ATTRIBUTING MODE COLLAPSE IN THE FINE-TUNING OF LARGE LANGUAGE MODELS

Laura O'Mahony[1,2], **Leo Grinsztajn**[1,3], **Hailey Schoelkopf**[1], **Stella Biderman**[1]
[1]EleutherAI, [2]University of Limerick, [3]SODA, INRIA Saclay
Correspondance to lauraa.omahony@ul.ie

ABSTRACT

Large language models (LLMs) are typically trained in two stages: first, pre-training on a large, diverse dataset for general-purpose language modeling capabilities, followed by a fine-tuning stage (often called "instruction tuning" or "alignment") on smaller, more curated datasets to adapt them to a specific task or downstream application, such as chat, or general instruction-following. It is a well-known anecdotal observation that instruction-tuned models have less output diversity, such as the infamous observation that ChatGPT cannot seem to generate more than a handful of jokes. A low output diversity means a model lacks the ability to generate varied outputs, which can be a limitation for many use cases. In this manuscript, we quantify how each step in a typical RLHF or instruction-tuning pipeline changes a model's diversity, for a series of models trained in a controlled fine-tuning setup and compare these models to some open-weight models. We distinguish between two categories of diversity in LLMs: token-level prediction diversity, and model output generation diversity. We find that the supervised fine-tuning and reward-based fine-tuning steps have different effects on these distinct diversity types. Our results have implications for better understanding the effects of instruction tuning on the diversity of language models.

1 INTRODUCTION

In recent years, reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020) and related techniques such as RLAIF (Bai et al., 2022b; Lee et al., 2023), direct preference optimisation (DPO) (Rafailov et al., 2023), and KTO (Ethayarajh et al., 2023) have become the premier paradigm for steering the behaviour of pretrained language models (PLMs) (Jain et al., 2023; Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023). These preference optimisation techniques, in contrast to multitask supervised fine-tuning (Sanh et al., 2021; Wei et al., 2021) ("SFT"), which requires ground-truth examples of desirable behaviour or correct task answers, involving collecting human preference labels of the quality of model completions across a range of prompts. These preference labels are then used to fine-tune a model directly (Rafailov et al., 2023; Ethayarajh et al., 2023) or to train a proxy model of quality or "reward" which is used to optimise a model with an RL algorithm, such as PPO (Schulman et al., 2017).

While there have been a few studies examining issues with RLHF (Casper et al., 2023), and analysing the choice of feedback type for RLHF (Bansal et al., 2023; Go et al., 2023), little is known about how different steps in the fine-tuning pipeline differentially impact model behaviour. Previous work found evidence that OpenAI's `text-davinci-002` exhibits more deterministic behaviour than davinci (janus, 2022), and Kirk et al. (2023) analysed the trade off between out-of-distribution generalisation and output diversity for Llama-7B base, SFT, and PPO. Lee et al. (2024) studied how the representations of GPT2 (Radford et al., 2019) change by comparing it before and after the DPO alignment algorithm is applied to remove toxicity.

In this paper, we build on janus (2022), who find evidence of "mode collapse", defined as an excessive and harmful reduction in output diversity, in OpenAI's `text-davinci-002` model. An example of this is shown in the appendix A.4, where models are asked to invent a new word, and Llama-2-chat repeats the term "Frumplenook" each time. Motivated by some well-understood results around mode collapse and mode covering in the RL literature (Ke et al., 2021), we examine the impacts of SFT and reward-based fine-tuning on the diversity of generations in LLMs. We fine-tune a collection of LLMs ranging in size from 70m parameters to 2.8 billion parameters in a controlled setting allowing us to contrast pure SFT and reward-based, DPO, fine-tuning approaches. We additionally do some analysis on Llama-2 and Llama-2-chat 7B parameter models (Touvron et al., 2023) as well as PPO and DPO versions. We break up our analysis by quantifying token prediction distributions and output diversity separately. We find that SFT is largely responsible for the drop in next-token prediction diversity metrics with reward-based fine-tuning having only a small additional effect. However, we find in our experiments that the reward-based fine-tuning step causes a dramatic decrease in output diversity which is not present in its SFT precursor version. Further, we find industrially fine-tuned models to have very poor output diversity results, and our results indicate a larger decrease in output diversity for prompts requiring creative or free-form output across all models. The larger diversity drop for creative tasks indicates that less-fine-tuned models may be particularly desirable for creative applications.

2 METHODS

2.1 MODELS AND SET UP

We use the Pythia model suite (Biderman et al., 2023b) for our experiments since it consists of 16+ models from 70M to 12B parameters trained in a highly controlled setup. Moreover, research has already given multiple insights using this model suite (Kaddour et al., 2023; Mosbach et al., 2023; Biderman et al., 2023a), and pretrained weights are available online. We SFT and subsequently DPO fine-tune 6 models (70m, 160m, 410m, 1b, 1.4b, and 2.8b parameters) with the same effective batch size, on up to 8 GPUs (Tesla A100 40G), a learning rate of $1.0e-6$, and a DPO KL penalty strength term $\beta = 0.1$. We train our models to be helpful assistants by training them for 1 epoch on the helpful subset[1] of the Anthropic helpfulness and harmlessness human preference dataset (Bai et al., 2022a), and the final checkpoints are used for evaluation.

We perform SFT with the `trlX` library (Havrilla et al., 2023) and subsequently perform DPO fine-tuning with the official repository (Rafailov et al., 2023). Finally, we evaluate all models with the Language Model Evaluation Harness (Gao et al., 2023). [2] See Appendix A.1 for a visualisation of the training pipeline and a summary of the experimental pipeline.

2.2 EVALUATION METRICS

To evaluate the diversity of the models described in Sec. 2.1, we use several diversity measures that are well-supported by prior work. Our first set of experiments involves analysing token level metrics such as the increase in surprise on data that is similar to the pretraining dataset. We calculate the perplexity, entropy (Leinster, 2021), and we compute further predicted token metrics such as the average top prediction confidence for each model. Our second set of experiments involves using a set of curated creative and factual prompts to measure model output diversity across a set of model output generations (the similarity of multiple model generations for each prompt). We consider self-BLEU (Zhu et al., 2018), Jaccard similarity using ngrams, normalized compression distance (NCD) (Cilibrasi & Vitányi, 2005), and cosine similarity for a range of temperatures. Each of these metrics produces a scalar measure of diversity. See A.5 for more details.

2.3 EXPERIMENTAL DESIGN

One of the motivators for the following quantitative experiments is a blog post by janus (2022), exploring the increase in deterministic and sometimes stereotyped behaviours from OpenAI's `davinci` model series. See Appendix A.2 for a replication of janus (2022)'s work. Due to a lack of transparency in training procedures, it is unclear how much each step in the fine-tuning pipeline causes this apparent reduction in diversity.

[1] https://huggingface.co/datasets/Dahoas/static-hh

[2] Models, code to perform controlled fine-tuning, and experiments are available at https://github.com/lauraaisling/trlx-pythia, https://github.com/lauraaisling/analyse-llms.

The Pythia suite of models is trained on the Pile (Biderman et al., 2022; Gao et al., 2020). We calculate the token diversity metrics on the Pile validation dataset (Gao et al., 2020), data that is similar but not a part of the PLM's training set.

To explore another facet of diversity metrics, we quantified a range of model generation based measures. In particular, we curated 20 creative prompts (e.g., "write a poem"), and 20 factual style prompts (e.g., "What is the capital of France?")[3]. See Appendix A.3 for the full list of prompts. We prompt the fine-tuned models with the prompt structure they expect, that is "Human: {prompt} Assistant: {start output}". We prompt the PLMs with "{prompt} Okay, here goes: {start output}", the form we considered most appropriate for the foundation model to continue. For each prompt, we generate 25 completions with a maximum of 70 new tokens, across 15 model temperatures. We calculate a range of metrics on these prompts for a maximum of 20 words.

3 Experimental Results

As mentioned in section 2.3, early experiments involved qualitative analysis of model predictions using various prompts. However, due to the qualitative nature of these experiments, and the lack of training details to assess how OpenAI models, as studied by janus (2022), compare to our pipeline, we failed to draw conclusions based on these results. See Appendix A.2 for some examples of this. In this section, we report the results of various diversity quantifications with the goal of attributing the decrease in token and output diversity to each step of the fine-tuning pipeline.

3.1 Token level diversity

Table 1 gives perplexity and entropy of all models on the same 10,000 documents from the Pile validation dataset (Gao et al., 2020). Across all model sizes, due to a distribution shift toward the fine-tuning data, we observe an expected increase in perplexity when predicting the validation set of pretraining data. We observe this to a larger extent for SFT fine-tuning, with DPO fine-tuning only slightly increasing the perplexity in comparison. Similarly, the prediction entropy of each model size calculated on this same dataset decreases significantly after SFT, and following DPO, it only marginally decreases.

On the same task, we calculate the mean output probabilities and assess the cumulative distribution function (CDF) of the token prediction probabilities. Given a vocabulary of n tokens $X_1, X_2, ..., X_n$, we formally define the empirical CDF as $F_e(x) = \sum_{i=0}^{x} P(X_i)$ where the tokens are ordered from most to least probable according to the model's average prediction on the text, that is, $P(X_1) \geq P(X_2) \geq ... \geq P(X_n)$, where $P(X)$ is the average probability the model assigns to predicting token X next. Intuitively, this is just taking the most to least probable tokens and calculating the number of tokens whose probabilities add up to some cutoff point, 0 to 1. We report the average proportion of tokens that make up 90% and 70%, respectively, of the total probability mass in Table 2. Across all model sizes, 4.1% of tokens add up to the 70% probability mass for the PLM, while for most models, 3.5-3.6% of tokens for SFT and DPO do the same. The proportions of tokens adding up to the 70th percentile Pythia-70m and Pythia-160m fine-tuned versions are substantially smaller, this could be due to the constant learning rate having varying effects by the model size. The proportion of tokens adding up to the 70th percentile for the PLM and SFT model shows a large decrease, while the DPO model shows a negligible further reduction from its SFT precursor. However, the pattern of DPO minimally decreasing diversity compared to SFT remains. The same can be said when looking at the percentage of vocabulary tokens adding up to 90th percentile. The average next token prediction confidence can be seen to vary by model size. The table shows the model has increased the confidence of its next token prediction after SFT fine-tuning, with little additional confidence gained from the DPO fine-tuning.

3.2 Model generation diversity

Figure 1 depicts our results relating to the diversity of model generations across various temperatures for factual and creative prompt types for versions of Pythia-2.8B. Observing the curves for self-BLEU for both factual and creative prompts, we see that the DPO models show considerably more

[3] This is inspired by Figure 21 in Touvron et al. (2023)

MODEL	PERPLEXITY			ENTROPY		
	PLM	SFT	DPO	PLM	SFT	DPO
PYTHIA-70M	18.33	30.07	30.28	3.31	1.89	1.96
PYTHIA-160M	12.89	16.18	16.61	3.02	2.23	2.22
PYTHIA-410M	9.23	9.69	10.02	2.52	2.28	2.18
PYTHIA-1B	8.13	8.41	8.59	2.33	2.15	2.06
PYTHIA-1.4B	7.55	7.87	8.15	2.23	2.01	1.88
PYTHIA-2.8B	6.86	7.11	7.35	2.09	1.92	1.80

Table 1: Model perplexity and entropy

	70%			90%			$y_{confidence}$		
MODEL	PLM	SFT	DPO	PLM	SFT	DPO	PLM	SFT	DPO
70M	0.0414	0.0213	0.0217	0.265	0.1673	0.1611	0.4404	0.6133	0.6006
160M	0.0408	0.0249	0.0257	0.2597	0.1605	0.1610	0.4805	0.5680	0.5629
410M	0.0405	0.0350	0.0351	0.2584	0.2328	0.2341	0.5266	0.5577	0.5693
1B	0.0406	0.0358	0.0358	0.2580	0.2360	0.2364	0.5444	0.5664	0.5769
1.4B	0.0406	0.0357	0.0351	0.2580	0.2364	0.2350	0.5543	0.5817	0.5967
2.8B	0.0402	0.0363	0.0366	0.2567	0.2389	0.2403	0.5688	0.5880	0.5996

Table 2: Proportion of all tokenizer tokens where the cumulative sum of average token probabilities is 70% or 90%, and next prediction average confidence.

"mode collapse" at the output generation level than at the token level as in section 3.1, with a notably lower diversity for creative type prompts for both metrics. In comparison, the SFT model has an output diversity closer to that of the PLM. Again for cosine similarity, the DPO model's outputs are notably less diverse for both prompt types. See Appendix A.3 for a depiction of the results with further metrics included and results for Llama-2-7B.

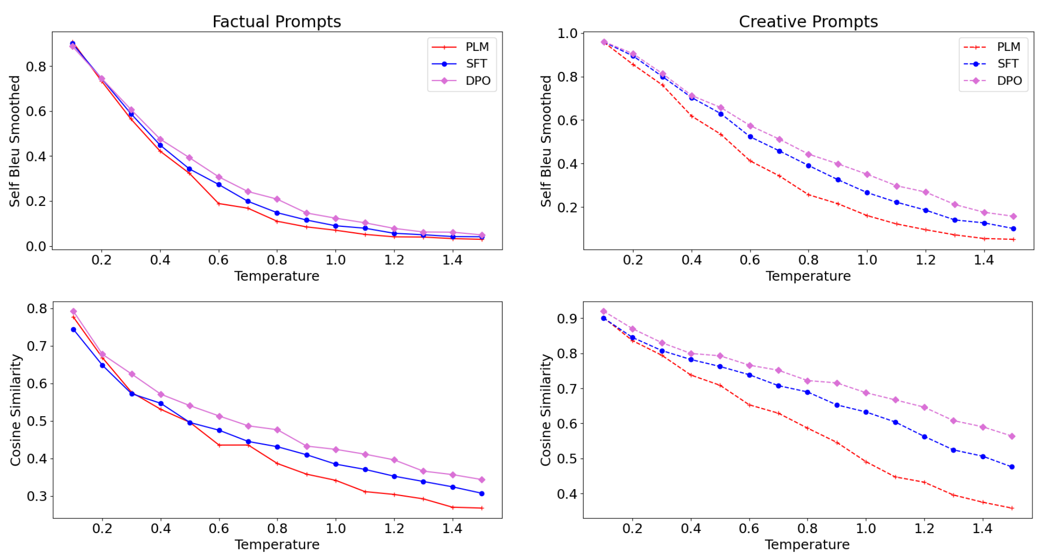

Figure 1: The self-BLEU and cosine similarity of model outputs for the factual (left) and creative (right) prompts for versions of Pythia-2.8B. See A.5 for details on the metrics.

4 Discussion and Conclusion

In this analysis, we carry out an empirical study on the effects of fine-tuning on LLM diversity metrics. Across various model sizes, we compared PLMs to SFT models and models further fine-tuned with DPO as a reward-based fine-tuning method. Studying DPO allowed us to study models ranging from 70 million to 2.8 billion parameters and keep all hyperparameters fixed across all model sizes. Future work could extend this analysis for PPO and other reward-based algorithms, albeit differences in hyperparameters would have to be considered.

Our results validate anecdotal claims of a reduction in output diversity or "mode collapse" brought about at some stage of the fine-tuning pipeline Jain et al. (2023); janus (2022). We attribute and quantify decreases in different categories of diversity metrics to particular steps in the fine-tuning pipeline. In particular, our token level diversity quantifications suggest that SFT is mainly responsible for a drop in next-token prediction diversity metrics, as further DPO reward-base fine-tuning, using the same dataset, has negligible additional effects. However, we specifically attribute a large decrease in output diversity to the reward-based step in the fine-tuning pipeline. We found this effect to be even more pronounced for creative-type prompts, but still existing for factual-type prompts. One potential explanation for this is the phenomenon of "mode-covering" versus "mode-seeking" (or collapsing) objectives–reward-based fine-tuning methods such as PPO and DPO include a reverse KL divergence penalty term, incentivising the minimisation of this divergence (Korbak et al., 2022), which is known from the RL literature to to collapse behaviour into one of several modes (Ke et al., 2021). However, as shown in Appendix 6, the industrially fine-tuned Llama-2-chat exhibits a far lower level of diversity across all output level metrics than any of the fine-tuned versions. Future work could involve comparing the drop in diversity for various levels of SFT and reward-based fine-tuning as well as adding a rejection-sampling step into the fine-tuning pipeline, as is performed in some industrially fine-tuned models, including Llama-2-chat Touvron et al. (2023). Finally, we observe a much larger drop in diversity for creative type tasks for SFT and DPO finetuning. In practical terms, this suggests that PLMs, and to some extent, SFT models give the best performance on tasks where we desire this type of diversity, especially for creative type tasks.

Acknowledgments

We would like to thank StabilityAI for access to compute which allowed us to undertake this work. We also thank many members of the EleutherAI community for fruitful discussions on ideas related to this project. LOM acknowledges the support of Science Foundation Ireland under Grant Number 18/CRT/6049.

References

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Hritik Bansal, John Dang, and Aditya Grover. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*, 2023.

Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*, 2022.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*, 2023a.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023b.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Rudi Cilibrasi and Paul MB Vitányi. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545, 2005.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. Human-centered loss functions (halos). Technical report, Contextual AI, 2023. https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL `https://zenodo.org/records/10256836`.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. *arXiv preprint arXiv:2302.08215*, 2023.

Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. trlX: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8578–8595, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.530. URL `https://aclanthology.org/2023.emnlp-main.530`.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2023.

janus. Mysteries of mode collapse. `https://www.lesswrong.com/posts/t9svvNPNmFf5Qa3TA/mysteries-of-mode-collapse`, 2022. Accessed: 20th January 2024.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pp. 313–329. Springer, 2021.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as Bayesian inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.77. URL https://aclanthology.org/2022.findings-emnlp.77.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Tom Leinster. *Entropy and diversity: the axiomatic approach*. Cambridge university press, 2021.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

A Appendix

A.1 Set up

Figure 2 shows an overview of the model training steps undertaken as well as some of the metrics used.

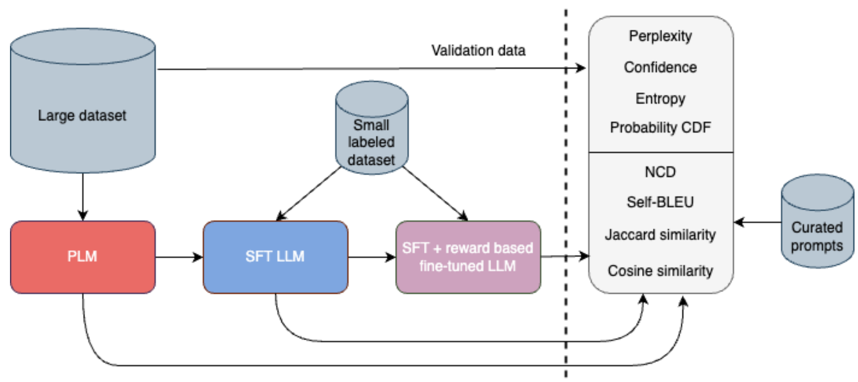

Figure 2: A typical fine-tuning workflow is shown on the left of the dashed line. Our experiments are performed across various model sizes on a PLM, a SFT version of the same model, and a further reward-based fine-tuning step applied to the SFT model. The experiments performed are shown in the right box.

A.2 Mode Collapse

Motivated by confusion over the exact training procedures for OpenAI's original davinci and text-davinci-002, we replicated the experiments by janus (2022) on our models trained in a controlled setup. We failed to conclude any clear patterns due to prompt sensitivity, noise, and potential attentional bias. In addition, industrial LLMs likely go through much more fine-tuning than in our experiments as we train on only 1 epoch of a publicly available dataset (Bai et al., 2022a). For Pythia-410m, figure 3 visualises the next token prediction confidence for the PLM, SFT, and DPO versions of the model. Figure 4 exemplifies a result on the task of picking a random integer. The model is prompted as follows:

"Q: Tell me a random integer between 0 and 100.
A: Ok, the integer is:"

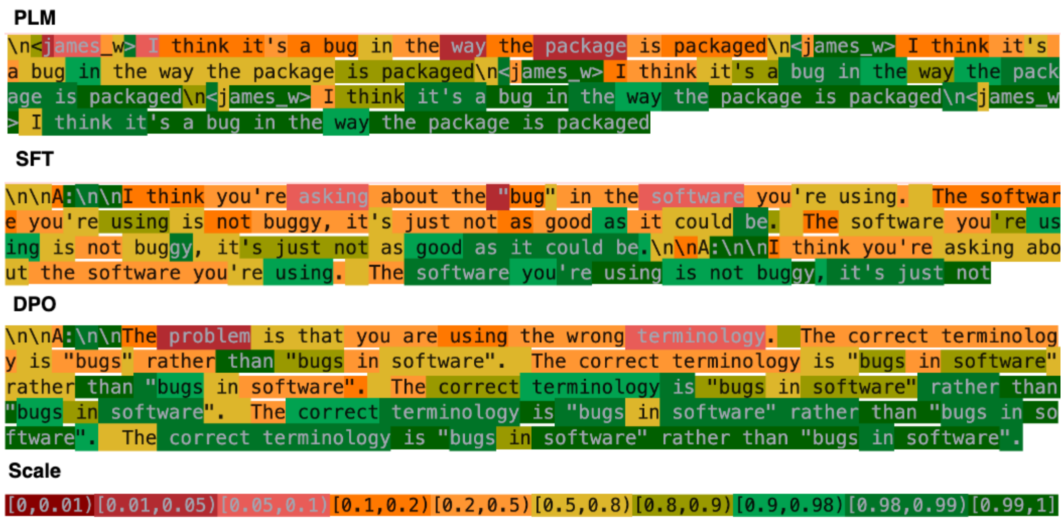

Figure 3: Pythia-410m PLM, SFT, and DPO confidence of maximum 100 new tokens predictions when prompted "are bugs real?"

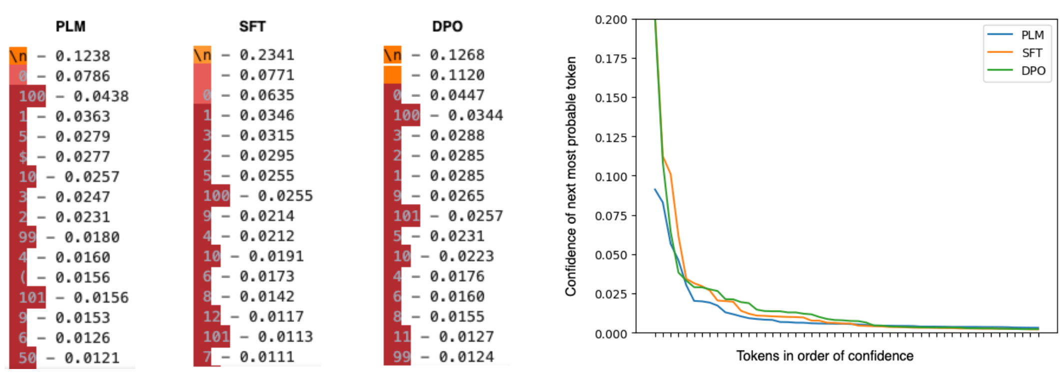

Figure 4: Pythia-410m PLM, SFT, and DPO when prompted "Q: Tell me a random integer between 0 and 100. A: Ok, the integer is:". The top next token predictions and confidences are shown on the left. A chart of these next token prediction confidences from highest to lowest is given on the right.

A.3 Creative and factual prompt completions further results

Figure 5 shows the results for the full set of output diversity metrics.

Figure 6 shows all output diversity metrics calculated on Llama-2-7B base, chat, and publicly available DPO and PPO fine-tuned versions [4]. The diversity of the Llama PLM is comparable with the diversity of the Pythia 2.8b parameter model we studied. There is no significant difference between the diversity of the PPO and DPO model versions. The diversity of the Llama chat model is far lower than that of Pythia-2.8b SFT and DPO versions. We do not have the recipe for fine-tuning Llama chat, however, we conjecture that this model has had significantly more SFT and reward-based fine-tuning applied than in our pipeline, and rejection sampling is applied in addition to our steps.

Table 3 gives a complete list of creative and factual prompts used.

[4] https://huggingface.co/ContextualAI/archangel_sft-dpo_llama7b, https://huggingface.co/ContextualAI/archangel_sft-ppo_llama7b

Table 3: Model creative and factual prompts for completion based diversity metrics.

CREATIVE PROMPTS
WRITE A POEM
TELL ME A JOKE
DESCRIBE THE FEELING OF LOVE
WRITE A STORY STARTING WITH 'ONCE UPON A TIME...'
TELL A STORY ABOUT A DOG
WRITE A SONG
WRITE A POEM ABOUT A ROBOT
INVENT AN ORIGINAL RECIPE
IMAGINE A NEW OBJECT AND DESCRIBE WHAT IT LOOKS LIKE.
IMAGINE A NEW PHILOSOPHY AND DESCRIBE IT.
CREATE A NEW GAME AND EXPLAIN THE RULES.
WRITE A NEW MYTH EXPLAINING THE ORIGIN OF RAINBOWS.
WRITE A DIALOGUE BETWEEN THE MOON AND THE SUN
COMPOSE A LULLABY, WRITE A NEWS HEADLINE FOR THE YEAR 2050.
INVENT A RIDDLE AND WRITE IT DOWN.
WRITE A STORY ABOUT TWO PEOPLE SEEING EACH OTHER FOR THE FIRST TIME.
WRITE A STORY ABOUT A PERSON WHO IS AFRAID OF THE DARK.
MAKE A NEW PUN ABOUT LLAMAS.
INVENT A NEW WORD AND DEFINE IT.
FACTUAL PROMPTS
WHAT IS THE CAPITAL OF FRANCE?
WHAT IS H2O COMMONLY KNOWN?
WHAT IS THE LARGEST COUNTRY IN THE WORLD?
HOW MANY DAYS ARE IN A YEAR?
WHAT IS THE LARGEST PLANET IN THE SOLAR SYSTEM?
WHAT IS THE LARGEST ANIMAL IN THE WORLD?
HOW DO YOU SAY HELLO IN SPANISH?
WHO WON THE 2018 WORLD CUP?
WHAT IS THE BIGGEST CITY IN EUROPE?
WHAT IS THE LARGEST COUNTRY IN AFRICA?
WHAT WAS THE LAST BATTLE OF NAPOLEON?
WHAT IS A NICKNAME FOR SOMEONE FROM NEW ZEALAND?
WHAT DO YOU CALL SOMEONE WHO STUDIES PLANTS?
WHO INVENTED THE TELEPHONE?
WHAT MAMMAL LAYS EGGS?
WHICH BONE IS THE LONGEST IN THE HUMAN BODY?
WHAT IS THE ANTHEM OF FRANCE?
WHO WROTE CANNERY ROW?
WHO WAS THE FIRST PRESIDENT OF THE UNITED STATES?
WHICH PAINTER PAINTED THE MONA LISA?

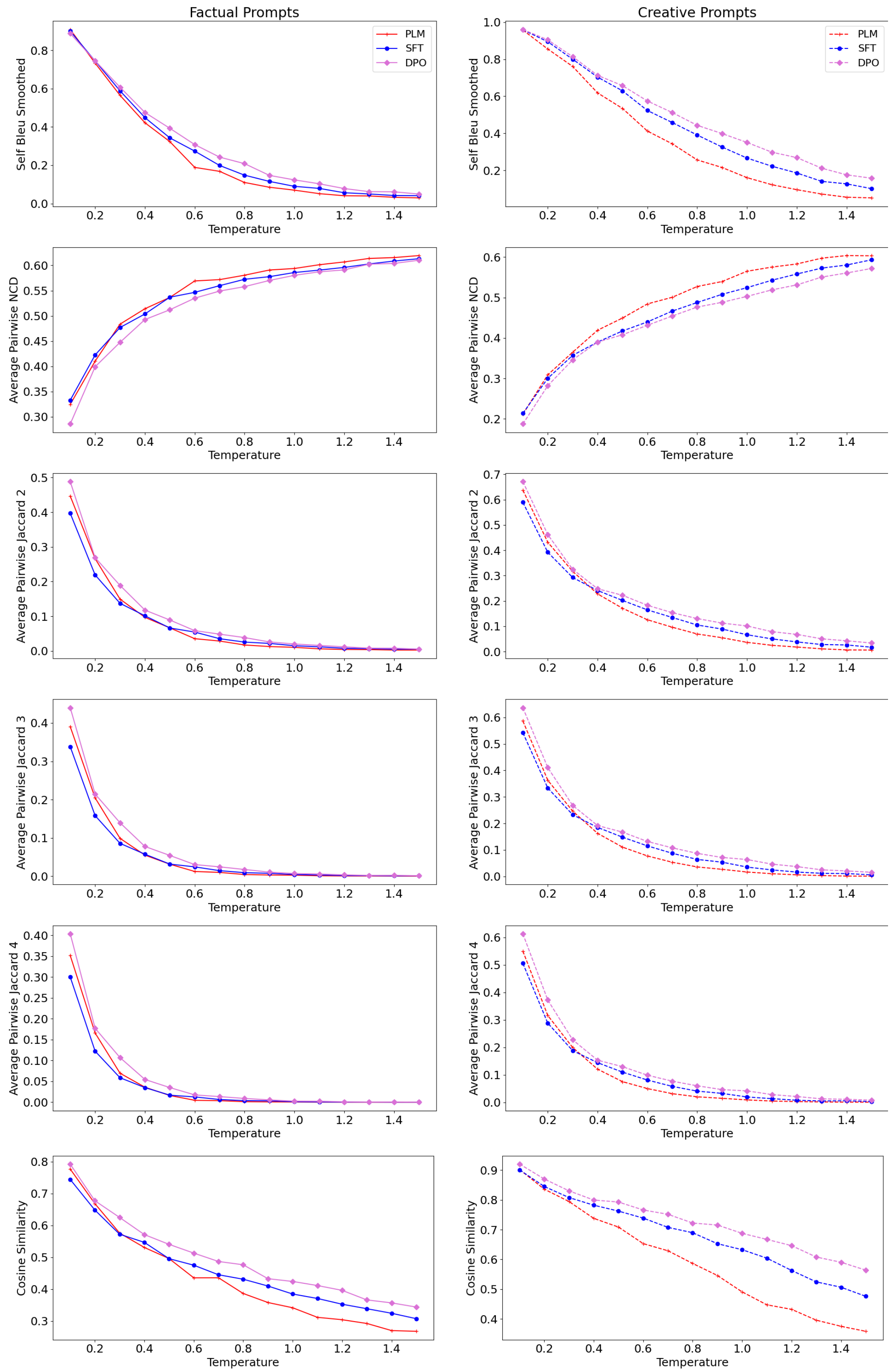

Figure 5: The left and right subfigures show the full set of diversity metrics for the factual and creative prompts respectively, across a range of temperatures on the x axis.

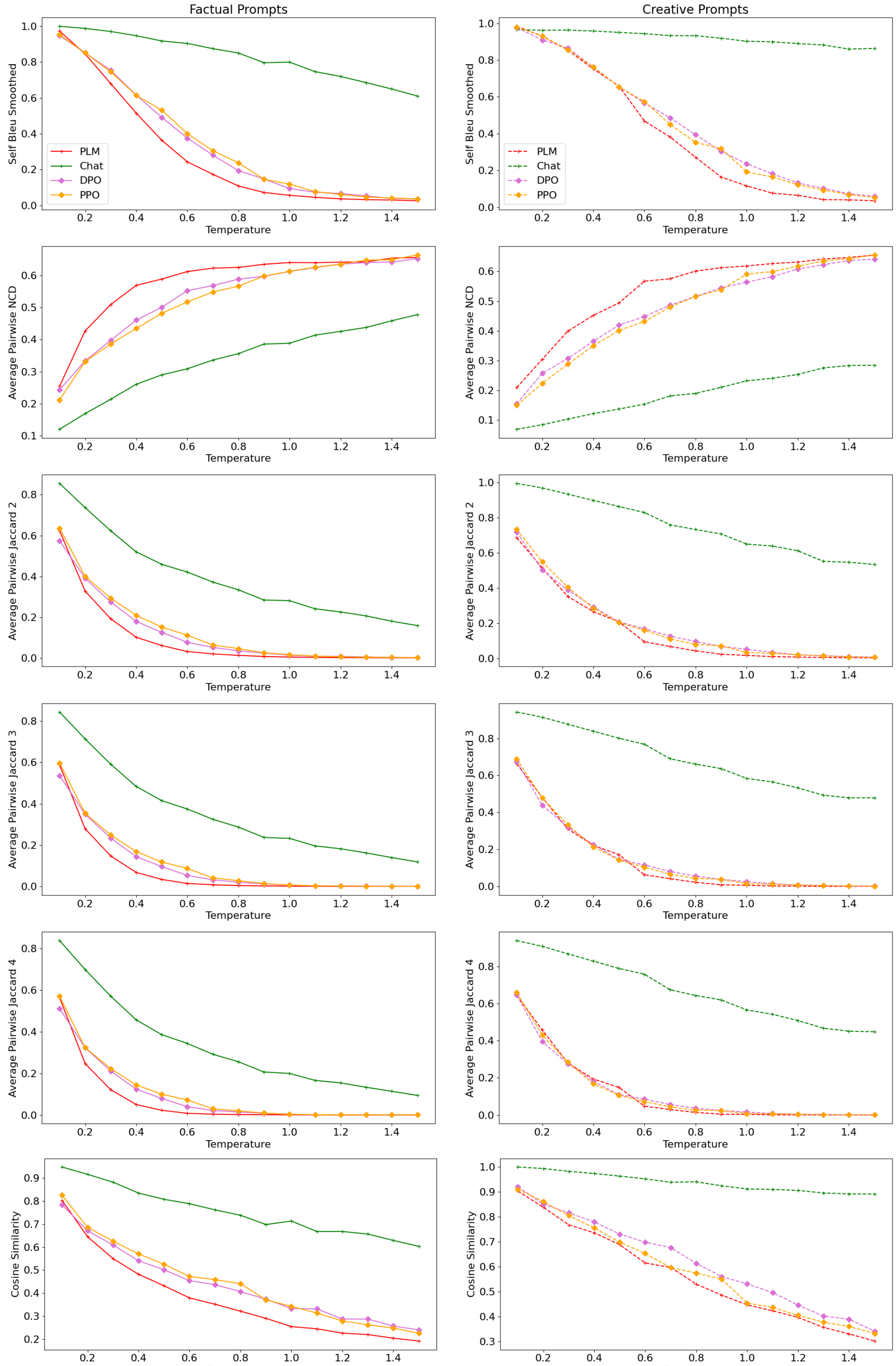

Figure 6: The diversity metrics for factual and creative prompts for Llama-2-7B base, chat, DPO, PPO.

A.4 EXAMPLES OF MODEL RESPONSES

Prompt: "Invent a new word and define it.", setting temperature 1.

- Llama-2-7B
  - 2-plyism is the attempt to convince oneself that a two-time convicted felon who was pardoned by the governor and a former vice-president in order to become a national hero and a symbol of American greatness is innocent. It's a new form of denial, in which 2-ply
  - Samay - the act of leaving a social network to the point of abandoning it completely. What's the best way to do it? There are several possibilities: A) The classic way: delete the app from your phone, sign out of
- Llama-2-7B-chat
  - Certainly! Here's a new word I've invented: "Flumplenook" (noun): a feeling of contentment and satisfaction that comes from achieving a difficult but worthwhile task, often accompanied by a sense of pride and accomplishment. Example sentence: "After finally completing
  - Sure, I can do that! Here is a new word I have invented: "Flumplenook" (noun): a feeling of contentment and satisfaction that comes from enjoying a simple, yet unexpected pleasure, such as a beautiful sunset or a delicious bowl of soup. Example sentence:
- Pythia-2.8b
  - An _action_ is an act of performing any action. As you can see on my previous posts, English has no good way of describing acts. So "an act" in "an action" is an attempt to describe the act without using a good name for that particular act. This also goes for "to act" for',
  - "Antecedents" means preceding or preceding in time. "Succedents" is the set of future events that have some influence on the antecedents. (The concept is more complicated) A:Ante and post can be used synonymously, just as "preceding"',
- Pythia-2.8b-SFT
  - na-che-ta, It's a word that means something like a 'way forward' for a person, as in "There are a number of ways that people can help each other."
  - I will create a special term and define it: Woo. And provide examples for how people use it: "My friend is Woo," "We think the world would be better with Woo" "I was talking with my friend" "It's Woo on the bus" "My friend is using
- Pythia-2.8b-DPO
  - A "new word" is a word that has been coined or created new by someone for the first time. An example of a new word might be "doodad" created by a person in a magazine or on the internet, but is still considered a new word. A new definition is a word that is used to define something or',
  - Panspermia – the theory that life can exist in outer space or in the void of space-time without being enclosed within a living cell. It is the concept that life can survive and evolve in outer space, on planetary surfaces, or in interstellar space for an indefinite period of time. There is also the theory that life could be

A.5 DETAILS ON DIVERSITY METRICS

We would like to thank Ryan Teehan for sharing with us code for computing various diversity metrics [5].

Cosine similarity We compute the average of the pairwise cosine similarity of the output embeddings. Embeddings are computed using the SBERTEmbedder model from SentenceTransformer (Reimers & Gurevych, 2019)).

[5] available at https://github.com/CarperAI/diversity_metrics

Self-BLEU The BLEU scores used for the Self-BLEU computation are done using the package NLTK (Bird et al., 2009) with default parameters (meaning that they are done from 1-grams to 4-grams), and NLTK smoothing function "method1". The n-grams are computed on words.