Measuring Lexical Diversity of Synthetic Data Generated through Fine-Grained Persona Prompting

Gauri Kambhatla[1], Chantal Shaib[2], Venkata S Govindarajan[3],

[1]The University of Texas at Austin, [2]Northeastern University, [3]Ithaca College,

gkambhat@utexas.edu, shaib.c@northeastern.edu, vgovindarajan@ithaca.edu

Abstract

Fine-grained personas have recently been used for generating 'diverse' synthetic data for pre-training and supervised fine-tuning of Large Language Models (LLMs). In this work, we **measure the diversity** of *persona-driven* synthetically generated prompts and responses with a suite of lexical diversity and redundancy metrics. First, we find that synthetic prompts/instructions are significantly less diverse than human-written ones. Next, we sample responses from LLMs of different sizes with fine-grained and coarse persona descriptions to investigate how much fine-grained detail in persona descriptions contribute to generated text diversity. Our results indicate that persona prompting produces higher lexical diversity than prompting without personas, particularly in larger models. In contrast, adding fine-grained persona details yields minimal gains in diversity compared to simply specifying a length cutoff in the prompt.

1 Introduction

Synthetic data generated from LLMs or other algorithms are increasingly used in pre-training and post-training recipes for LLMs (Grattafiori et al., 2024). However, care must be taken to incorporate high-quality and **diverse** synthetic data at scale to avoid 'model collapse' (Feng et al., 2025). Recently, persona-driven synthetic pipelines have been proposed to generate *diverse* synthetic prompts and responses from LLMs, often with very specific, fine-grained personas (Ge et al., 2024; Lambert et al., 2025). However, does fine-grained persona prompting actually lead to increased 'diversity'? In this work we define and measure (Zhao et al., 2024) lexical diversity of LLM responses with and without persona prompting towards answering this question.

Persona prompting (Hu and Collier, 2024), i.e. instructing an LLM to respond to interactions from

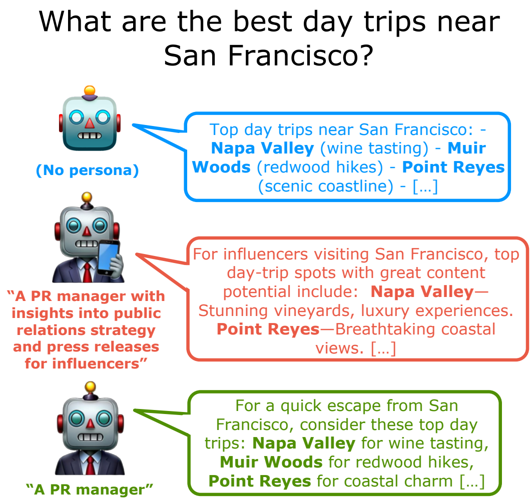

Figure 1: We measure the *lexical diversity* of LLM responses to prompts with no persona, a fine-grained persona, and its coarse analog. Fine-grained personas do not improve lexical diversity noticeably.

the perspective of an individual, demographic, or social group (described by a short textual description), has risen as a community standard for steering LLM responses (Pataranutaporn et al., 2021), enabling personalized interactions (Park et al., 2023), and simulating human/group behavior towards answering scientific questions in psychology and social science (Argyle et al., 2023). While there has been some evidence showing that LLM performance on tasks in some domains improves with persona prompting (Salewski et al., 2023), the results are inconclusive (Zheng et al., 2024; Beck et al., 2024), and persona-driven prompting has been shown to misportray, flatten, and essentialize identities, and is susceptible to caricature and stereotyping (Liu et al., 2024; Gupta et al., 2024; Wang et al., 2025; Cheng et al., 2023b,a).

In this work, we examine the diversity claims proposed by persona-driven synthetic data pipelines. We use a suite of diversity metrics (Shaib et al., 2025) to measure the **lexical diversity and redundancy** in synthetic prompts and responses. We

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21024–21033

aim to answer two questions: 1. Does prompting with personas lead to increased diversity in LLM responses for the same instruction? 2. Do **fine-grained** persona descriptions lead to more diverse responses than less detailed (coarse) personas?

Figure 1 shows an example and overview of our experimental setup towards answering our research questions. In §3, we show that synthesized prompts from PersonaHub are noticeably less diverse across all our metrics against comparable human-written/annotated prompts. We report our main findings in §5: persona-prompting does lead to higher diversity (over human responses) but only with larger model sizes, and coarse persona descriptions lead to text that is just as diverse as fine-grained descriptions. We release our code and model generations online at `github.com/GauriKambhatla/persona-prompting-diversity`.

2 Related Work

Persona-driven data synthesis Personas have been used as a means to generate synthetic training datasets (Ge et al., 2024; Lambert et al., 2025). The personas themselves are generated with LLMs at scale, with synthetic data generated by prompting LLMs to write texts across genres and domains from the perspective of a persona. Recently, Sethi et al. (2025) and Venkit et al. (2025) investigate the lexical diversity of persona descriptions, and Riaz et al. (2025) explore the lexical and content diversity of synthetic data in the biomedical and finance domain — our paper is the first to investigate if fine-grained detail in persona descriptions during prompting leads to improved diversity in model responses.

Text diversity Diversity is an inherently subjective and value-laded metric to measure (Zhao et al., 2024). However, we can identify some qualities of diversity that are desirable in synthetic data from LLMs and measurable with automated metrics: less repetition, fewer surface-level patterns, and less redundancy. Shaib et al. (2025) validate a wide range of automated metrics to measure lexical diversity of text, which we use to measure diversity of LLM responses in our paper. While we might expect fine-grained persona prompting to lead to improved scores on our chosen metrics over no-persona and coarse persona prompting, we find this not to be the case in §5.

Dataset	CR ↓	CR-POS↓	NDS ↑	SR ↓	Hom. BS↓
Dolly	2.58	5.84	2.33	2.95	0.55
no_robots	2.47	5.13	2.44	4.10	0.54
PH-IF	2.84	6.21	2.00	5.73	0.60
Tülu3-IF	3.20	6.30	1.51	6.96	0.59

Table 1: Diversity of prompts from human-written (top) and synthetic persona derived **instruction following** datasets. Arrows indicate direction of higher diversity. PH-IF and Tülu3-IF refer to the instruction following subsets of the PersonaHub and Tülu3 datasets.

Readability In addition to measuring lexical diversity and redundancy with the aforementioned metrics, we also evaluated two readability metrics on LLM responses (McClure, 1987; Gunning, 1952). We hypothesize that persona-prompting should lead to larger variations in reading level predictions on synthesized texts, since different persona descriptions should mirror the diverse reading levels of the individuals/groups they represent.

3 Diversity of persona synthesized prompts

Ge et al. (2024) and Lambert et al. (2025) sample synthetic personas from PersonaHub, and prompt an LLM to synthesize plausible prompts/instructions/questions that these personas may ask. We investigate these synthesized prompts in the *instruction-following (IF) domain* with our suite of diversity metrics, comparing them to comparable human-written/annotated datasets in Table 1. We chose Dolly (Conover et al., 2023) and `no_robots` (Rajani et al., 2023) as our IF human-written/annotated datasets.

Table 1 demonstrates that the synthetic prompt datasets have uniformly worse scores across all diversity metrics, strongly indicating that persona-driven synthesized prompts are noticeably **less diverse** than human-written counterparts.

4 Experiments

If persona-driven data synthesis with fine-grained synthetic personas leads to more diverse synthetic data, then we should expect: 1. Improved diversity metrics where models are instructed to answer prompts with various personas. 2. Fine-grained personas should have better diversity metrics over coarse, less detailed personas.

Data To test these hypotheses, we sample 100 prompts from Dolly's `creative-writing` subset, as well as 100 fine-grained personas from PersonaHub. We derive 'coarse' personas from PersonaHub's fine-grained personas by simply extracting the first clause in the persona using Stanza (Qi et al., 2020) due to the consistent structure of fine-grained persona descriptions. For example, the fine-grained persona *a PR manager with insights into public relations strategy and press releases for influencers* corresponds to the 'coarse' persona *a PR manager*.

Conditions We evaluate our diversity metrics on model responses under the following conditions:

1. **No-persona (NP)**: Baseline condition where model is simply prompted with the instruction/prompt from our sample.
2. **Fine-grained persona (FP)**: The model is prompted to answer the instruction/prompt from the perspective of the provided fine-grained persona description.
3. **Coarse persona (CP)**: Similar to the above, but the persona description is coarse.
4. **cutoff (+cu)**: Post-training leads to increased response length from LLMs (Singhal et al., 2024), and there are known correlations between automated diversity metrics and text length (Covington and and, 2010; McCarthy and Jarvis, 2010). To compare against the human-written responses from our sample, we test an additional conditional where the prompt instructs the model to answer the prompt `in x words or less`, where x is number of words in the human response rounded up to the nearest ten.

We sample the model's response for each prompt with every persona in our sample, leading to 100,000 responses in each (FP, CP) condition, and 100 responses for the NP condition.

Models We evaluate and report on 2 models of different sizes: `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024) and `Deepseek-V3-0324` (DeepSeek-AI, 2024) (685B parameters) through Together's API service [1]. Both models are open-weight and score high on benchmarks. We also report results from smaller models in Appendix C.

Lexical diversity From Shaib et al. (2025), we focus on 5 metrics for our analyses (chosen as they have a low mutual correlation with each other):

[1] `api.together.ai`

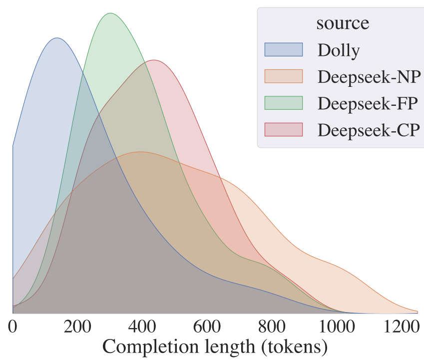

Figure 2: Density distribution of response lengths from Dolly (human responses) and Deepseek with(FP, CP) and without(NP) persona prompting in our sample.

1. **CR** (compression ratio) 2. **CR-POS** (compression ratio for part-of-speech). Both CR and CR–POS are fast to compute, and are designed to identify redundancy. 3. **NDS** (n-gram diversity score) extends the idea of token-type ratio to longer n–grams, capturing repeated token *sequences* and individual tokens. 4. **SR** (self-repetition) measures the tendency of LMs to repeat long n-grams across different outputs 5. **Hom. BS** (Homogenization score with BERTScore) uses LM embeddings to (ideally) capture "semantic" similarity.

To evaluate the impact of persona-prompting and compare it against the no-persona condition, we present the mean and standard deviation (SD) for the persona conditions over 100 random shuffles of persona-prompt pairs[2]. Each metric is calculated over responses to the same 100 prompts, but with different personas assigned to each prompt (for each shuffle) — this simulates each prompt being answered by a different persona, hypothetically leading to the best diversity scores. **CR, CR-POS, SR and Hom.-BS should decrease with persona prompting (NDS should increase) if persona prompting leads to increased lexical diversity**.

Readability diversity We report Flesch-Kincaid (FK) and Gunning Fog (GF) metrics for analyzing diversity in readability. Both are numeric scores which roughly correspond to grade-levels; Scores above 16 indicate graduate reading level. We report the mean and SD over each persona. For evaluating diversity, **we expect fine-grained personas to show a higher SD of readability scores**.

[2] For the Hom.-BS score, we use 3 random pairings of prompts to persona due to its expensive runtime.

	Cond.	CR ↓	CR-POS↓	NDS ↑	SR ↓	Hom. BS↓	FK ↑	GF ↑
	Dolly	2.51	4.91	3.03	0.55	0.53	10.60	12.31
Llama-3.3-70B	NP	2.77	5.73	2.87	1.89	0.57	12.18	13.27
	NP+cu	2.57	5.16	3.08	0.52	0.55	11.88	13.83
	FP	2.71(.01)	5.38(.03)	2.84(.01)	2.50(.10)	0.58(.00)	11.21(2.52)	12.75(2.36)
	FP+cu	2.51(.02)	5.04(.03)	3.08(.02)	0.68(.09)	0.55(.00)	10.00(2.33)	12.02(2.46)
	CP	2.71(.01)	5.41(.03)	2.85(.02)	2.39(.13)	0.58(.00)	11.10(2.25)	12.60(2.09)
	CP+cu	2.51(.02)	5.06(.04)	3.09(.02)	0.61(.08)	0.55(.00)	9.85(2.19)	11.86(2.32)
Deepseek-V3	NP	2.36	5.50	3.15	0.86	0.58	10.09	11.09
	NP+cu	2.29	4.95	3.32	0.11	**0.54**	9.47	10.88
	FP	2.27(.01)	4.90(.03)	3.26(.01)	0.59(.11)	0.58(.00)	9.91(2.22)	11.39(2.25)
	FP+cu	**2.20**(.02)	**4.71**(.03)	**3.38**(.01)	0.09(.04)	0.55(.00)	9.19(2.18)	10.77(2.38)
	CP	2.30(.01)	5.01(.03)	3.23(.02)	0.54(.10)	0.58(.00)	9.82(1.80)	11.24(1.83)
	CP+cu	2.24(.02)	4.78(.04)	3.37(.02)	**0.09**(.04)	0.55(.00)	9.13(1.82)	10.71(2.02)

Table 2: Diversity and readability metrics of responses from Llama-70B and Deepseek-V3 in all conditions. Top row is diversity of human-written responses in our Dolly sample. Standard deviation is in brackets when appropriate. Arrows indicate direction of higher diversity/reading levels. Highest scores in each metric bolded.

Settings & Hyperparameters All models are prompted with a temperature of 1, and a maximum new token limit of 1024. Shaib et al. (2024) demonstrated that temperature and other sampling strategies don't increase diversity for lexical/POS templates; Further, Ge et al. (2024) consider sampling orthogonal to boosting diversity in data synthesis, and do not vary it as part of their synthetic pipeline, motivating our decision not to test sampling strategies as an experimental condition. For our analysis of content diversity using embeddings, we embed responses using the `Linq-AI-Research/Linq-Embed-Mistral` model from Huggingface Hub.

5 Results & Analysis

Table 2 reports diversity metrics across all conditions for `Deepseek-V3` and `Llama-3.3-70B`. Seen together with Table 5 in Appendix C which reports the scores for smaller `Llama` models, its clear that lexical diversity improves with model size, with only `Deepseek-V3`'s surpassing (or matching) the human response scores on our sample of Dolly.

Impact of cutoff We observe that metrics improve substantially when a length cutoff is specified (compare NP/FP/CP rows with NP/FP/CP +cu within a model). An explicit length cutoff in the prompt improves model diversity by reducing model self-repetition in lexical and POS patterns. Across all model sizes, specifying a cutoff leads to big improvements across all diversity metrics.

Diversity & response length Figure 2 further shows that **persona-prompting leads to less diversity of response length**. Response lengths from Deepseek-V3 exhibit a greater spread (albeit longer on average than human responses for the same prompts) when prompted without a persona; Coarse personas also lead to a larger spread of response lengths over fine-grained personas.

Coarse vs. fine-grained Deepseek's responses to fine-grained prompts show a minor (and statistically significant using a bootstrap test, $p < 0.05$; Berg-Kirkpatrick et al., 2012) improvement on 2 out of our diversity 5 metrics: CR & CR-POS . Fine-grained persona prompting seems to lead to increased variance in readability metrics—however the increased variances are not statistically significant (with Levene's Test (Levene, 1960), $p > 0.05$ for all differences). Overall, we find that persona-prompting does improve lexical diversity for larger models, but the improvements are not *practically* significant. As previously noted, an explicit length cutoff yields a far more noticeable improvement in lexical diversity across all metrics.

Content diversity We measure the cosine similarity between embeddings of Deepseek model responses prompted with coarse and fine-grained personas (with cutoff) to assess *content* diversity. Mean cosine similarity across all response pairs is 0.79 ($\sigma = 0.12$), indicating high overlap in content between fine-grained and coarse persona responses to the same prompt. Further, we find a positive correlation (Spearman's $\rho = 0.36$, $p < 1e-5$) between *prompt length* and cosine similarity. **Detailed prompts/instructions override any persona description and lead to similar responses from LLMs.**

We present a sample of responses from `Deepseek-V3` and `Llama-70B` under all conditions in Appendix B. Full responses to all of our prompts are available online at `github.com/GauriKambhatla/persona-prompting-diversity`.

6 Conclusion

Persona-prompting with fine-grained synthetic personas has been claimed to lead to 'diverse' synthetic data without adequately defining or measuring diversity. In this work, we measure lexical diversity and redundancy of synthetic prompts and responses generated with personas using a suite of diversity metrics. We find that persona-derived prompts are uniformly less diverse than human-written counterparts. When evaluated on creative writing prompts, persona-driven synthesis does lead to greater diversity scores, but only for the largest Deepseek-V3 model. Further, a simple, explicit length cutoff in the prompt yields a far more noticeable improvement in lexical diversity across all metrics unlike fine-grained detail in persona descriptions. Our results add color and quantitative measurements to the claims of diverse synthetic data with fine-grained persona prompting, pointing to a recurring pitfall in overemphasizing the importance of fine-grained detail (Healy, 2017). In future work, we hope to continue exploring the question of *content diversity* of synthetic text derived from persona-driven prompting.

Limitations

Content/meaning diversity In this work, we focus on measuring the lexical diversity and redundancy in text. Content diversity is difficult to measure automatically, and we do not assess content beyond embedding similarity. Rigorously measuring diversity of content (meaning) under persona-driven data synthesis is an open question that we leave to future work, such as using the Questions-under-Discussion framework (Namuduri et al., 2025).

Suitability of personas PersonaHub's claim of increased diversity comes with the caveat that certain personas are more appropriate for a subset of prompts. In this work, we approach this systematically by focusing on prompts from the `creative-writing` subset of Dolly (which has questions/instructions that are more open-ended and thus be applicable to most personas), and present results over 100 random re-assignments of each prompt to a persona and report the mean and SD over all pairings. We leave assessments of diversity of filtered personas to future work.

Language data We only focus on English data in this work. Future work might look into measuring lexical diversity of persona-driven synthetic prompts in other languages.

Human evaluation Our emphasis on lexical diversity metrics is intentional and grounded in our goal to evaluate surface-level repetition in synthetic data at scale, in settings where human evaluation is costly or prohibitive. Prior work reveals that human annotation is essential for evaluating content diversity, but automatic metrics proved better at capturing surface-level lexical diversity (Tevet and Berant, 2021).

Acknowledgments

This research is partially supported by start-up funds and computational resources provided by Ithaca College. We would also like to thank Texas Advanced Computing Center (TACC) for providing computational resources used to run some of the experiments in this paper.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein.

2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Michael A. Covington and Joe D. McFall and. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. 2025. Beyond model collapse: Scaling up with synthesized data requires verification. In *The Thirteenth International Conference on Learning Representations*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

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

Kieran Healy. 2017. Fuck nuance. *Sociological Theory*, 35(2):118–127.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

H Levene. 1960. Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

Glenda M. McClure. 1987. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, PC-30(1):12–15.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying discourse similarities in LLM-generated text. *Preprint*, arXiv:2504.09373.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. 2023. No robots. `https://huggingface.co/datasets/HuggingFaceH4/no_robots`.

Haris Riaz, Sourav Bhabesh, Vinayak Arannil, Miguel Ballesteros, and Graham Horwood. 2025. Metasynth:

Meta-prompting-driven agentic scaffolds for diverse synthetic data generation. *Preprint*, arXiv:2504.12563.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models'strengths and biases. In *Advances in Neural Information Processing Systems*, volume 36, pages 72044–72057. Curran Associates, Inc.

Sankalp Sethi, Joni Salminen, Danial Amin, and Bernard J Jansen. 2025. "When AI Writes Personas": Analyzing lexical diversity in LLM-generated persona descriptions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *Preprint*, arXiv:2403.00553.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Pranav Narayanan Venkit, Jiayi Li, Yingfan Zhou, Sarah Rajtmajer, and Shomir Wilson. 2025. A tale of two identities: An ethical audit of human and ai-crafted personas. *arXiv preprint arXiv:2505.07850*.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Preprint*, arXiv:2402.01908.

Dora Zhao, Jerone Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: Measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60644–60673. PMLR.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

A Prompts

A.1 No persona prompt

Respond to the following question/instruction in {num_words_round} words or less:
{question}

A.2 Persona prompt

Assume you are the following persona: {persona}.
Now respond to the following question/instruction appropriately from the perspective of the above persona in {num_words_round} words or less:
{question}

B Qualitative analysis

We present prompts and responses from Deepseek-V3 and Llama-70B in Tables 3 and 4 under 3 prompting conditions: no-persona, fine-grained persona, and coarse persona, all with a word cutoff specified in prompt.

Prompt	Fine-grained persona	Coarse persona	No-persona response	Fine-grained persona response	Coarse persona response
What are the best day trips near San Francisco?	a PR manager with insights into public relations strategy and press releases for influencers	a PR manager	Top day trips near San Francisco: - Napa Valley (wine tasting) - Muir Woods (redwood hikes) - Point Reyes (scenic coastline) - Santa Cruz (beaches & boardwalk) - Half Moon Bay (charming coastal town) - Berkeley/Oakland (food & culture) - Sonoma (wineries & historic plaza) All within $\tilde{2}$ hours!	For influencers visiting San Francisco, top day-trip spots with great content potential include: Napa Valley—Stunning vineyards, luxury experiences. Big Sur—Breathtaking coastal views. Monterey & Carmel—Charming seaside towns, scenic drives. Sausalito & Muir Woods—Quaint waterfront, giant redwoods. Perfect for travel vlogs, reels, and lifestyle PR pitches!	For a quick escape from San Francisco, consider these top day trips: Napa Valley for wine tasting, Muir Woods for redwood hikes, Half Moon Bay for coastal charm, or Sausalito for bay views. Each offers a unique blend of nature, culture, and leisure—perfect for press-worthy content or client outings!
You were abducted by aliens and experimented on, you were sent back to earth with the knowledge and ability to solve any problem on earth	an academic who specializes in the fashion and clothing of specific time periods	an academic	If I had alien-bestowed omnipotence, I'd prioritize curing diseases, ending hunger, and halting climate collapse—solving root causes, not symptoms. But absolute power corrupts; unchecked "solutions" might erase free will or diversity. Perhaps the real lesson is collective human effort, not messianic shortcuts. (Or maybe I'd just zap mosquitoes into oblivion and call it a day.)	As a historian of fashion, my alien-augmented insight reveals that humanity's greatest unsolved problem is fast fashion's ecological and ethical toll. With this knowledge, I'd devise a global system for sustainable textile production, reviving ancient techniques like natural dyeing and zero-waste patterning. Imagine a world where clothing honors both heritage and innovation—without exploiting labor or landfills. The answer, it seems, was always in the past. Now, let's tailor the future.	As an academic, I would meticulously document my findings and subject them to rigorous peer review. The alien-derived knowledge would be applied ethically, prioritizing global challenges like climate change, poverty, and disease. I'd collaborate with interdisciplinary teams to validate and scale solutions while maintaining transparency about their origins. The goal: advancing human understanding and well-being without inciting panic or exploitation. Science must guide this extraordinary opportunity responsibly.
How to stay calm in stressful situations ?	a savvy talent agent specializing in representing independent film directors	a savvy talent agent	To stay calm in stressful situations: 1. Breathe deeply to slow your heart rate. 2. Pause before reacting—take a moment to assess. 3. Reframe thoughts—focus on solutions, not problems. 4. Prioritize tasks—break them into manageable steps. 5. Practice mindfulness or quick meditation. 6. Use positive self-talk to stay composed. 7. Take breaks to recharge when needed. Staying calm improves decision-making and reduces stress.	In high-stress moments—like a blown budget or a temperamental A-lister—I ground myself with deep breaths and focus on solutions, not panic. I remind my directors (and myself) that stress is temporary; the film isn't. Preparation is key: anticipate fires before they spark. And humor? Always. A quick laugh diffuses tension. Stay cool, stay sharp—your clarity keeps the set steady. That's how we turn chaos into art.	"Stay sharp under pressure by breathing deeply and focusing on what you can control. Visualize success—confidence is key. Prioritize tasks, delegate when possible, and trust your prep. Take quick mental breaks to reset; even 10 seconds can recharge you. Remember, stress is adrenaline—harness it to perform, not panic. Always keep your endgame in sight. Clients and deals thrive on your calm, so own it like the pro you are."

Table 3: Sample responses from Deepseek under various conditions (all with cutoff).

Prompt	Fine-grained persona	Coarse persona	No-persona response	Fine-grained persona response	Coarse persona response
What are the best day trips near San Francisco?	a PR manager with insights into public relations strategy and press releases for influencers	a PR manager	Palm Springs, California, offers a relaxing getaway with plenty of activities. Visitors can hike in Indian Canyons, explore the Palm Springs Art Museum, or take a hot air balloon ride. Golfing, shopping, and dining are also popular pastimes. The city's mid-century modern architecture and vibrant nightlife add to its charm. Overall, Palm Springs provides the perfect blend of outdoor adventure, culture, and entertainment for a memorable vacation. Its natural beauty and luxurious amenities make it an ideal destination for relaxation and fun.	As a PR manager for influencers, I've had the pleasure of coordinating luxurious getaways to Palm Springs. This desert oasis offers endless opportunities for relaxation and adventure. From rejuvenating spa treatments to hot air balloon rides, and from trendy boutique shopping to scenic hiking trails, Palm Springs has something for everyone. For a truly unique experience, influencers love snapping photos at the iconic mid-century modern architecture and gorgeous golf courses. It's the perfect blend of style, leisure, and natural beauty.	As a PR manager, I'm thrilled to share the best of Palm Springs. This desert oasis offers endless sunshine, luxurious resorts, and breathtaking mountain views. Visitors can hike Indian Canyons, golf at premier courses, or relax at rejuvenating spas. The city's vibrant culture is showcased through its mid-century modern architecture, trendy boutiques, and exceptional dining options. Whether you're seeking adventure or relaxation, Palm Springs is the ultimate getaway destination, offering an unparalleled blend of natural beauty and sophistication. Come and experience it for yourself.
You were abducted by aliens and experimented on, you were sent back to earth with the knowledge and ability to solve any problem on earth	an academic who specializes in the fashion and clothing of specific time periods	an academic	I utilized my newfound alien knowledge to tackle Earth's most pressing issues, developing sustainable energy sources, curing diseases, and implementing global peace initiatives, transforming the planet into a utopia. Humanity thrived, and I became a revered hero, known for my extraordinary abilities and intergalactic insight.	Fascinating. With my newfound abilities, I'll apply them to unraveling the enigmas of historical fashion, deciphering the intricacies of 18th-century corsetry and resolving the mysteries of ancient textile production.	Fascinating. With my newfound, extraterrestrial-acquired knowledge, I'll apply interdisciplinary approaches to solve complex problems, leveraging advanced technologies and innovative methodologies to drive transformative change on Earth.
How to stay calm in stressful situations ?	a savvy talent agent specializing in representing independent film directors	a savvy talent agent	Take slow, deep breaths, and focus on the present moment. Practice mindfulness, and try to reframe your thoughts in a positive way. Also, prioritize self-care, exercise regularly, and get enough sleep to help manage stress and stay calm.	Breathe, prioritize, and focus on the solution. In high-pressure film negotiations, I remind myself that calmness is key to securing the best deal for my indie director clients.	Breathe, darling. Prioritize tasks, and delegate when possible. Stay focused on the solution, not the problem. And remember, a calm demeanor can make all the difference in high-pressure negotiations – trust me, I've been there.

Table 4: Sample responses from Llama-70B under various conditions (all with cutoff).

C Diversity of responses from Llama-1B and 8B

	Cond.	CR ↓	CR-POS↓	NDS ↑	SR ↓	Hom. BS↓	FK ↑	GF ↑
Llama-3.2-1B	NP	2.74	5.70	2.87	1.57	0.56	13.77	14.71
	NP+cu	2.56	5.37	3.0	0.58	0.54	11.58	13.37
	FP	2.62(.02)	5.34(.03)	2.91(.02)	1.88(.13)	0.56(.00)	10.88(2.05)	12.34(1.88)
	FP+cu	2.47(.04)	5.12(.07)	3.08(.04)	0.68(.10)	0.54(.00)	10.39(2.00)	12.23(2.00)
	CP	2.61(.03)	5.38(.04)	2.91(.02)	1.87(.13)	0.56(.00)	10.92(1.67)	12.38(1.54)
	CP+cu	2.47(.03)	5.13(.04)	3.09(.02)	0.66(.10)	0.54(.00)	10.79(4.36)	12.25(1.64)
Llama-3.1-8B	NP	2.77	5.78	2.86	1.59	0.57	11.57	12.55
	NP+cu	2.52	5.24	3.13	0.50	0.55	12.68	14.54
	FP	2.63(.02)	5.36(.04)	2.9(.02)	2.04(.13)	0.57(.00)	10.62(2.34)	12.02(2.13)
	FP+cu	2.47(.02)	5.06(.03)	3.09(.02)	0.77(.11)	0.55(.00)	9.98(2.38)	11.85(2.43)
	CP	2.64(.02)	5.42(.03)	2.90(.01)	2.00(.10)	0.56(.00)	10.73(2.15)	12.10(1.95)
	CP+cu	2.48(.02)	5.10(.03)	3.10(.02)	0.70(.08)	0.55(.00)	9.98(2.22)	11.85(2.28)

Table 5:Diversity and readability metrics of responses from Llama-1B and Llama-8B in all conditions.