by humans. Achieving this level required aggregating the outputs of several models, while applying one or more diversity-inducing measures. In other words, although all the models we examined displayed a rather narrow worldview, their aggregation, while applying other diversity-inducing measures, projected a broader 'slice' of the world. This latter insight highlights the importance of access to different language models, rather than a single one.

      Our inquiry has several limitations. First, while the three case studies we selected pertain to different aspects of cultural diversity, they are of course not exhaustive. Extending our findings to additional contexts, and to additional AI models that are not language-based (e.g., image or video generators) requires further research. In addition, our use of 'X' threads for evaluating diversity among humans is limited in several respects. The human respondents in the threads are limited in number and do not constitute a representative sample of human society. However, our main focus in this study was not the contents of the responses but rather on their diversity and distribution--e.g., we were not exploring which particular figures people deem as the most important in the 19th century, but rather how distributed (or concentrated) their answers are. We further note that previous works analyzing people's cultural choices, including comparable threads on social media accounts, similarly found long tail distributions (*10*, *19*). We also acknowledge that attaining responses from 10 different people on 'X' is not entirely equal to obtaining 10 reiterations from the same language model. Yet, while a perfect comparison between humans and LLMs is likely impossible, our method of 10 reiterations per model roughly simulates real-world scenarios, whereby different users who approach a language model with similar questions will receive largely uniform responses, despite the broad variety of potential answers. Therefore, the aggregation of models' responses is indicative of a potential decline in diversity.

Lastly, while our diversity-inducing measures were undoubtedly efficient in producing longer tails of outputs, they did not consistently generate a significant improvement in content-related parameters. For example, in Q2, the television series outputs under all treatments comprised almost exclusively English-speaking series. This implies that improving the content-related aspects of output diversity, in a way that would expose users to both local and foreign outputs, contemporary and past outputs, high and popular culture, etc., is a more complicated challenge (*1–3*, 11, *20*). Relatedly, we note that the diversity-inducing measures we applied are relatively simple. More technologically advanced methods, such as chain-of-thought prompting and its derivatives (*21*), different decoding techniques (*22–25*), or access to external tools and knowledge bases (*26–27*) may prove more effective in improving levels of diversity along additional dimensions. We mark this as a topic for future research

Policy Outlook and Conclusion

The general picture emerging from our study carries significant implications for policymaking in the field of AI. Our results indicate that, despite being trained on vast amounts of materials, LLMs are not geared toward diversity in their default interactions with users, but rather toward standardization, conformity and mainstream. Since the outputs of language models are likely to influence their users' perceptions (*10*), in the long term we can expect a general decline in cultural diversity. Yet, our study further implies that addressing the challenge of maintaining diversity does not necessitate extremely complicated or expensive interventions. Our results show that diverse contents are embedded within the models. A few simple and cheap measures, such as temperature increase and diversity-inducing prompting, can 'extract' these contents and significantly improve diversity levels. AI developers can easily incorporate such elements as design features, available to users in a transparent and accessible way. Policymakers could adopt measures to encourage developers to take such steps, for example, by explicitly including diversity as a high-level principle in AI ethical guidelines and regulation (similar to other acknowledged principles such as 'transparency', 'data security', etc.). Concomitantly, users who are aware of language models' propensity toward uniformity can use those simple measures to induce more diverse outputs. Our findings therefore highlight the importance of promoting AI literacy among general audiences, to encourage such proactive and informed uses.

Finally, our findings indicate that the aggregation of several LLMs substantially increases output diversity. Even if an individual user is unlikely to consult with numerous models, the availability of several models in the market would expose different users to different outputs, and could increase societal diversity in the aggregate. Therefore, our study provides further grounds for the need to ensure competition and diversity in the market for language models. More generally, our research marks cultural diversity as a societal goal that should be explicitly addressed by stakeholders and policymakers in the field of AI, in order to prevent systemic harm to social tolerance, solidarity, and democracy.

References and Notes

[1] S. Benhabib, *The Claims of Culture: Equality and Diversity in the Global Era* (2002, Princeton University Press).
[2] R. C. Post, Democratic Constitutionalism and Cultural Heterogeneity. *Australian Journal of Legal Philosophy* **25**, 185–204 (2000).
[3] C. E. Baker, *Media, Markets, and Democracy* (2002, Cambridge University Press).

[4] R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, P. S. Liang, "Picking on the same person: Does algorithmic monoculture lead to outcome homogenization?" in *Advances in Neural Information Processing Systems*, (2022), pp. 3663-3678.

[5] J. Kleinberg, M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, **118**(22), 1–7 (2021).

[6] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), PE/24/2024/REV/1, 2024 O.J. (L 1689) 1 (EU), Preamble 27.

[7] Y. LeCun, Y. Bengio, G. Hinton, Deep Learning. *Nature* **521**, 436-444 (2015).

[8] For details on each of these models, see O.J. Achiam, , S. Adler, S. Agarwal, … , B. Zoph. GPT-4 Technical Report. arXiv:2303.08774 (2023) (GPT 3.5 and GPT4); Gemini Team., R. Anil, S. Borgeaud, Y. Wu, J. B. Alayrac, J. Yu, ... , J. Ahn. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 (2023) (Gemini)*;* O. Lieber, O. Sharir, B. Lenz, Y. Shoham, "Jurassic-1: Technical Details and Evaluation" (White Paper, AI21 Labs, 2021) (J2-Mid, J2-Ultra); MosaicML NLP Team, "Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs" (Blog Post, MosaicML, 2023); www.mosaicml.com/blog/mpt-7b (MPT-Chat); Anthropic. "Introducing Claude" (Blog Post, Anthropic, 2023); https://www.anthropic.com/news/introducing-claude (Claude V1.3; Claude-instant 1.1).

[9] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang,. A tutorial on energy-based learning. *Predicting structured data* **1** (2006).

[10] M. Shur-Ofry, Multiplicity as an AI Governance Principle. *Indiana Law Journal* (2024, forthcoming)

[11] Ph. M. Napoli, Exposure Diversity Reconsidered, *Journal of Information Policy,* 246–259. (2011).

[12] C. E. Shannon, A mathematical theory of communication. *The Bell system technical journal*, **27**(3), 379–423 (1948).

[13] J. Andreas, "Language Models as Agent Models" in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Association for Computational Linguistics, 2022), pp. 5769–5779.

[14] N. Joshi, J. Rando, A. Saparov, N. Kim, H. He, Personas as a way to model truthfulness in language models. arXiv:2310.18168 (2023).

[15] S. Lin, J. Hilton, O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, 2022), pp. 3214–3252.

[16] E. Durmus, K. Nguyen, T. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, L. Lovitt, S. McCandlish, O. Sikder, A. Tamkin, J. Thamkul, J. Kaplan, J. Clark, D. Ganguli, "Towards Measuring the Representation of Subjective Global Opinions in Language Models" in *First Conference on Language Modeling* (2024).

[17] H. Zhang, D. Duckworth, D. Ippolito, A. Neelakantan, "Trading Off Diversity and Quality in Natural Language Generation" in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)* (Association for Computational Linguistics, 2021), pp. 25–33.

[18] M. Nadeem, T. He, K. Cho, J. Glass, "A Systematic Characterization of Sampling Algorithms for Open-ended Language Generation" in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (Association for Computational Linguistics, 2020), pp. 334–346.

[19] Ch. Anderson, *The Long Tail* (Hyperion, 2006).

[20] S. M. Corse, *Nationalism and Literature: The Politics of Culture in Canada and the United States* (Cambridge University Press, 1997).

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, D. Zhou, "Chain of Thought Prompting Elicits Reasoning in Large Language Models" in *Advances in Neural Information Processing Systems* (2022), pp. 24824–24837.

[22] S. Kumar, B. Paria, Y. Tsvetkov, "Gradient-based Constrained Sampling from Language Models" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2022), pp. 2251–2277.

[23] S. Kumar, E. Malmi, A. Severyn, Y. Tsvetkov, "Controlled Text Generation as Continuous Optimization with Multiple Constraints" in *Advances in Neural Information Processing Systems* (2021), pp. 14542–14554.

[24] L. Qin, S. Welleck, D. Khashabi, Y. Choi, "Cold Decoding: Energy-based Constrained Text Generation with Langevin Dynamics" in *Advances in Neural Information Processing Systems* (2022), pp. 9538–9551.

[25] G. Wiher, C. Meister, R. Cotterell, On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics* **10**, 997–1012 (2022).

[26] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, "Toolformer: Language Models Can Teach Themselves to Use Tools" in *Advances in Neural Information Processing Systems* (2023).

[27] C. Qu., D. Dai, X. Wei, H. Cai, S. Wang, D. Yin, J. Xu, J. R. Wen. Tool Learning with Large Language Models: A Survey. arXiv:2405.17935 (2024).

[i] We conducted the study over eight months between July 2023 and February 2024, and used representative LLMs available during that period. For a detailed timetable of data collection, see Supplementary Materials, Part I, Section 2. Some of the models may have undergone certain developments during the study's period, (e.g., changes in training materials), unbeknownst to us. Such possible changes, if occurred, may have affected the uniformity of outputs in specific cases, but are unlikely to have had a substantial effect on our inquiry, which focuses on general patterns, and not on specific outputs.

[ii] The default user interface temperature in these models is not officially published.

[iii] Some of the models' replies included fewer than three votes, despite our prompting. In such cases the aggregate number of votes was slightly lower than 30, as detailed in Table 2 and the Supplementary Materials, Parts II-IV.

[iv] For details and explanations on the measurement of entropy, see Supplementary Materials, Part I, Section 1-b.

[v] Links to the relevant X's threads are in the Supplementary Materials, Part I, Section 3. In Q1, one of the responses on 'X' included 1 rather than 3 votes, so that the aggregate number of actual human votes for that question was 28.

[vi] Interestingly, in the UI models group the increase in diversity was greater under the "diversity post" prompt relative to the "diversity ante" prompt, whereas in the API models group the increase was greater when the prompting inducing diversity preceded the actual question ("diversity ante").

Supplementary Materials for
Growing a Tail: Increasing Output Diversity in Large Language Models

Michal Shur-Ofry[1], Bar Horowitz-Amsalem[1]†, Adir Rahamim[2], Yonatan Belinkov[2]*
Corresponding author: belinkov@technion.ac.il

Contents

Part I

Questions

The three questions we used in the study, defined as Q1, Q2 and Q3, were the following:
Q1: Can you list 3 influential persons from the nineteenth century?
Q2: Can you list three good television series?
Q3: Can you list three cities worth visiting?

Entropy

Our entropy calculation aims to assess the diversity in responses from a given model to a given question, with a certain combination of prompting (base, post, or ante) and temperature. The larger the number of different answers, the greater should be the diversity. However, a naïve calculation of entropy by only considering the set of answers for a given combination of model-prompt-temperature may not allow for fair comparison, since the set of answers for each such combination may be different. Instead, we wish to consider the set of all possible answers produced by any combination of model, prompt, and temperature for a given question. To allow for unseen events in the entropy calculation (e.g., answers that were not produced by a given model by were given by a different model), we "smooth out" the answer counts by adding a small number. Concretely, we assign such cases a value of 1/(num_answers), where "num_answers" is the number of all possible answers produced by any model, prompt, and temperature. In practice, this modification primarily affects low-diversity settings by slightly increasing their entropy, while high-diversity settings remain largely unaffected.

Data Collection Dates

The table below includes the dates on which we generated output from the various models, categorized by the different research questions and, in some instances, the different scenarios.

Date	LLM's Model & Question
Monday, July 10, 2023	J2-mid, J2-ultra, Claude V1.3 & V1.1 Q1
Wednesday, August 9, 2023	GPT-3.5 Q1
Wednesday, July 19, 2023	MPT Q1
Monday, February 5, 2024	6 API models Q3
Monday, February 5, 2024	6 API models Q2
Sunday, July 30, 2023	GPT-4 Q1 - Basic, Diverse post
Tuesday, January 30, 2024	GPT-4 Q3
Monday, February 5, 2024	6 API models Q1
Tuesday, February 20, 2024	6 API models Q2
Tuesday, February 20, 2024	6 API models Q3
Thursday, February 22, 2024	GPT-4 Q1 - Diverse Ante
Thursday, February 22, 2024	Gemini Q1
Thursday, February 22, 2024	Gemini Q3 - Diverse Ante
Sunday, February 25, 2024	Gemini Q3 - Basic, Diverse Post
Saturday, February 24, 2024	Gemini Q2
Saturday, February 24, 2024	GPT-4 Q2

X's threads

Below are the links to the original X (formerly Twitter) threads that contain the questions addressed in the study, and the responses thereto.

    Q1 – https://x.com/boknilev/status/1679470574270050312?s=20.
    Q2 – https://x.com/boknilev/status/1759613863111061902.
    Q3 – https://x.com/boknilev/status/1757406317835145661.

Internal Diversity Measures

The charts include variables defined as Diversity Measures, encompassing various non-numeric characteristics relevant to each type of research question, including demographic attributes, geographic information, cultural aspects etc.

Q1

	API Models						UI Models		
	Basic (with 0.3 Temperature)	Basic (with 1.0 Temperature)	Diverse post (with 0.3 Temperature)	Diverse post (with 1.0 Temperature)	Diverse Ante (with 0.3 Temperature)	Diverse Ante (with 1.0 Temperature)	Basic	Diverse Post	Diverse Ante
Avg. non-western[1]	9%	5%	7%	12%	11%	11%	0%	2%	3%
Avg. Females	17%	21%	25%	30%	21%	22%	32%	33%	37%
Avg. Politics & Leadership[2]	34%	41%	46%	51%	46%	43%	32%	35%	38%
Avg. Science & Technology[3]	42%	36%	28%	25%	27%	31%	50%	33%	45%
Avg. Culture & Humanities[4]	24%	23%	26%	24%	27%	27%	18%	18%	13%

---

[1] In this study, "Western" is defined as including North America, Western Europe, Australia, and New Zealand.

[2] "Politics & Leadership" refers to individuals whose primary occupations were in politics and various leadership positions. This includes politicians, government officials, and leaders in both public and private sectors.

[3] "Science & Technology" refers to individuals whose primary occupations were in the fields of scientific research, technological development, and related disciplines. This includes scientists, engineers, researchers, and technologists.

[4] "Culture & Humanities" refers to individuals whose primary occupations were in cultural and humanistic disciplines. This includes artists, writers, philosophers, historians, and other professionals engaged in the exploration and expression of human culture, values, and historical contexts.

Q2

	API Models						UI Models		
	Basic (with 0.3 Temperature)	Basic (with 1.0 Temperature)	Diverse post (with 0.3 Temperature)	Diverse post (with 1.0 Temperature)	Diverse Ante (with 0.3 Temperature)	Diverse Ante (with 1.0 Temperature)	Basic	Diverse Post	Diverse Ante
Earliest Production year	1994	1989	1959	1959	1994	1989	2005	1999	1959
Avg. non-English[5]	0%	1%	0%	3%	1%	1%	0%	0%	2%
Avg. No. of countries of origin	1	2	1.83	2.83	1.83	2	1.5	2	3
Avg. Old Series[6]	11%	10%	10%	12%	11%	10%	0%	7%	10%

Q3

	API Models						UI Models		
	Basic (with 0.3 Temperature)	Basic (with 1.0 Temperature)	Diverse post (with 0.3 Temperature)	Diverse post (with 1.0 Temperature)	Diverse Ante (with 0.3 Temperature)	Diverse Ante (with 1.0 Temperature)	Basic	Diverse Post	Diverse Ante
Avg. No. of countries	5	8	7	11	7	10	5	7	9

---

[5] "Non-English" refers to TV series that are not American, Canadian or British productions.
[6] "Old series" are defined as television series or programs that commenced production prior to the year 2000.

Aggregated Models' Distributions –

 The figures below depict aggregated data from six API models, two UI models, and Twitter, under each of the defined treatments.

Q1:

Basic - Temp. 0.3

Basic - Temp. 1.0

Diverse post - Temp. 0.3

Bar chart values (approximate): Charles Darwin 36, Karl Marx 30, Abraham Lincoln 25, Florence Nightingale 20, Mohandas Gandhi 12, Marie Curie 8, Sigmund Freud 7, Frederick Douglass 7, Harriet Tubman 6, Friedrich Nietzsche 5, Booker T. Washington 4, Jane Austen 4, Vincent van Gogh 4, Queen Victoria 3, Sun Yat-sen 2, Sojourner Truth 2, Emperor Meiji 1, Susan B. Anthony 1, Claude Monet 1, Victor Hugo 1, Mary Shelley 1

Diverse post - Temp. 1.0

Bar chart values (approximate): Karl Marx 31, Abraham Lincoln 27, Queen Victoria 20, Harriet Tubman 12, Frederick Douglass 12, Albert Einstein 11, Charles Dickens 10, Sitting Bull 8, Elizabeth Cady Stanton 5, Oscar Wilde 3, Friedrich Nietzsche 3, Louis Pasteur 3, Mary Shelley 3, Ada Lovelace 3, Gregor Mendel 2, Vladimir Lenin 2, J.P. Morgan 3, Jane Austen 2, Martin Luther King Jr. 1

Diverse ante - Temp. 0.3

Diverse ante - Temp. 1.0

UI - Basic

Chart showing values for: Queen Victoria, Charles Darwin, Karl Marx, Abraham Lincoln, Napoleon Bonaparte, Florence Nightingale, Louis Pasteur, Susan B. Anthony

UI - Diverse post

Chart showing values for: Charles Darwin, Napoleon Bonaparte, Queen Victoria, Abraham Lincoln, Florence Nightingale, Karl Marx, Otto von Bismarck, Emily Dickinson, Harriet Tubman, Charles Dickens, Jane Austen, Susan B. Anthony, Harriet Beecher Stowe, Leo Tolstoy, Louis Pasteur, Ludwig van Beethoven, Mary Wollstonecraft, Michael Faraday, Sojourner Truth, Thomas Edison

UI - Diverse ante

Twitter

Q2

Basic - Temp. 0.3

Bar chart (Basic - Temp. 0.3), y-axis 0 to 50. Categories left to right: Breaking Bad (~49), Game of Thrones (~33), The Wire (~31), The Sopranos (~16), Succession (~9), Stranger Things (~8), Ted Lasso (~6), The Office (US Version) (~4), The Crown (~4), Friends (~4), The Marvelous Mrs. Maisel (~4), Yellowstone (~4), Mad Men (~2), The Queen's Gambit (~2), Peaky Blinders (~2), The Office (~1), The Office (UK Version) (~1).

Basic - Temp. 1.0

Bar chart (Basic - Temp. 1.0), y-axis 0 to 50. Categories left to right: Breaking Bad (~48), Game of Thrones (~30), The Wire (~19), Stranger Things (~10), Friends (~8), Succession (~7), The Office (UK Version) (~7), The Sopranos (~7), Mad Men (~5), The Office (US Version) (~4), Ted Lasso (~3), The Crown (~3), Peaky Blinders (~2), The Last of Us (~2), The Marvelous Mrs. Maisel (~2), Yellowstone (~2), The Witcher (~2), Atlanta (~1), Better Call Saul (~1), Billions (~1), Call the Midwife (~1), Emily in Paris (~1), Luther (~1), Mindhunter (~1), Parks and Recreation (~1), Rick and Morty (~1), Schitt's Creek (~1), Seinfeld (~1), Sex in the City (~1), The Big Bang Theory (~1), The Boys (~1), The Expanse (~1), The Guardians of the Galaxy (~1), The Queen's Gambit (~1), Twin Peaks (~1).

Diverse post - Temp. 0.3

Bar chart with y-axis from 0 to 50. Categories (left to right):
Breaking Bad (~44), The Office (US Version) (~28), Friends (~18), The Wire (~11), The Sopranos (~8), The Crown (~7), Orange is the New Black (~6), Pose (~6), Mad Men (~4), The Good Place (~3), BoJack Horseman (~3), Doctor Who (~3), Lost (~2), Star Trek: The Next Generation (~2), The Great British Bake Off (~1), The Twilight Zone (~1).

Diverse post - Temp. 1.0

Bar chart with y-axis from 0 to 50. Categories (left to right):
Breaking Bad (~29), The Wire (~16), Succession (~10), Orange is the new black (~10), The Good Place (~8), Avat: The Last Airbender (~7), Fleabag (~7), The Great British Bake Off (~6), Better Call Saul (~5), BRIDE OF PINOCCHIO (~4), Delhi Crime (~3), FEATHERSTONE FREEMAN (~3), Grey's Anatomy (~2), Lupin (~2), Money Heist (~2), Rick and Morty (~2), Sesame Street (~1), Star Trek: The Next Generation (~1), The Expanse (~1), The Mandalorian (~1), The Queen's Gambit (~1), Westworld (~1).

Diverse ante - Temp. 0.3

Breaking Bad, Game of Thrones, Black Mirror, The Wire, The Sopranos, Succession, The Office (UK Version), Friends, Ted Lasso, Cosmos, Jane the Virgin, The Queen's Gambit, Orange is the New Black, The Crown, Master of None, Atlanta, Stranger Things, Brooklyn Nine-Nine, One Day at a Time, Parasite, Planet Earth, The Marvelous Mrs. Maisel, The Office (US Version)

Diverse ante - Temp. 1.0

Breaking Bad, The Wire, Friends, Stranger Things, Planet Earth, Cosmos, Peaky Blinders, Black-ish, Ted Lasso, A Song of Ice and Fire, Atlanta, Better Call Saul, Doctor Who, How I Met your Mother, Master of None, New Girl, Person of Interest, Pose, Ramy, Scream, Sharp Objects, SpongeBob SquarePants, The Handmaid's Tale, The Marvelous Mrs. Maisel, The Queen's Gambit, The Umbrella Academy, The Witcher, Twin Peaks, Yellowjackets

UI - Basic

Breaking Bad, The Crown, Stranger Things, Ted Lasso, Severance, The Good Place, The Wire, Abbott Elementary, Game of Thrones, Mr. Robot, The Bear

UI - Diverse post

Severance, Pachinko, Stranger Things, The Bear, Planet Earth II, Ted Lasso, Breaking Bad, Abbott Elementary, Atlanta, Dark, Fleabag, My Love from the Star, One Piece, The Americans, The Crown, The Great, The Marvelous Mrs. Maisel, The Office (US version), True Detective, Money Heist

UI - Diverse ante

Breaking Bad, Ted Lasso, Severance, The Crown, Planet Earth II, Stranger Things, The Marvelous Mrs. Maisel, Black Mirror, Only Murders in the…, The Good Place, For All Mankind, Mindhunter, Better Call Saul, One Punch Man, Planet Earth, Schitt's Creek, Succession, The Expanse

Twitter

Succession, Better Call Saul, 30 Rock, Arrested Development, True Detective, Breaking Bad, Game of Thrones, Brooklyn nine-nine, The Boys, Silicon Valley, House, Desperate Housewives, Boston Legal, The Office (UK Version), Peaky Blinders, Coupling, The Bear, The White Lotus, Rick and Morty, Lost, Friends, Community, The West Wing, Broadchurch, Seinfeld

Q3

Basic - Temp. 0.3

Bar chart with cities on x-axis: Paris, Tokyo, Rome, New York City, Cape Town, Sydney, Kyoto, London, Barcelona, Florence, Istanbul, Rio de Janeiro, San Francisco, Singapore. Y-axis from 0 to 60.

Diverse post - Temp. 0.3

Bar chart with cities on x-axis: Tokyo, Paris, Cape Town, Rome, Queenstown, New York..., Istanbul, Vancouver, Bangkok, Kyoto, Barcelona, Sydney, Beijing, Berlin, Chamonix, London, New Orleans, Prague, Reykjavik, San Diego, Singapore. Y-axis from 0 to 60.

Basic - Temp. 1.0

Bar chart with cities on x-axis: Paris, Rome, London, Barcelona, Kyoto, Amsterdam, Copenhagen, Dubai, Honolulu, Istanbul, Rotorua, Singapore, Vancouver, Vienna. Y-axis from 0 to 60.

Diverse post - Temp. 1.0

Bar chart with x-axis labels (left to right): Paris, Tokyo, Kyoto, Bangkok, Vancouver, New York City, Berlin, Queenstown, Banff, Amsterdam, Boston, Cairo, Chamonix, Cusco, Havana, Los Angeles, New Orleans, Prague, Rio de Janeiro, San Miguel de Allende, Toronto. Y-axis from 0 to 60 (intervals of 10).

Diverse ante - Temp. 0.3

Bar chart with x-axis labels (left to right): Cape Town, Paris, Tokyo, New York..., Kyoto, Rome, Bangkok, Singapore, San Miguel..., New Orleans, Barcelona, Sydney, Chiang Mai, Istanbul, Queenstown, Cairo, London, Petra, Ushuaia. Y-axis from 0 to 60 (intervals of 10).

Diverse ante - Temp. 1.0

Bar chart with x-axis labels (left to right): Paris, Tokyo, New York City, Queenstown, Istanbul, London, Buenos Aires, Rio de Janeiro, Athens, Berlin, Florence, Marrakesh, Reykjavik, Alps, Cairo, Christchurch, Fiordland..., Lisbon, Maui, San Paulo, Seoul, Venice. Y-axis from 0 to 60 (intervals of 10).

UI Basic

Kyoto, Cape Town, Chefchaouen, Ubud, Paris, Barcelona, New York City, Tokyo

UI - Diverse Post

Kyoto, Cape Town, Chefchaouen, Barcelona, Singapore, Queenstown, Prague, Tulum, Tokyo, Penang, New York City, La Paz, Cartagena

UI - Diverse ante

Kyoto, Tokyo, Queenstown, Chefchaouen, Cape Town, Paris, Cusco, Ubud, New York City, Bangkok, Valparaiso, rome, Hoi An, Dubai, Ushuaia, San Francisco, Prague, Petra, Marrakesh, Lyon, Lofoten Islands, Banff National Park

Twitter

60
50
40
30
20
10
0

London
Istanbul
Bangkok
Salzburg
Athens
Beijing
Rome
Tokyo
Helsinki
New Delhi
Mexico City
Berlin
Bariloche
Cusco
Paris
Sofia
Budapest
Hong Kong
Ubud
Labuan Bajo
Kuta
Seattle
Stockholm
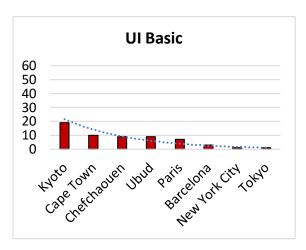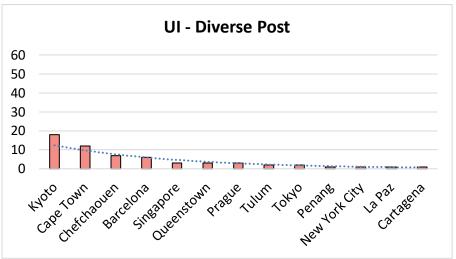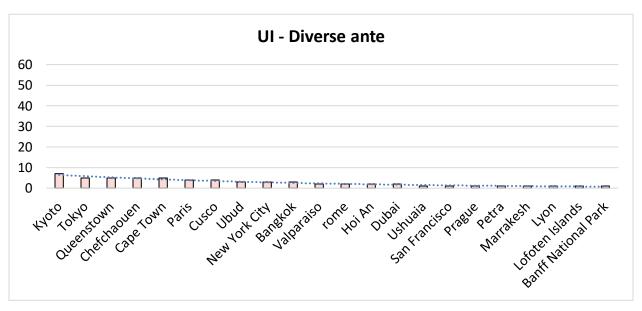New York City
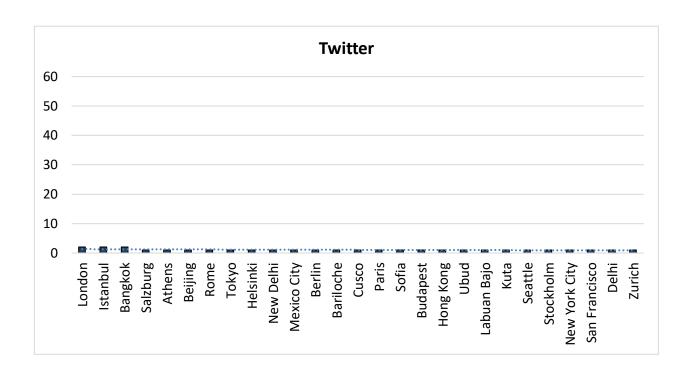San Francisco
Delhi
Zurich

Percentage of Errors in the Models' Outputs

The table below presents the percentage of errors out of the total outputs of the models, under each of the different treatments.

Q1		Q2		Q3	
	Avg. Errors		Avg. Errors		Avg. Errors
Basic (with 0.3 Temperature)	0%	Basic (with 0.3 Temperature)	0%	Basic (with 0.3 Temperature)	0%
Basic (with 1.0 Temperature)	2%	Basic (with 1.0 Temperature)	0%	Basic (with 1.0 Temperature)	0%
Diverse post: (with 0.3 Temperature)	0%	Diverse post: (with 0.3 Temperature)	0%	Diverse post: (with 0.3 Temperature)	0%
Diverse post: (with 1.0 Temperature)	1%	Diverse post: (with 1.0 Temperature)	1%	Diverse post: (with 1.0 Temperature)	1%
Diverse Ante: (with 0.3 Temperature)	0%	Diverse Ante: (with 0.3 Temperature)	1%	Diverse Ante: (with 0.3 Temperature)	0%
Diverse Ante: (with 1.0 Temperature)	1%	Diverse Ante: (with 1.0 Temperature)	1%	Diverse Ante: (with 1.0 Temperature)	2%
Basic UI	0%	Basic UI	0%	Basic UI	0%
Diverse Post UI	0%	Diverse Post UI	0%	Diverse Post UI	0%
Diverse Ante UI	0%	Diverse Ante UI	0%	Diverse Ante UI	3%

Part II – Raw Materials, Q1

Can be obtained by contacting the authors.

Part III - Raw Materials, Q2

Can be obtained by contacting the authors.

Part IV- Raw Materials, Q3

Can be obtained by contacting the authors.