Learning to vary: Teaching LMs to reproduce human linguistic variability in next-word prediction

Tobias Groot
University of Amsterdam
tobias.groot@student.uva.nl

Salo Lacunes
University of Amsterdam
salo.lacunes@student.uva.nl

Evgenia Ilia
University of Amsterdam
e.ilia@uva.nl

Abstract

Natural language generation (NLG) tasks are often subject to inherent variability; *e.g.* predicting the next word given a context has multiple valid responses, evident when asking multiple humans to complete the task. While having language models (LMs) that are aligned pluralistically, so that they are able to reproduce well the inherent diversity in perspectives of an entire population of interest is clearly beneficial, Ilia and Aziz (2024) show that LMs do not reproduce this type of linguistic variability well. They speculate this inability might stem from the lack of consistent training of LMs with data reflecting this type of inherent variability. As such, we investigate whether training LMs on multiple plausible word continuations per context can improve their ability to reproduce human linguistic variability for next-word prediction. We employ fine-tuning techniques for pre-trained and instruction-tuned models; and demonstrate their potential when fine-tuning GPT-2 and Mistral-7B-IT, using Provo Corpus. Our evaluation, which measures divergence among empirically estimated human and model next-word distributions across contexts before and after fine-tuning, shows that our multi-label fine-tuning improves the LMs' ability to reproduce linguistic variability; both for contexts that admit higher and lower variability.

1 Introduction

Inherent variability in natural language generation (NLG) tasks might arise from ambiguity or varying perspectives (Plank, 2022; Baan et al., 2023). For example, when predicting the next word given a context, multiple plausible and valid continuations exist; a task whose linguistic variability we can appreciate by asking a human population to complete it (Luke and Christianson, 2018). We can also appreciate this type of linguistic variability for autoregressive language models (LMs) that generate text by sampling from next-token (*i.e.* subword unit) distributions conditioned on preceding tokens (Vaswani et al., 2017). We achieve that by viewing such distributions as a representation of the model's uncertainty over continuations given a prefix (Ilia and Aziz, 2024; Guo et al., 2024; Tevet and Berant, 2020). It is often valuable for models to reproduce such variability, particularly in open-ended NLG tasks, where multiple responses can be plausible. Whereas this variability contributes to making LMs more robust (Sheng et al., 2008; Peterson et al., 2019; Uma et al., 2021; Kurniawan et al., 2025) and more representative of the linguistic diversity of human populations of interest (Sorensen et al., 2024; Muscato et al., 2025b), it has been shown that the variability LMs exhibit does not always align with the one humans exhibit (Pavlick and Kwiatkowski, 2019; Ma et al., 2025; Shaib et al., 2024). For next word prediction, Ilia and Aziz (2024) identify this misalignment and speculate it might stem from inconsistent exposure of LMs to training data reflecting such variability.

As such, we investigate whether training LMs with multiple observations of the next word per context will improve their ability to reproduce human variability. While previous fine-tuning work utilising multiple references per instance focused on classification tasks (Peterson et al., 2019; Uma et al., 2021; Rajeswar et al., 2022), our work focuses on next-word prediction, a generative task. Similar to Eisape et al. (2020), who employ a form of multi-label distillation in next word prediction, we also employ a technique to fine-tune pre-trained LMs and extend to instruction-tuned LMs. For the former, we alter the training signal, and for the latter we exploit a training data augmentation method to ensure that variability is observed.

We employ these fine-tuning techniques for GPT-2 (Radford et al., 2019), a pre-trained model, and Mistral-7B-IT (Jiang et al., 2023), an instruction-tuned model. When evaluating, by measuring divergence among empirically estimated human and model next-word distributions across contexts, be-

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 73–88
November 9, 2025

fore and after fine-tuning, we find that fine-tuning with multiple labels per instance improves those LMs' ability to reproduce linguistic variability, across contexts of varying open-endedness. Additional ablations measure performance when varying the number of training labels per instance; and with a preliminary analysis we measure the trade-off in performance in tasks that admit no plausible variability. For that, we handcraft a small evaluation dataset using a knowledge-based question answering dataset (Berant et al., 2013).[1]

2 Related Work

Human label variation in natural language processing (NLP) tasks is often dismissed as noise (Paun et al., 2022; Ferracane et al., 2021). However, multiple responses can be plausible, especially relevant to ambiguous or open-ended tasks or prompts (Plank, 2022; Baan et al., 2023; Weber-Genzel et al., 2024; Nie et al., 2020; Aroyo and Welty, 2015). Embracing this plausible variation as part of NLP systems, which could make them more fair (Deng et al., 2023; Muscato et al., 2025b) and robust (Peterson et al., 2019; Sheng et al., 2008), involves altering all stages of our systems' development pipelines: from dataset creation, collecting multiple labels per prompt (Luke and Christianson, 2018; Nie et al., 2020, *i.a.*), to training, utilising these labels during the learning phase (Rodríguez-Barroso et al., 2024; Aroyo and Welty, 2012; Padmakumar et al., 2024, *i.a.*), and evaluation, comparing models' responses to multiple human references (Baan et al., 2022; Ilia and Aziz, 2024, *i.a.*).

Our approach aims to embrace plausible variability during training. Rather than collapsing annotations into a single ground truth (Paun et al., 2022), we incorporate multiple plausible references. The idea of multi-label fine-tuning has been adopted in image-classification (Peterson et al., 2019; Aurpa et al., 2024; Rajeswar et al., 2022), as well as in NLP, primarily for classification (Uma et al., 2021; Jung et al., 2023; He and Xia, 2018; Beţianu et al., 2024; Li et al., 2024; Zhang et al., 2024a; Li et al., 2025; Muscato et al., 2025a). Additionally, recent efforts have applied instruction fine-tuning for multi-label text classification tasks (Siddiqui et al., 2024; Yin et al., 2024) and tasks with restricted outcome spaces, such as sampling from discrete distributions (Zhang et al., 2024b). Our work focuses on a generative task, (*i.e.*, that of predicting complete wordforms by stringing together tokens), with a countably infinite outcome space (*i.e.*, all possible wordforms from a finite set of tokens). Eisape et al. (2020) explores a form of multi-label distillation in next-word prediction for an LSTM model. We also explore a form of multi-label distillation for transformer-based models, extending our investigation to instruction tuned LMs.

3 Methodology

We exploit simple yet intuitive fine-tuning techniques, depending on the LMs' previous training. These require a set of contexts $C = \{c_1, ..., c_N\}$, where for each context c_i, we have a set of human next-word references $W_i = \{w_{i1}, ..., w_{iM}\}$[2]:

Fine-tuning pre-trained LMs Autoregressive LMs are trained using cross-entropy between a target and the model's next-token distribution given context c ($p(\cdot|c)$ and $q(\cdot|c)$ resp.). This corresponds to searching for the maximum likelihood estimate (MLE). When training on a corpus with a single continuation (*i.e.* the next corpus token w^*), p is a deterministic distribution centered on w^*, leading to the following loss:

$$L_{\text{Label}} = -\log q(w^* \mid c). \quad (1)$$

When multiple word continuations are available, we replace this deterministic distribution with the empirically estimated distribution (using W_i), where the probability of a word given c_i, $p(w|c_i)$, equals its relative frequency in W_i. This results in the following loss, which comprises generalized cross entropy (Jurafsky and Martin, 2025):

$$L_{\text{Var}} = -\sum_{w \in \mathcal{V}} p(w \mid c_i) \log q(w \mid c_i), \quad (2)$$

where $\mathcal{V}$ is the vocabulary.[3] Since words may consist of multiple tokens, to obtain $q(w \mid c_i)$ we re-express the model's token-level probabilities over complete words.[4] For a word w with tokenization $\tau(w) = (t_1, \ldots, t_n)$, we compute:

$$q(w \mid c_i) = \prod_{j=1}^{n} q(t_j \mid c_i, t_1, \ldots, t_{j-1}), \quad (3)$$

where $q(t_j \mid \cdot)$ is the probability of token t_j under the model, given the context and preceding tokens.

[1] Code available at: GitHub repository

[2] M might vary accross contexts.

[3] Words that actually contribute to the loss, i.e. non-zero terms, are words in the set of human samples, W_i for c_i.

[4] Humans predicted *word* continuations, not tokens; so the outcome space of $p(w \mid c_i)$ is over complete words, and we must ensure that $q(w \mid c_i)$ is expressed over the same space.

Fine-tuning instruction-tuned LMs Instruction-tuned models underwent additional training to cater a rather conversational format and adhere to task instructions. We sample responses from the model's conditional predictive distribution (CPD) given a prompt, *i.e.* an instruction and an example. For our task, we sample response r containing a predicted word from $q(r|(I, c_i))$, where the prompt includes instruction I requesting a word continuation given a prefix, and the example context c_i. So as to utilise multiple labels, we employ the following training data augmentation technique: for each context c_i in C, we construct the prompt (I, c_i) and for each word w_j in W_i, we create a training datapoint where w_j is a response to (I, c_i). This entails that c_i will appear multiple times with different continuations as per their frequency in W_i. We train using L_{Label}.[5] See Appendx A for prompts.

4 Experiments

Models & Datasets. We fine-tune pre-trained GPT-2 (124M; Radford et al. (2019)) and instruction-tuned Mistral-7B-Instruct-v0.3 (7.25B; Jiang et al. (2023)), which we refer to as Mistral-7B-IT. Both models are fine-tuned using Provo Corpus (Luke and Christianson, 2018), which contains 55 text passages (2687 total contexts). Each prefix is annotated with an average of 40 human annotations predicting the word following it. We split the dataset randomly at the paragraph level (to avoid partial passage leaks between train and test sets). 80% is for training, of which 10% is reserved for validation; and the remaining for testing.

Training Configuration. Both models were fine-tuned using the Adam optimizer (Kingma, 2014). For GPT-2: we train for 3 epochs, using a learning rate of $1e^{-5}$ and a batch size of 16. For Mistral-7b-IT: we train for 4 epochs using Low-Rank Adaptation (Dettmers et al., 2023, LoRA) with a learning rate of $1e^{-4}$ with a batch size of 32. We train on 3 random seeds; training details in Appendix B.

Metrics. Following Ilia and Aziz (2024): for each context, we measure the divergence between the human and model CPDs given a context using total variation distance (TVD) (Rudin, 1987).[6] TVD quantifies the difference between two probability distributions by summing the absolute differences in the probabilities they assign to the same event. A higher TVD indicates greater disagreement between human and model CPDs (*i.e.*, poorer alignment with human linguistic variability), whereas a lower TVD indicates less disagreement (*i.e.* better alignment with human linguistic variability). In order to compute TVD, we need estimates of the human and model CPDs ($p(w|c)$ and $q(w|c)$ respectively). As done in Ilia and Aziz (2024): (1) for $p(w|c)$, we estimate it via Monte Carlo, with $p(w|c)$ equaling the relative frequency of w in all human samples, and (2) for $q(w|c)$, we estimate it via Monte Carlo, by sampling 40 sequences from the model long enough to contain a full word, slice it, and compute $q(w|c)$ (or $q(w|(I, c))$) as the relative frequency of w in all sampled words.

Mean TVD ± SD (↓)		
Model	**GPT-2**	**Mistral-7b-IT**
Base	0.607±0.001	0.812±0.002
1-Shot	N/A	0.784±0.002
FT (Orig. corpus)	0.612±0.002	0.805±0.001
FT (Maj. label)	0.556±0.005	0.563±0.002
FT (Mul. labels)	**0.550**±0.003	**0.499**±0.006
Oracle	0.443±0.002	0.443±0.002

Table 1: Mean and standard deviation of TVD averages across test contexts for three seeds.

Baselines & Upper Bounds. We compare the distribution of TVD values across contexts before and after fine-tuning, where improved performance would mean a shift towards lower TVD values (*i.e.* less disagreement with human CPDs). For the instruction-tuned model, we add a 1-shot baseline, where the prompt includes an example of a context and word references (details in Appendix A). As another baseline, we fine tune models with Provo's original corpus passages (*i.e.* one continuation per prefix), imitating models' usual training. Lastly, to estimate the best performance we can expect from our models, which essentially is to mimic human divergence, we establish a baseline for the expected level of disagreement from humans for a context. We split human responses in two disjoint groups and measure their CPDs' TVD ('Oracle' baseline).

5 Results

Main results As shown in Table 1, both models fine-tuned with multiple labels (FT (Mul. labels)) achieve a notably lower mean TVD compared to other baselines (Base, 1-Shot and FT (Orig. Cor-

[5] Constructing the dataset in this way (one prompt-response pair for each word annotation for every context) using L_{Label} is similar to learning $q(r|(I, c))$ with L_{Var}.

[6] $\text{TVD}(p, q) = \frac{1}{2} \sum_w |p(w|c_i) - q(w|c_i)|$

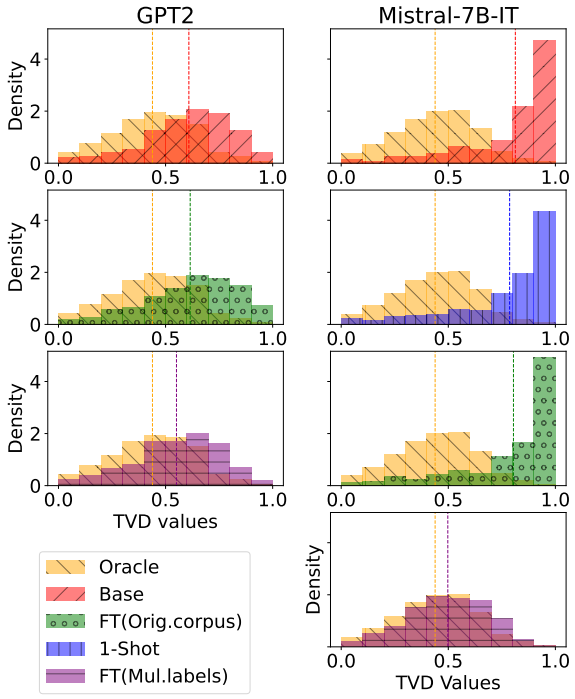


Figure 1: Distribution of TVD scores (for 1 seed) across contexts. For both GPT-2 and Mistral-7B-IT; fine-tuning shifts the TVD distribution towards the Oracle baseline, suggesting better linguistic alignment with humans.

pus)). We also observe how FT (Orig. corpus)'s performance is very similar to the Base model. This simultaneously indicates that our improved performance does not stem from an out-of-distribution effect between Provo Corpus and the models' training data. Figure 1, which shows the histogram of TVD values for all models and baselines (for 1 seed), confirms that; FT (Mul. labels) models' TVD distributions shift towards the Oracle distributions, indicating that models improve at reproducing human linguistic variability. For other seeds, we see similar patterns; see Appendix C.

When and how do models improve? To understand the effects of our fine-tuning, we analyze changes in TVD. We visualise the models' changes in performance against context open-endedness (as measured by the TVD between human oracles; lower TVD indicating more 'restrictive' contexts), allowing us to grasp if performance gains arise in contexts that admit higher or lower variability. In Figure 4 (Appendix D), negative TVD differences between fine-tuned and base models (indicating gains) occur at all levels of contexts' open-endedness. We also assess whether models improve at predicting words that humans predicted (regardless of frequency). We plot the fraction of unique human predictions that were also predicted by the models before and after fine-tuning (Figure 7; Appendix D). Fine-tuned Mistral-7B-IT's ability to predict unique human words (along with its CPDs' 'diversity'; Figure 6, Appendix D) improves substantially (details and analyses in Appendix D).

Is the entire response distribution useful? When gathering datasets with multiple labels, disagreement can be discarded as noise and the most common response is used as ground truth. Aiming to assess whether retaining the entire response distribution is useful, we fine-tune a model on Provo Corpus using only the majority response (FT (Maj. Label) in Table 1). We find that, FT (Maj. Label) surpasses the performance of FT (Orig. Corpus), which is not entirely surprising: the corpus word is a single observation, while the majority vote exploits in a sense multiple labels. This is intuitively in line with analysis revealing that performance gains seem to relate with less open-ended contexts (Figure 8; Appendix E). Nonetheless, FT (Mul. labels) outperforms FT (Maj. label), with moderate gains for GPT2 and more notable gains for Mistral-7B-IT; indicating the utility of retaining all labels.

Number of labels ablation. We analyse how the number of labels used to fine-tune the model affect the model's performance. We fine-tune GPT2 using a varying number of labels each time (1,2,4,16 and 32; randomly sampled from available annotations). Figure 9 of Appendix F shows that 16 samples are sufficient for substantial performance improvements; for more details, see Appendix F.

Impact on tasks without data uncertainty. Whereas optimising for a task that admits inherent variability (*i.e.* next-word prediction) might improve the model's ability to reproduce such variability; the effect of this on tasks that admit no variability is unclear. To assess that, we test the models' performance before and after fine-tuning on knowledge-based question answering (a task admitting no plausible variability), adapted for next-word prediction. For that, we handcraft examples from a subset of WebQuestions (Berant et al., 2013); details and examples in Appendix G. For each context, we sample 40 responses and measure how often responses exactly match the reference. As shown in Table 3 of Appendix G, fine-tuning on multi-label data moderately improves the low performance of GPT2, but worsens the performance

of Mistral-7B-IT; highlighting a potential trade-off in performance between tasks that do and do not admit variability, when optimising for the latter.

6 Conclusion

This study examines whether fine-tuning with multiple labels per instance has the potential to enhance models' ability to reproduce linguistic variability in next word prediction. We show improved performance for a smaller pre-trained language model (GPT-2) and a larger instruction-tuned model (Mistral-7b-IT) across contexts that admit varying levels of plausible variability. Our findings highlight both the potential and possible limitations of such fine-tuning, paving the way for further advancements in modeling linguistic variation.

7 Limitations

We hereby discuss various limitations of our study: we fine-tune using Provo Corpus, which is a relatively small dataset with a limited number of human annotations per prefix. The high cost of obtaining data with multiple references means that such data is scarce and not available at large scale. However, we show that even with a limited amount of contexts and a limited amount of annotations per context that are well-curated and of high-quality it is possible to observe performance improvements. Simultaneously, as the field of synthetic data generations is becoming increasingly popular; we can entertain the idea that future work exploits such synthetic labels, and a model that has been fine-tuned to embrace variability, such as the ones we present in this study, could comprise generators for such synthetic annotations. Additionally, for our training and evaluation, we assumed all human annotations to be draws from the same underlying distribution; which is not an assumption that is easy to verify. We also observed a trade-off between capturing variability well and performance on tasks with a single correct answer; with future work potentially focusing on methods that could balance-off better such trade-offs. Additionally, due to resource constraints, we were only able to include in our study only two (relatively small) models that were trained for English. Despite focusing on a generative task, we only focused on next word prediction. Transferring this to the sequence level might be non-trivial and come with its own challenges. However, we hope that our study inspires future work in this research direction, aiming to embrace inherent variability as part of the training of LMs, and tackle challenges related to this field.

Acknowledgements

Evgenia Ilia is supported by the EU's Horizon Europe research and innovation programme (grant agreement No. 101070631, UTTER). The experiments and findings presented in this paper were conducted as part of a research project fostered within the NLP 2 course of the MSc AI programme of the University of Amsterdam (2024-2025 edition), coordinated by Ana Lucic.

References

Lora Aroyo and Chris Welty. 2012. Harnessing disagreement for event semantics. *Detection, Representation, and Exploitation of Events in the Semantic Web*, 31.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Tanjim Taharat Aurpa, Md Shoaib Ahmed, Md Mahbubur Rahman, and Md Golam Moazzam. 2024. Instructnet: A novel approach for multi-label instruction classification through advanced deep learning. *Plos one*, 19(10):e0311161.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. *arXiv preprint arXiv:2210.16133*.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Miruna Bețianu, Abele Mălan, Marco Aldinucci, Robert Birke, and Lydia Chen. 2024. Dallmi: Domain adaption for llm-based multi-label classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 277–289. Springer.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *arXiv preprint arXiv:2305.14663*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. Association for Computational Linguistics (ACL).

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7*, pages 250–259. Springer.

Evgenia Ilia and Wilker Aziz. 2024. Predict the next word: Humans exhibit uncertainty in this task and language models _. *arXiv preprint arXiv:2402.17527*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Taehee Jung, Joo-Kyung Kim, Sungjin Lee, and Dongyeop Kang. 2023. Cluster-guided label generation in extreme multi-label classification. *arXiv preprint arXiv:2302.09150*.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models (Chapter 10)*, 3rd edition. Online manuscript released January 12, 2025.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2025. Training and evaluating with human label variation: An empirical study. *arXiv preprint arXiv:2502.01891*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025. Preserving diversity in supervised fine-tuning of large language models. *Preprint*, arXiv:2408.16673.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. *arXiv preprint arXiv:2505.17510*.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and Tommaso Cucinotta. 2025a. Embracing diversity: A multi-perspective approach with soft labels. *arXiv preprint arXiv:2503.00489*.

Benedetta Muscato, Lucia Passaro, Gizem Gezici, and Fosca Giannotti. 2025b. Perspectives in play: A multi-perspective approach for more inclusive nlp systems. *arXiv preprint arXiv:2506.20209*.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

Vishakh Padmakumar, Chuanyang Jin, Hannah Rose Kirk, and He He. 2024. Beyond the binary: Capturing diverse preferences with reward regularization. *arXiv preprint arXiv:2412.03822*.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical methods for annotation analysis*. Springer Nature.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626.

Barbara Plank. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

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

Sai Rajeswar, Pau Rodriguez, Soumye Singhal, David Vazquez, and Aaron Courville. 2022. Multi-label iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4783–4793.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Walter Rudin. 1987. *Real and complex analysis*. McGraw-Hill, Inc.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. *arXiv preprint arXiv:2407.00211*.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Muhammad Hammad Fahim Siddiqui, Diana Inkpen, and Alexander Gelbukh. 2024. Instruction tuning of llms for multi-label emotionclassification in social media content. In *Proceedings of the Canadian Conference on Artificial Intelligence*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. A roadmap to pluralistic alignment. *CoRR*.

Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. *arXiv preprint arXiv:2004.02990*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation. *Preprint*, arXiv:2403.01931.

Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*.

Jinghui Zhang, Dandan Qiao, Mochen Yang, and Qiang Wei. 2024a. Regurgitative training: The value of real data in training large language models. *arXiv preprint arXiv:2407.12835*.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J Zico Kolter, and Daphne Ippolito. 2024b. Forcing diffuse distributions out of language models. In *First Conference on Language Modeling*.

A Prompts for Baselines

When constructing the training set and evaluating our models, we present the relevant prompts:

Base prompt. To assess the performance of the non fine-tuned models, we prompt them repeatedly for the next word prediction task. The prompt includes an instruction to predict a next-word continuation and the given context at a time.

Prompt:

```
Instruction: Return one plausible next
 word for the following context.
 Context: <CONTEXT>
 Continuation:
```

When creating training prompt-response pairs, the prompt is identical to before, and the responses are words from the set of human references.

Response:

```
<HUMAN_REFERENCE>
```

1-Shot prompt. As a performance baseline we have one-shot prompting, which includes the instruction, an example from the training set, and the given context at a time:

```
Instruction: This is an example of a
 context and some plausible next word
continuations. given by a group of 39
 people: Context: There are now
 rumblings that, Continuations:
 [are, are, are, are, are, are, are,
 can, can, can, can, can, sound,
 sound, sound, sound, shake, shake,
 shake, the, the, the, have, have,
 our, our, someone, someone,
 appear, ca, cause, come, make,
 occur, people, say, suggest,
 tumble, we]. Following this
 example, return only one plausible
  next word for the following context.
 Context: <context>
 Continuation:
```

B QLoRa

Table 2 shows the configuration used for finetuning the Mistral-7B-IT model.

Parameter	Value
QLoRA	
r	8
LoRA α	16
LoRA dropout	0.05
Task type	Causal Language Modeling
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
Quantization	
Load in 4-bit	True
4-bit quantization type	nf4
Double quantization	True
Compute data type	bfloat16

Table 2: LoRA and 4-bit quantization configuration parameters.

C Main results

We present Figure 1, which comprises the results on the test set for one of the three random seeds we trained on. We observe similar trends for the remaining seeds; which we present in Figure 2. This is confirmed when plotting the differences between the TVD of the model and human CPDs and the TVD among the human oracle CPDs, as observed in Figure 3.

D Analysis of model performance changes

In order to understand how fine tuning has affected the model performance. We perform various analyses. We visualise the models' changes in performance against context open-endedness. We approximate that using the TVD between human oracles. We assume that a lower TVD, reflecting lower disagreement among human populations, indicates more 'restrictive' contexts, while a higher TVD, indicates contexts that admit a higher level of plausible variability. We plot changes in performance by computing the differences between the TVD of the fine tuned model and human CPD and the TVD of the non fine tuned model and human CPD. Results are shown in Figure 4 (showing all contexts) and Figure 5 (showing only contexts for which performance improved, *i.e.* negative differences in TVD values). We observe how improvements occur across contexts of varying open-endedness (*i.e.* varying TVD among oracles values).

To gain further insight as to how fine tuning has affected our models, we plot the entropy values of the empirically estimated model CPDs across contexts before and after fine-tuning. Results can be seen in Figure 6. For GPT2, we observe how the entorpy of the model's empirically estimated CPDs were not impacted very substantially. We observe only a slight shift towards lower entropy values (*i.e.* peakier distributions); which means that model predictions might be slightly more confident, while also being better better aligned with human linguistic variability. On the contrary, the fine tuned Mistral-7B-IT model's entropy values shift substantially towards higher values, demonstrating that now the model is making more diverse predictions (which are also better aligned with human linguistic variability, as evident by our main findings).

Lastly, we assess whether models improve at predicting words that humans predicted (regardless of their frequency), as a means to approximate how wel their lexical diversity aligns with that of our assessed human population. We plot the fraction of unique human predictions that were also predicted by the models before and after fine-tuning with multiple labels (Figure 7). Higher values indicating a more highly aligned lexical diversity. We find that GPT2's lexical diversity remained relatively similar to before fine tuning, but for Mistral-7B-IT we see a clear rightward shift in the distribution of unique word coverage for the fine-tuned model. This indicates that the fine-tuned model predicts a greater number of relevant unique words per context compared to the non-fine-tuned baseline.

E Analysis of model fine-tuned with majority label

Similar to Appendix D, we analyse the changes in performance of the model fine tuned with the majority label compared to the base model. We visualise the models' (FT (Maj.Label)) changes in performance against context open-endedness. We approximate that using the TVD between human oracles. We assume that a lower TVD, reflecting lower disagreement among human populations, indicates more 'restrictive' contexts, while a higher TVD, indicates contexts that admit a higher level of plausible variability. We plot changes in performance by computing the differences between the TVD of the fine tuned model (FT Maj. label) and human CPD and the TVD of the base model and human CPD. Results are shown in Figure 8. When comparing with the corresponding plots for FT (Mul.label) in Figure 4, we observe how improvements occur for contexts that admit lower plausible variabiltiy (*i.e.* lower TVD among oracles values; steeper regression line towards lower Oracle TVD values for lower negative differences/performance

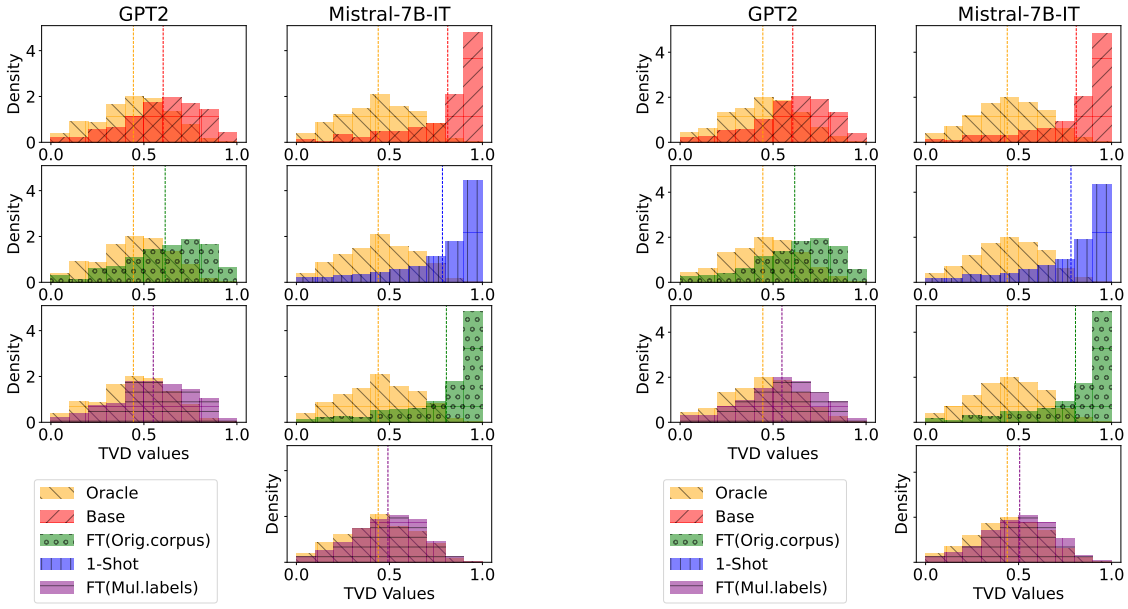


Figure 2: Distribution of TVD scores across contexts, for the two remaining seeds not presented in the main paper. For both GPT-2 and Mistral-7B-IT; fine-tuning shifts the TVD distribution toward the Oracle baseline, suggesting improved alignment with human linguistic variability.

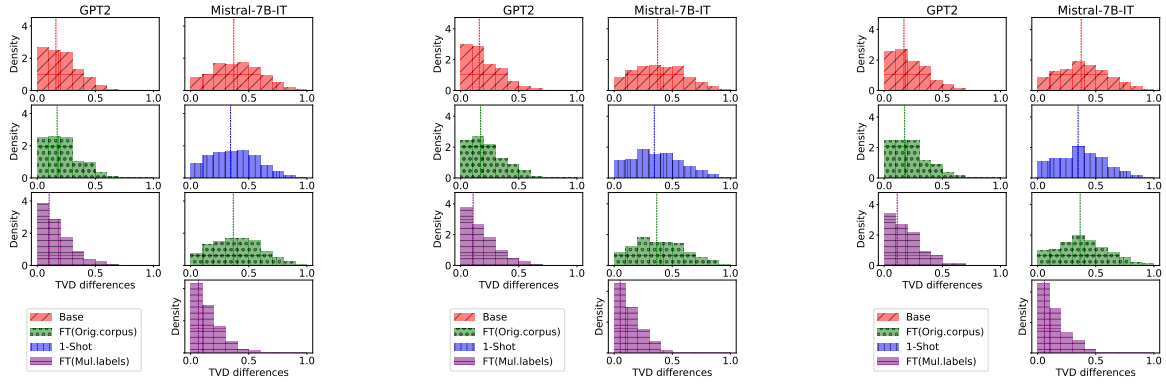


Figure 3: Distribution of differences of TVD scores between the model and the human CPDs and the oracle CPDs, for all 3 seeds. For both GPT-2 and Mistral-7B-IT; fine-tuning shifts the TVD distribution towards smaller differences, confirming previous findings.

gains).

F Varying training labels per instance Study

We perform an ablation to understand the number of labels that is necessary to obtain substantial performance gains. We perform this ablation study only for GPT2, given computational constrains (Mistral-7B-IT is a much larger model, and fine-tuning it repeatedly is computationally prohobited). We sample 1,2,4,16 and 32 labels given our available annotations and fine tune GPT2 given the subsequent training sets. We then perform the same evaluation as for the rest of our analysis and present the average TVD of the test set, against the label set size per instance in Figure 9. Scores for 16 and 32 samples are nearly identical, and very similar to the score obtained when training on all available labels (40 on average per prompt). These results suggest that around 16 labels per instance

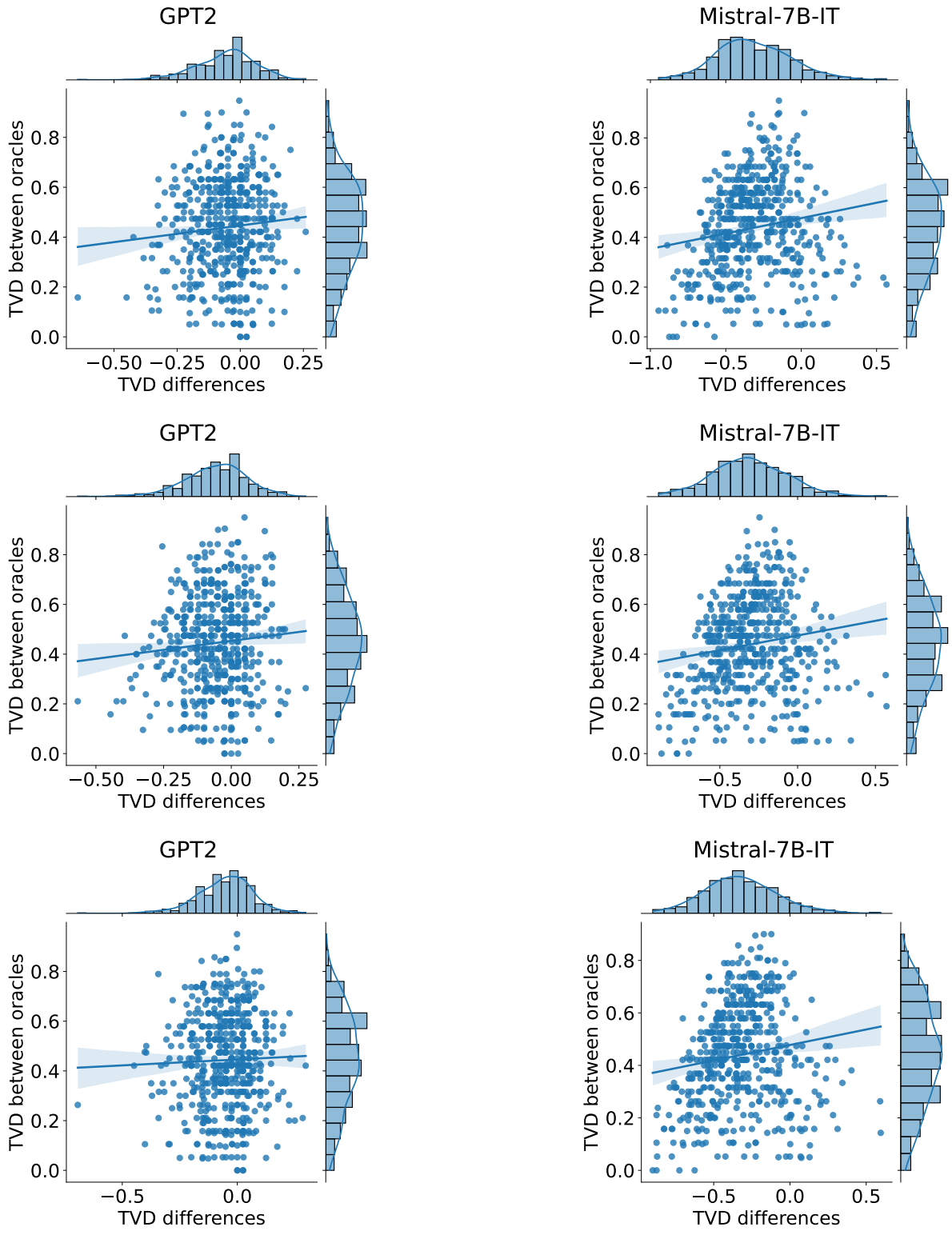


Figure 4: Distribution of differences of TVD scores between the fine tuned model and the human CPDs minus the TVD of the non fine tuned model and the human CPDs, against TVD among oracles. Performance gains (negative differences) for both models occur across contexts of varying open-endedness (with lower TVD indicating more 'restricted' contexts).

are sufficient to observe significant performance gains.

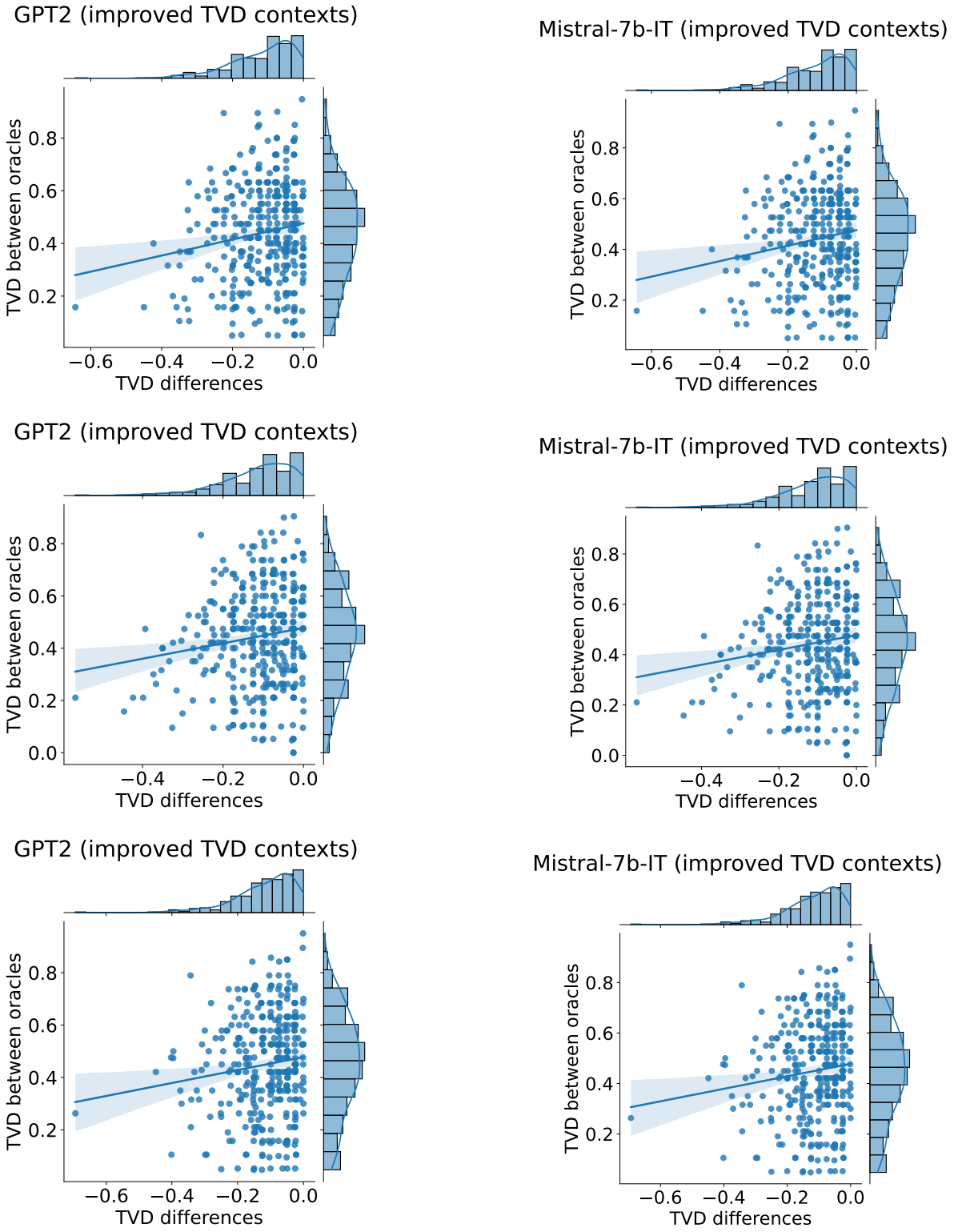
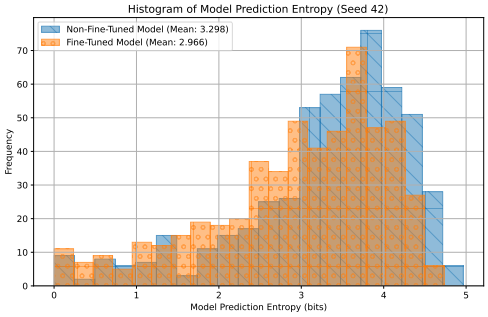
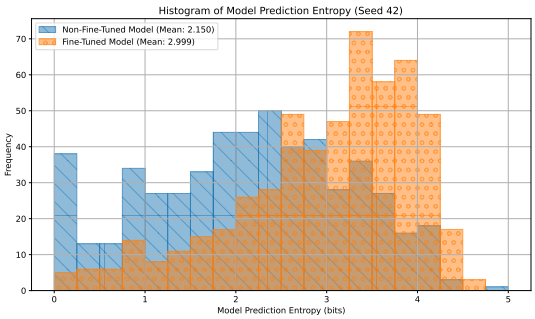


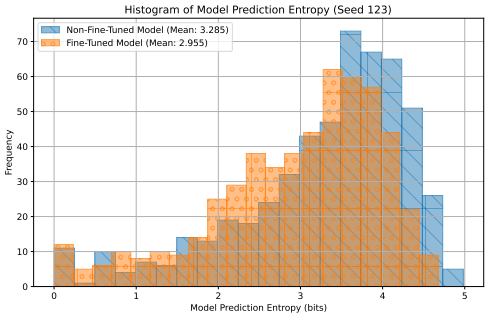
Figure 5: Distribution of differences of TVD scores between the fine tuned model and the human CPDs minus the TVD of the non fine tuned model and the human CPDs, against TVD among oracles. In this case, we only plot datapoints for which we observed improvements (*i.e.* negative differences) for both models. Similarly, we observe that gains occur across contexts of varying open-endedness (with lower TVD indicating more 'restricted' contexts).



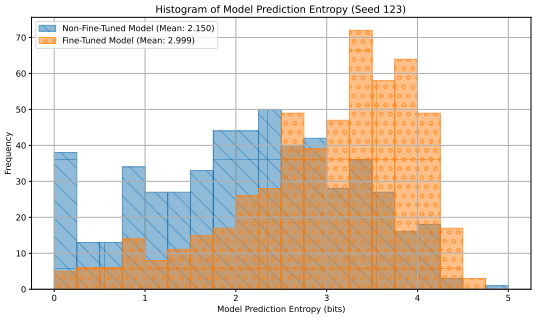
(a) GPT-2 (Seed 42)



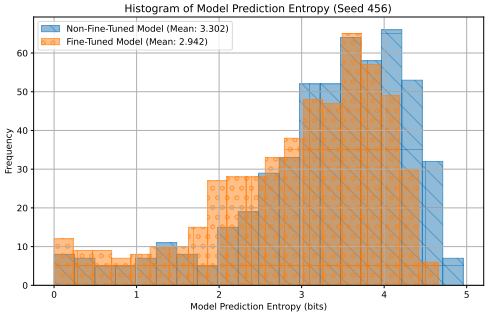
(b) Mistral-7B (Seed 42)



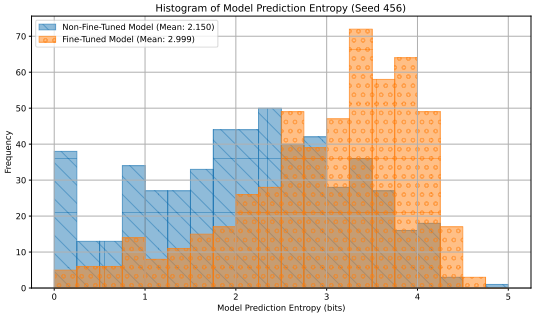
(c) GPT-2 (Seed 123)



(d) Mistral-7B (Seed 123)

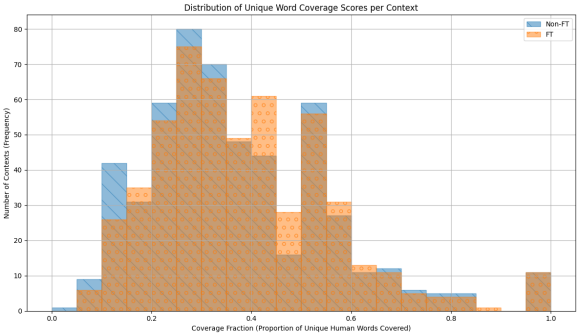


(e) GPT-2 (Seed 456)

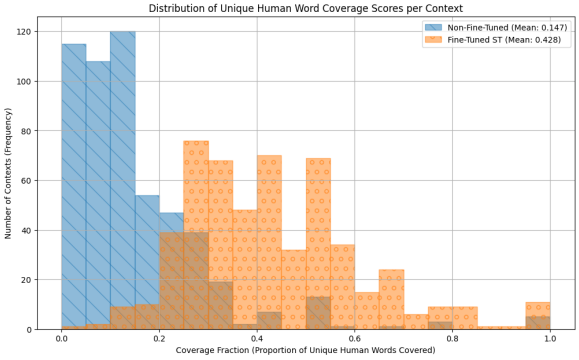


(f) Mistral-7B (Seed 456)

Figure 6: Entropy of model predictions before an after finetuning.



(a) GPT-2



(b) Mistral-7B

Figure 7: Unique word coverage across models. Fine-tuning with multiple labels per instance increases lexical diversity more compared to hard-targets (majority vote).

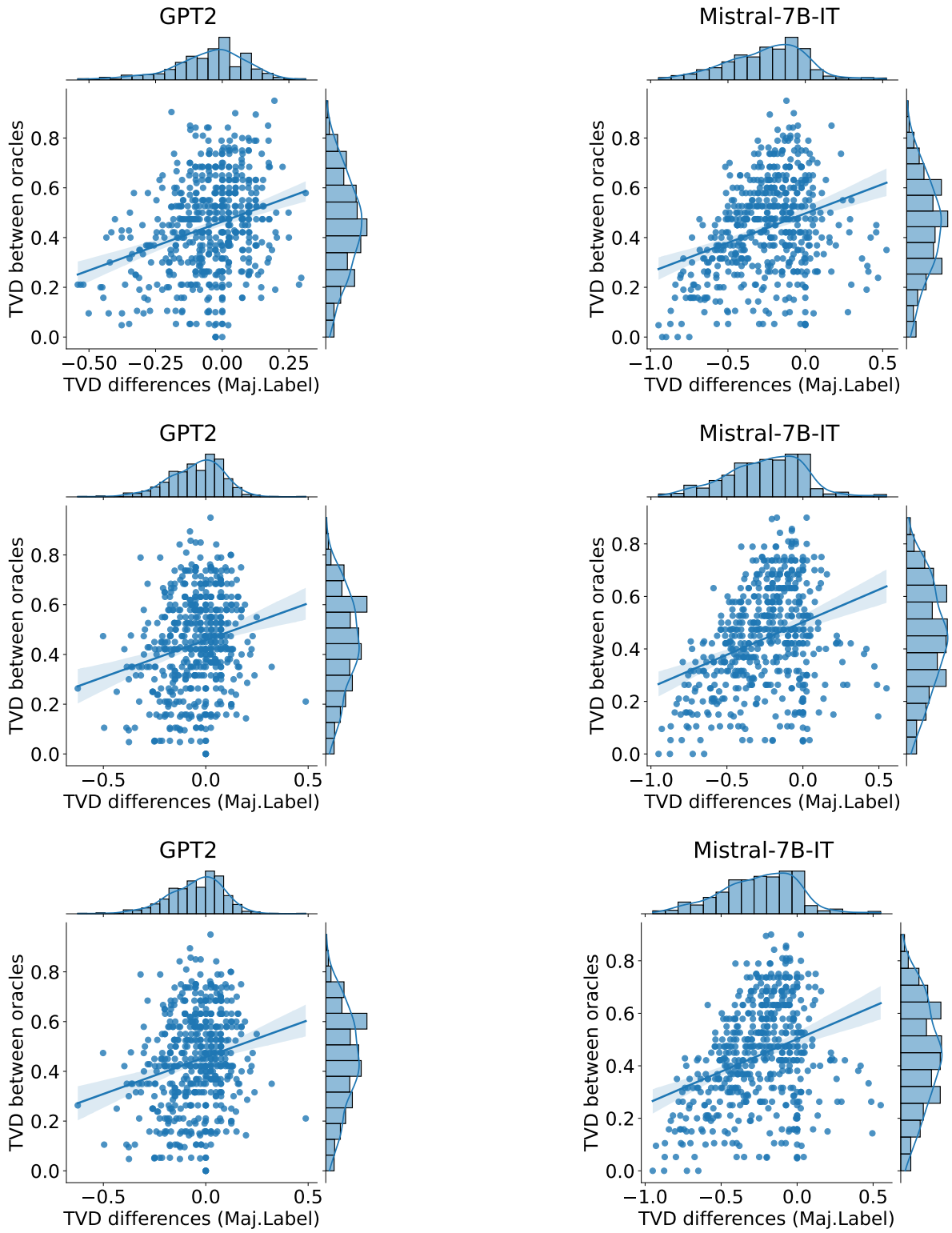


Figure 8: Distribution of differences of TVD scores between the fine tuned model and the human CPDs minus the TVD of the non fine tuned model and the human CPDs, against TVD among oracles. Performance gains (negative differences) for both models occur across contexts of varying open-endedness (with lower TVD indicating more 'restricted' contexts).

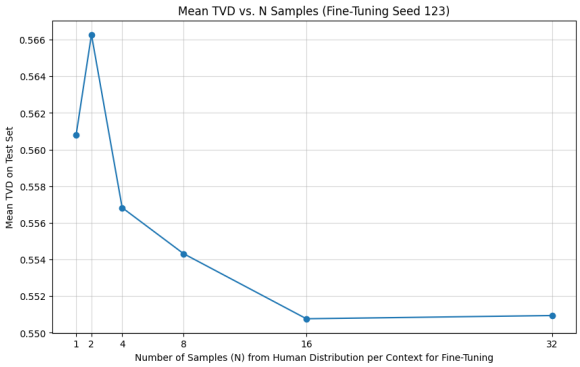


Figure 9: Mean TVD by number of samples per context. Performance improves with more samples, plateauing after 16.

G Analysis for QA task without variability

Whereas optimising for a task that admits inherent variability (*i.e.* next-word prediction) might improve the model's ability to reproduce such variability better; the effect of this on tasks that admit no variability is unclear. We test the models' performance on knowledge-based question answering (which is a task that admits no plausible variability), adapted as a next-word prediction task. We create a small evaluation dataset based on a knoweldge-based question answering dataset, WebQuestions (Berant et al., 2013). We create a subset of 55 handpicked contexts, chosen to include a variety of topics ranging from science, history and pop culture, each rephrased into next-word prediction tasks. We demonstrate 3 randomly chosen examples below:

Prompt:

```
Instruction: Return one plausible next
 word for the following context.
Context: The  first country to invade
 poland in ww2 was
 Continuation:
```

Target:

```
Germany
```

Prompt:

```
Instruction: Return one plausible next
 word for the following context.
 Context: the organelle responsible
for atp production and storage is the
 Continuation:
```

Target:

```
mitochondrion
```

Prompt:

```
Instruction: Return one plausible next
 word for the following context.
 Context: darth vader's star
 destroyer was called
 Continuation:
```

Target:

```
Devastor
```

We also evaluate model performance using the original questions. For Mistral-7B-IT, the instruction was modified into: **QA-Prompt:**

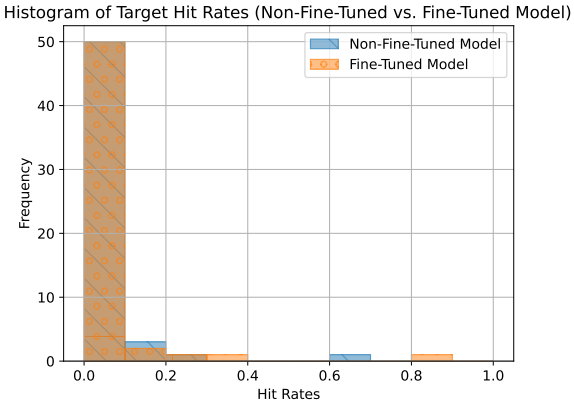
```
Instruction: Answer the following
question with one word only
Context: What country first invaded
poland in ww2?
Continuation:
```

Mean Hit Rate ± SD		
Model	**GPT-2**	**Mistral-7B**
Base	0.032 ± 0.002	0.590±0.005
FT (Orig.corpus)	0.030 ± 0.002	0.229±0.005
FT (Mul.labels)	0.041 ± 0.002	0.127±0.005

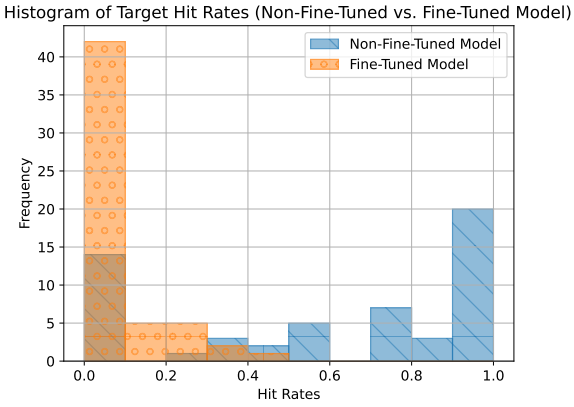
Table 3: Mean target hit rate for 40 samples per context across three seeds with standard deviation, for both GPT-2 and Mistral-7B.

We compare the base model, the model fine tuned with the original corpus (so as to account for the impact of training on Provo corpus, a potentially different domain) and the model that was fine tuned with multiple labels in their ability to generate the correct answer to the question (phrased as a next-word prediction task). To evaluate the performance, for each context, we sample 40 responses and measure how often responses exactly match the reference, denoted as hit rate.

Table 3 shows the results of this evaluation. GPT-2 shows a slight increase in hit rate after finetuning, although its overall performance remains poor, and Mistral-7B-IT's performance also drops, more substantially. However, we cannot rule out the effect of other confounders in the data or optimisation process that might have incidentally impacted the performance changes and are not relevant to the multiplicity of responses. Hence, we approach these preliminary results with cautiousness, and hope to inspire future work that investigates this more extensively. Supplementary histograms of hit-rates across contexts can be seen in Figure 12.

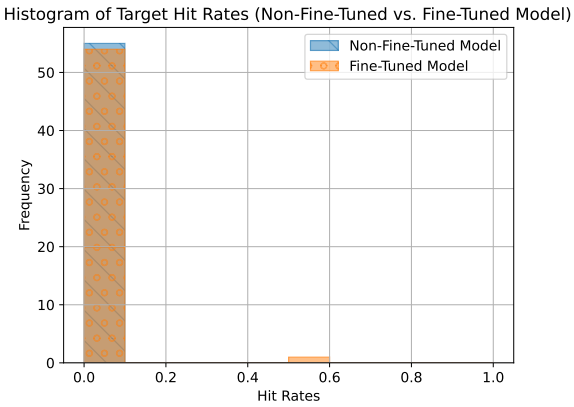


(a) GPT-2

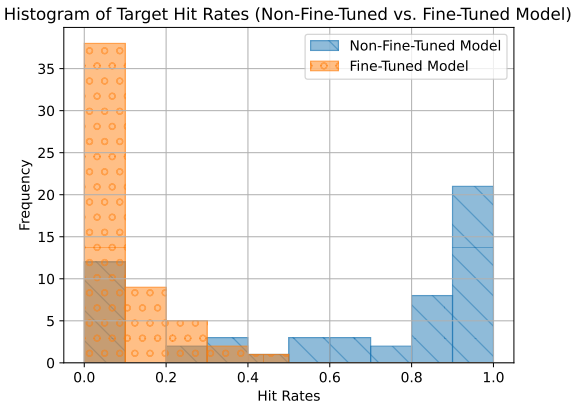


(b) Mistral-7B

Figure 10: Hit rates on gold target label before and after finetuning. Averaged across 3 seeds.

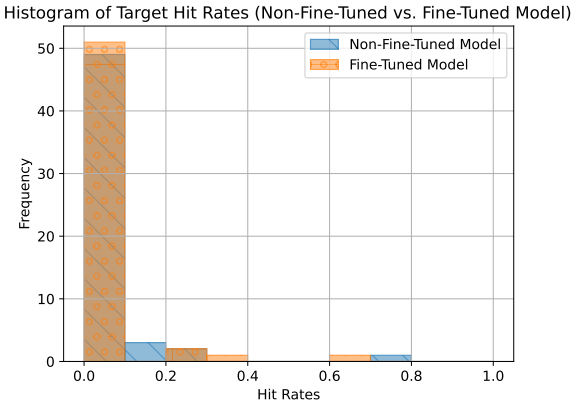


(a) GPT-2

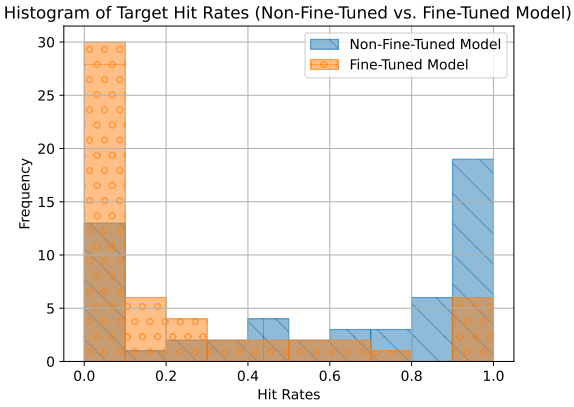


(b) Mistral-7B

Figure 11: Hit rates on gold target label when prompted in the original QA format, before and after finetuning. Averaged across 3 seeds.



(a) GPT-2



(b) Mistral-7B

Figure 12: Hit rates on gold target label after finetuning on hard targets (corpus). Averaged across 3 seeds.