PEFT-U: Parameter-Efficient Fine-Tuning for User Personalization

Anonymous NAACL submission

Abstract

The recent emergence of Large Language Models (LLMs) has heralded a new era of human-AI interaction. These sophisticated models, exemplified by Chat-GPT and its successors, have exhibited remarkable capabilities in language understanding. However, as these LLMs have undergone exponential growth, a crucial dimension that remains understudied is the personalization of these models. Large foundation models such as GPT-3 etc. focus on creating a universal model that serves a broad range of tasks and users. This approach emphasizes the model's generalization capabilities, treating users as a collective rather than as distinct individuals. While practical for many common applications, this one-size-fits-all approach often fails to address the rich tapestry of human diversity and individual needs. To explore this issue we introduce the **PEFT-U Benchmark**: a new dataset for building and evaluating NLP models for user personalization. PEFT-U consists of a series of user-centered tasks containing diverse and individualized expressions where the preferences of users can potentially differ for the same input. Using PEFT-U, we explore the challenge of efficiently personalizing LLMs to accommodate user-specific preferences in the context of diverse user-centered tasks.

1 Introduction

Large Language Models (LLMs) have shown tremendous capability in performing complex tasks such as reasoning, summarization, creative writing, etc. Through the scaling of these models, both in size ($>$ 1B parameters) and data ($>$ 1 Trillion tokens) these models have achieved remarkable performance on a wide range of natural language understanding and generation tasks (Touvron et al., 2023; Henighan et al., 2020). However, even as the generalization capability of LLMs has grown, one crucial dimension that has been understudied is the personalization of these models (Salemi et al., 2023; Kazienko et al., 2023).

At its core, personalization is about tailoring AI-driven interactions to the individual preferences, needs, and idiosyncrasies of each user (Salemi et al., 2023; Welch et al., 2022; Clarke et al., 2023; Kang et al., 2022). In many real-world scenarios, users have unique preferences, contexts, and expectations, which are currently incapable of being effectively accommodated by the generalized LLMs available today. These traditional LLMs have predominantly adhered to a "one-size-fits-all" approach (Touvron et al., 2023; Clarke et al., 2022; OpenAI, 2023; Bubeck et al., 2023), offering a single, uniform model capable of serving all users and tasks alike. While this approach is undoubtedly valuable in many scenarios, it falls short when it comes to accommodating the rich tapestry of human diversity where people are not uniform, and their linguistic and communicative preferences vary widely (Li et al., 2023; Zhang et al., 2018).

Existing works in NLP have highlighted the need for user perspective in language modeling particularly when dealing with intrinsically subjective applications such as Hate Speech Detection and Sentiment Analysis where differing perspectives are common (Mostafazadeh Davani et al., 2022; Sang and Stanton, 2021; Geva et al., 2019; Kanclerz et al., 2022; Welch et al., 2022). Research has shown that accounting for user perspective and personalization is essential for building robust and effective models (Mostafazadeh Davani et al., 2022; Sang and Stanton, 2021; Geva et al., 2019; Kanclerz et al., 2022; Welch et al., 2022). However, despite this glaring need, existing resources fail to model and cater to these differing perspectives. When curated, NLP resources tend to have an intrinsic bias towards the majority perspective due to their reliance on voting for ground truth. As such they fail to adequately represent diverse user preferences and individualized expressions, further contributing to a lack of personalization (Mostafazadeh Davani et al., 2022; Sang and Stan-

ton, 2021; Geva et al., 2019; Kanclerz et al., 2022).

To combat these challenges we introduce the **PEFT-U Benchmark**. PEFT-U consists of over 13+ personalized tasks and 15k+ users across domains such as Hate Speech, Sentiment/Emotion, and Humor. In contrast to other resources, the PEFT-U benchmark uniquely tests complex scenarios where LLMs are faced with differing user perspectives even when facing the same input. To the best of our knowledge, this benchmark is the first of its kind to focus on modeling user preferences in NLP with an emphasis on identical inputs that require different model outputs depending upon the user. Using PEFT-U we explore a range of strategies for efficiently compartmentalizing user perspectives. In particular, we implement and empirically analyze a series of personalized prompting approaches (non-parametric) vs tuning and compartmentalizing user-level knowledge (parametric) for personalized tasks showing that personalized models are crucial to providing users with more accurate results representative of their actual perspectives. We publicly release our code, models, and benchmark[1].

2 PEFT-U Benchmark

The PEFT-U benchmark aims to assess the efficacy of language models in producing personalized outputs based on user-specific information.

Data Collection To generate high-quality data samples representative of differing user perspectives we focus on curating subjective problems where the model is forced to respect the user's points of view e.g. Hate Speech Detection. Typically NLP resources for these problem areas employ an annotation process where correctness is determined via majority vote and outliers are discarded. This often results in the overlooking of the subtleties of the user's perspective, ultimately leading to potential group biases and inaccuracies in the data. In contrast, we reconstruct these data sources framing individual annotators as distinct users to capture these important nuances. To avoid the possible influence of noisy/bad annotators we take into account their contribution level to the annotation process and discard low-quality users. Additionally, we discard users with less than $n = 10$ samples in their training and test sets respectively. As shown in table 1 PEFT-U consists of 13+ personalized tasks and 15k+ users with each task gaining a maximum Krippendorff's alpha (α) of 0.5 (Krippendorff, 2011) across the domains of Hate/Abuse, Humor, and Emotion/Sentiment as shown in table 1. For each dataset, we construct a unique instruction-style prompt to guide the LLM to generate the desired output. More details on each of the specific datasets, our reconstruction process, and their prompts are provided in Appendix A.

User Disagreement As shown in table 1, we enforce that all personalized tasks must obtain a Krippendorff's alpha score of ($\alpha \leq 0.5$). This requirement is created to assess the ability to capture differing user perspectives even when facing the same input. Krippendorff's alpha (α) is a reliability coefficient developed to measure the agreement among annotators. When used in data curation, data is usually considered reliable when ($\alpha \geq 0.8$). By enforcing all datasets to have low agreement scores, we force the model to rely on respective user information to generate its output.

3 Modularity + Personalization

When exploring the problem of personalization, one possible solution would be the allocation of a dedicated LLM for each user. However, deploying a separate personalized model for each user would incur significant compute costs in production. In addition, the balance between embedding generalized and personalized knowledge in these models remains unclear. Thus providing such a solution is prohibitive in this era of large language models. Recent works in Modular Deep Learning (Liu et al., 2022a; Pfeiffer et al., 2023; Hu et al., 2021; Houlsby et al., 2019), seek to optimize and further tune these LLMs without having to update all the model parameters. These methods typically introduce a small number of additional parameters and update these parameters in a model while freezing the remaining weights, thus limiting the computational resources required. This is often done to enable multi-task learning or to introduce new updates in the training data without the need for training from scratch.

In our experiment setting, we shift this paradigm from multi-task learning to multi-user learning focusing on the research question of *"How to efficiently personalize large language models for subjective tasks?"*. As such, we empirically analyze personalized prompting approaches (non-parametric) vs efficiently tuning and compartmentalizing user-level knowledge (parametric) for per-

[1] https://github.com

Dataset		# Users	Avg # examples per user	Avg # users per text	Krippendorff's Alpha
Domain	**Name**				
Hate+Abuse	HateXplain (Mathew et al., 2022)	253	238.9	3.0	0.46
	GabHate (Kennedy et al., 2018)	18	4807.16	3.12	0.24
	MeasuringHateSpeech (Sachdeva et al., 2022)	7912	17.13	3.42	0.47
	TweetEval (Röttger et al., 2022)	20	200.0	200.0	0.14
	UnhealthyConversations (Price et al., 2020)	588	386.77	4.64	0.28
	WikiDetox Aggression (Wulczyn et al., 2017)	4053	336.84	11.78	0.43
Sentiment	GoEmotion (Demszky et al., 2020)	82	1859.95	2.83	0.44
	StudEmo (Ngo et al., 2022)	25	296.44	1.43	0.18
	Subjective Discourse (Ferracane et al., 2021)*	68	9.26	6.20	0.50/0.01/0.01
Humor	Cockamamie (Gultchin et al., 2019)	1878	489.33	7.65	0.08
	EPIC (Frenda et al., 2023)	74	191.51	4.72	0.19

Table 1: PEFT-U Benchmark: We design a large-scale benchmark for personalized model training and evaluation consisting of 13+ personalized tasks across 15k+ users with each task obtaining a Krippendorff's alpha (α) < 0.5 per task. *Subjective Discourse consists of 3 different sentiment tasks.

sonalized tasks.

4 Benchmark Evaluation

To quantify the challenge the PEFT-U benchmark presents, we evaluate the performance of a range of parameter-efficient methods compared to zero/few-shot prompting approaches.

Methods We implement and evaluate 7 different parameter-efficient methods for personalizing LLMs using Flan-T5 (Chung et al., 2022). These methods consist of:

1) **Zero-shot/Few-shot Prompting**: Using $k = 3$ random samples of user data we construct an instruction-style prompt for inference.

2) **LoRa** (Hu et al., 2021): injects trainable rank decomposition matrices into each layer of the Transformer architecture.

3) **Adapters** (Houlsby et al., 2019) add a trainable bottleneck layer after the feedforward network in each Transformer layer.

4) **Prompt Tuning** (Lester et al., 2021) introduces an additional k tunable tokens per downstream task prepended to the input text and trained end-to-end.

5) **Prefix-Tuning** (Li and Liang, 2021) prepends task-specific trainable vectors to the input of multi-head attention in each Transformer layer that is attended to as virtual tokens.

6) **P-Tuning** (Liu et al., 2022b) employs trainable continuous prompt embeddings in concatenation with discrete prompts.

7) **IA^3** (Liu et al., 2022a) introduces trainable vectors l_w into different components of the transformer which perform element-wise rescaling of inner model activations.

Training We train all models with AdamW (Loshchilov and Hutter, 2019) and a weight decay of 0.01 on NVIDIA RTX 3090 24GB GPUs. We use a learning rate of 2e-5, batch size of 16, and a linear learning rate warmup over the first 10% steps with a cosine schedule for 8 epochs over multiple runs with varied random seeds.

5 Results

Evaluation Metrics We consider two performance metrics: (1) average per-user accuracy per task and (2) average accuracy for all tasks.

5.1 Few/Zero-shot vs PEFT

Table 2 shows our results analyzing existing PEFT methods in comparison to few/zero-shot prompting techniques. From these results, we show that personalizing models is crucial to providing users with more accurate results representative of their actual perspectives. Notably, zero/few-shot prompting falls short of adequately representing user viewpoints compared to its trained counterparts being outperformed on average by all methods except for Prompt Tuning. Across all methods, results show that Adapters performs the best outperforming on 12 out of the 13 PEFT-U tasks and achieving an overall accuracy score of 64.4% compared to 59.5% of LoRa in 2nd. The presented results underscore the complexity of the PEFT-U benchmark, revealing the challenges inherent in achieving consistently high performance across diverse tasks and datasets. While personalized fine-tuning methods exhibit superior accuracy compared to traditional

Method	Size	#Params	Dataset													Average
			Hate+Abuse						Sentiment		Humor		Sentiment			
			Hate Xplain	Gab Hate	MHS	Tweet Eval	UHC	Wiki Detox	Go Emotion	Stud Emo	Cockamamie	EPIC	SD(D)	SD(QS)	SD(RS)	
Zero/Few-Shot (k=3)	-	0	47.4	81.9	26.1	61.5	55.2	61.4	53.3	27.2	97.1	63.3	16.0	20.5	11.3	47.9
LoRA	3.8M	880K	54.9	88.6	27.4	69.5	73.4	81.0	66.3	67.7	97.2	67.9	29.4	30.7	18.9	59.5
Prefix Tuning	1.5M	370K	48.8	85.4	**45.5**	61.8	66.0	72.7	60.1	32.9	97.1	66.7	7.7	11.7	8.1	51.1
P-Tuning	128K	1.8M	48.7	81.9	29.2	61.3	56.8	62.3	53.8	27.6	97.0	62.8	16.6	19.6	11.1	48.4
Prompt Tuning	128K	31K	49.3	82.8	28.7	59.5	56.3	63.3	52.1	27.6	97.1	61.1	10.5	20.5	9.9	47.6
Adapters	14M	3.5M	**59.0**	**89.1**	38.9	**70.5**	**77.5**	**84.0**	**68.4**	**83.7**	**97.2**	**68.5**	**35.3**	**39.0**	**25.6**	**64.4**
IA^3	450K	111K	48.6	86.7	26.5	62.3	61.6	70.0	58.9	27.6	97.1	64.3	19.0	24.1	13.5	50.8

Table 2: Results of PEFT Methods on the PEFT-U Benchmark: This table shows the macro accuracy of each PEFT method in comparison to Zero/Few-shot prompting on Flan-T5 (Chung et al., 2022).

Method	Original Params	Adjusted Params	Original Acc	Acc w/ Reduced Params
LoRA	880K	111K	*69.5*	**66.2**
Prefix Tuning	370K	111K	*61.8*	61.5
P-Tuning	1.6M	111K	*61.7*	61.7
Prompt Tuning	15K	111K	*59.8*	50.0
Adapters	3.5M	111K	*70.5*	64.2
IA^3	111K	-	*62.3*	-

Table 3: Results on TweetEval task for PEFT-U with equal number of trainable parameters.

few/zero-shot prompting techniques, the variations in performance among different PEFT methods as well as the performance on datasets such as Subjective Discourse and MeasuringHateSpeech indicate that the benchmark presents a multifaceted challenge. The nuances of user personalization, model size, and parameter tuning significantly impact the effectiveness of these methods. This observed diversity in performance across methods suggests that there is no one-size-fits-all solution, and further investigation is imperative.

5.2 Impact of Number of Parameters

Given the performance of Adapters, we sought to understand whether its performance was due to the number of trainable parameters. As such, we systematically varied the parameter count across each method on the TweetEval Task. Notably, we observed nuanced patterns across different PEFT methods. As shown in table 3, with reduced parameters all methods except for P-tuning suffered a decrease in overall performance. However, LoRa with equal trainable parameters was able to outperform Adapters.

6 Related Works

Prior works have highlighted the need for user perspective particularly when dealing with intrinsically subjective applications where differing perspectives are common (Mostafazadeh Davani et al., 2022; Sang and Stanton, 2021; Geva et al., 2019; Kanclerz et al., 2022; Welch et al., 2022). Research has shown that accounting for user perspective and personalization is essential for building robust and effective models (Mostafazadeh Davani et al., 2022; Sang and Stanton, 2021; Geva et al., 2019; Kanclerz et al., 2022; Welch et al., 2022). However, despite this glaring need, existing resources fail to model and cater to these differing perspectives. When curated, NLP resources tend to have an intrinsic bias towards the majority perspective due to their reliance on voting for ground truth. As such they fail to adequately represent diverse user preferences and individualized expressions, further contributing to a lack of personalization (Mostafazadeh Davani et al., 2022; Sang and Stanton, 2021; Geva et al., 2019; Kanclerz et al., 2022). Other benchmarks such as Salemi et al. (2023) highlight the importance of personalization LLMs, however, PEFT-U uniquely factors cases of conflicting user perspectives when exploring personalization in addition to considering the compute constraints.

7 Conclusion

This work addresses a critical gap in NLP concerning the personalization of LLMs. While LLMs have achieved remarkable performance in various tasks, their generalization capabilities have predominantly followed a "one-size-fits-all" paradigm. This approach falls short of meeting the diverse linguistic and communicative preferences of individual users. The PEFT-U Benchmark introduced in this paper serves as an effort to evaluate the personalization capabilities of LLMs across a spectrum of tasks. PEFT-U, presents a unique challenge by emphasizing scenarios where identical inputs necessitate diverse model outputs. The reported results showcase the inherent challenges posed by the PEFT-U benchmark and advocate for the continued exploration of effective personalization strategies.

8 Limitations

The PEFT-U Benchmark while designed to capture diverse user perspectives, may not fully represent the intricacies of all real-world communication scenarios. The dataset's construction involved a careful curation process, but the authors acknowledge that the complexities of individual preferences and linguistic nuances are vast and varied. In this work, user perspective is modeled solely based on the user's output preferences. Factors such as age, gender, and other potentially important demographics are not considered.

In addition, the personalization methodologies explored in this study may not encompass the entire spectrum of potential approaches. The field of NLP is dynamic, and emerging techniques could offer alternative solutions to the challenges presented. Personalization in LLMs is an evolving research area, as such there may be relevant strategies released recently that were not covered in this work.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376, Toronto, Canada. Association for Computational Linguistics.

Christopher Clarke, Joseph Peper, Karthik Krishnamurthy, Walter Talamonti, Kevin Leach, Walter Lasecki, Yiping Kang, Lingjia Tang, and Jason Mars. 2022. One agent to rule them all: Towards multi-agent conversational AI. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3258–3267, Dublin, Ireland. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

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, Online. Association for Computational Linguistics.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Limor Gultchin, Genevieve Patterson, Nancy Baym, Nathaniel Swinger, and Adam Kalai. 2019. Humor in word embeddings: Cockamamie gobbledegook for nincompoops. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2474–2483. PMLR.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling laws for autoregressive generative modeling.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages

37–45, Marseille, France. European Language Resources Association.

Yiping Kang, Ashish Mahendra, Christopher Clarke, Lingjia Tang, and Jason Mars. 2022. Towards personalized intelligence at scale.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Y. Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Aadila Hussain, A. Lara, O G., Asmaa Al Omary, C. G. Park, C. Wang, X Wang, Y. Zhang, and Morteza Dehghani. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach llms to personalize – an approach inspired by writing education.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Anh Ngo, Agri Candri, Teddy Ferdinan, Jan Kocon, and Wojciech Korczynski. 2022. StudEmo: A non-aggregated review dataset for personalized emotion recognition. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 46–55, Marseille, France. European Language Resources Association.

OpenAI. 2023. Gpt-4 technical report.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2023. Modular deep learning.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization.

Yisi Sang and Jeffrey Stanton. 2021. The origin and value of disagreement among data labelers: A case study of the individual difference in hate speech annotation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten,

Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Additional Dataset Details

In this section, we detail the datasets in our PEFT-U benchmark, including dataset construction, representative samples, and task instructions.

A.1 Dataset Details & Construction

We include datasets in various domains, including:

- **HateXplain** (Mathew et al., 2022) contains posts on social media. Each post is classified into 3 classes: hate, offensive, or normal. The dataset additionally contained annotations for the hate speech target community and the rationales. We consider the post texts and the classification labels only.
- **GabHate** (Kennedy et al., 2018) has 28K posts from the social media platform Gab. Each post is annotated using a hierarchical coding typology of hate-based rhetoric, with hierarchical labels indicating dehumanizing and violent speech, vulgarity and offensive language, and targeted groups. We only consider the top-level binary classification on hate speech.
- **MeasuringHateSpeech** (Sachdeva et al., 2022) contains 50K social media comments from various platforms. They are labeled by a large number (11K) of Crowdsource workers on Amazon Mechanical Turk[2]. Each comment is annotated with 10 ordinal labels: sentiment, disrespect, insult, attacking/defending, humiliation, inferior/superior status, dehumanization, violence, genocide, and a 3-valued hate speech benchmark label. This dataset adjusts for annotators' perspectives by aggregating the labels via faceted Rasch measurement theory (RMT). We use the comment text and the 3-valued hate speech label only.
- **TweetEval** (Röttger et al., 2022) comprises 200 Twitter posts, each annotated by 20 annotator groups of 3. Annotators were given a short definition of hate speech only, which encourages the subjectivity of annotators. The labels are binary classifications of hatefulness.
- **UnhealthyConversations** (Price et al., 2020) consists of 44K comments labeled by crowdsource workers as either "healthy" or "unhealthy". It also contains six potentially unhealthy sub-attributes: (1) hostile; (2) antagonistic, insulting, provocative, or trolling; (3) dismissive; (4) condescending or patronizing; (5) sarcastic; and/or (6) an unfair generalization. We only consider the top-level binary classification of healthy conversations.
- **WikiDetox Aggression** (Wulczyn et al., 2017) is a large collection of 100K online comments to English Wikipedia where crowd-source annotators label whether each comment is a personal attack.
- **GoEmotion** (Demszky et al., 2020) is a large annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We reduce the emotion categories into 6 coarse categories with Ekman-style grouping. Each comment can have multiple emotion labels. We drop texts with no labels annotated and texts labeled as 'neutral'.
- **StudEmo** (Ngo et al., 2022) comprises 5K customer reviews annotated by 25 people for 10 emotion dimensions: eight emotion dimensions from Plutchik's model plus valence and arousal. Valence has an intensity range of $[-3, +3]$ whereas each remaining category has a range of $[0, 3]$. We treat the problem

[2] https://www.mturk.com

Dataset		# Unique Texts	Labels
Domain	**Name**		
Hate+Abuse	HateXplain	20K	[hateful, offensive, normal]
	GabHate	28K	[Hateful, Non-hateful]
	Measuring HateSpeech	50K	Hate speech scale: [0, 1, 2]
	TweetEval	200	[Hateful, Non-hateful]
	Unhealthy Conversations	44K	[healthy, unhealthy]
	WikiDetox Aggression	100K	[Aggressive, Normal]
Sentiment	GoEmotion	58K	[anger, disgust, fear, joy, sadness, surprise]
	StudEmo	5K	[joy, trust, anticipation, surprise, fear, sadness, anger, disgust, valence, arousal]
	Subjective Discourse (response) Subjective Discourse (question sentiment) Subjective Discourse (response sentiment)	1K	[answer+direct, answer+over-answer, shift+correct, shift+dodge, can't answer+honest, can't answer+lying] [very-negative, negative, somewhat-negative, neutral, somewhat-positive, positive, very-positive] [very-negative, negative, somewhat-negative, neutral, somewhat-positive, positive, very-positive]
Humor	Cockamamie	120K	[humorous, not-humorous]
	EPIC	3K	[Ironic, Non-ironic]

Table 4: Additional details on the PEFT-U Benchmark.

as multi-class classification where we keep categories with intensity value ≥ 1 as positive labels.

- **Subjective Discourse** (Ferracane et al., 2021) consists of 1,000 question-response pairs from 20 witness testimonials in U.S. congressional hearings. The study collects subjective interpretations from the crowdsource workers about the conversations on three aspects: the question, the response, and an explanation. In particular, the annotator provides subjective assessments of the conversation acts and communicative intents of the responses, forming 6 response labels. Each annotator also rates their sentiment toward the politicians and the witness on a 7-point scale. We leverage the response labels, and the sentiments toward the politicians and witnesses to form 3 dataset versions. To construct the text part, we join (question speaker detail, question text, response speaker detail, response text) by newlines.

- **Cockamamie** (Gultchin et al., 2019) includes 120K words and phrases from GNEWS. The words and phrases are annotated by crowd workers fluent in English on whether they are humorous. 1,500 words are further annotated on six binary humor features. We leverage the words and the initial binary "humorous" labels only.

- **EPIC** (Frenda et al., 2023) is made of 3K short social media (post, reply) pairs collected from Twitter and Reddit across five regional varieties of English. Annotators from five different counties were asked to provide a binary label (either *Irony* or *not-Irony*) for the *Reply* text given the context provided by *Post*. We template the (Post, Reply) pairs by adding "message" and "reply" prefixes.

We summarize additional details for each dataset in Table 4. We split each user's data into train/dev/test sets by 80/10/10 splits.

A.2 Representative Samples

In this section, we show representative samples for each dataset where different user perspectives result in different labels for the same text input.

Warning: Offensive content is inevitable for datasets in the Hate+Abuse domain.

A.2.1 Representative Sample for HateXplain

Text. right definitely not going back to the fag hag thing it was fun for a bit in my <number> but outrageously gay people are just too high strung for me it back to that reason i can not be a lesbian
Labels.
- normal
- hatespeech
- offensive

A.2.2 Representative Sample for MeasuringHateSpeech

Text. @L1LTR4P fucking losers wetbacks #SorryNotSorry
Labels.
- 0
- 1
- 2

A.2.3 Representative Sample for GabHate

Text. #Breaking: Claiming bias against conservatives, @FreedomWatchUSA has filed a class action lawsuit against Google, Facebook, Instagram, Twitter, and Apple.
Labels.
- Non-hateful
- Hateful

A.2.4 Representative Sample for TweetEval

Text. [USER] fuck Brett Farve redneck ass, he stuckup he don't give a damn lol he be on campus acting like he the shit
Labels.
- Hateful
- Non-hateful

A.2.5 Representative Sample for Unhealthy Conversations

Text. They are poor because they are reliant on the drug trade and reliant on the drug trade because they are then poor. That cycle can be broken.
Labels.
- healthy
- unhealthy

A.2.6 Representative Sample for WikiDetox Aggression

Text. == Dougbiznatch again! ==

Hey I'm back. Gonna vandalize all day and no one can stop me! As you can tell I can't be stopped by banning. I'll be seeing alo tof you and the rest of the blacklisted admins for the next couple of weeks =P
Dougbiznatch
Labels.
- Aggressive
- Normal

A.2.7 Representative Sample for GoEmotion

Text. Is he dead now from a tragic drowning accident? Asking for a friend.
Labels.
- [sadness]
- [surprise]

A.2.8 Representative Sample for StudEmo

Text. We got to the Cycle On Hostel by chance in the middle of the night. There wasn't a single place available in other places. . . ...and it's very good that we didn't get to another place! First of all, great service: people who are open to others, nice and smiling, who help us with every time of day and night. Spacious hostel, rooms and bathrooms are clean. And besides all that - the location - nothing to add, nothing to take away. Literally 5 minutes away from Neptune and at the same time the building is situated in such a place that at night it is quiet despite such a short distance from the busiest street where it is full of tourists and children. If we will still have a chance to spend the night in Gdansk, we will surely come to Cycle On again. With a clear conscience I recommend
Labels.
- [trust, anticipation, valence, arousal]
- [joy, trust, valence, arousal]

A.2.9 Representative Sample for Subjective Discourse (response)

Text. politician: JIM JORDAN, Ohio

Okay. Well, this is put out by the Exempt Organizations Division, same division where all these problems took place over the last three years. It came out, again, just five days after the comment period on the proposed rule ended at the end of February, and I want to just highlight a few of the questions that are asked. So there is a category that says what if the IRS needs more information about your (c)(4) application? New sample questions. So if we could put them up side-by-side. Now, the first slide are the targeted questions that TIGTA said were inappropriate and that you agree are inappropriate, Judith Kendall agreed are inappropriate.

Those are those questions. And now, just five days after the proposed rule comment period ends, you issue a newsletter from the Exempt Organizations Division highlighting the new questions you are going to ask, and I just want to look how similar the two questions are. Let's just take the second category, whether an officer or director, etcetera, has run or will run for public office. The new question says this: Do you support a candidate for public office who is one of your founders, officers, or board members? It is basically the same. This reminds me of when I was in grade school and the teachers told us you shouldn't plagiarize, so you change a few words and basically plagiarize. This is the same thing. So here is what I don't understand. If you are trying to comply with the TIGTA report, if the new (c)(4) rule is a way to deal with what the audit said and not as what I believe is a continuation of the project Lois Lerner started, why are you asking the same darn questions?

witness: The Hon. John Koskinen, Commissioner, Internal Revenue Service

As I noted, I haven't seen that and can't read it on the chart. I would be delighted to sit down and go over all of those questions with you and with the exempt organizations. All of the TIGTA report didn't blanket say you should never ask questions about this. Thank you for the chart.

Labels.

- can't answer+lying
- can't answer+honest
- shift+dodge

A.2.10 Representative Sample for Subjective Discourse (question sentiment)

Text. politician: RANDY NEUGEBAUER, Texas

And, as you're aware, Section 972 of Dodd-Frank requires an issuer of securities to disclose the annual proxy statement, the reason why the issuer has chosen to allow the same person to serve as the board chairman and the CEO. This year, Wells states that your dual role is a result of your extensive experience and knowledge regarding the company and provides the most efficient leadership of the board and the company. Mr. Stumpf, do you think it's a good idea for the same person to serve as both chairman of the board and CEO?

witness: Stumpf, John G., Chairman and Chief Executive Officer, Wells Fargo and Company

Thank you, Congressman, for that question. In our company, we have 14 outside directors. We have a lead director. All directors are New York Stock Exchange-independent, by their standards. I am not a member of any standing committee of that board. The independent directors and the lead director help set the agenda for the boards. They always have meetings that are in executive session without me. And as you probably read about what happened this weekend–because we filed an 8-K yesterday about actions that they took as an independent board, and I was not part of that. So the board acts quite independently.

Labels.

- negative
- neutral
- somewhat-negative
- somewhat-positive

A.2.11 Representative Sample for Subjective Discourse (response sentiment)

Text. politician: JERROLD NADLER, New York, Chairman

It is our understanding that at least one briefing occurred between December–between your decision not to recuse yourself on December 19th and 6 days later, Christmas day. Is that correct? Simple enough question, yes or no.

witness: Matthew G. Whitaker, Acting Attorney General, U.S. Department of Justice Oral Testimony

Mr. Chairman, I, again, what is the basis for your question? You are saying that it is your understanding—-

Labels.

- negative
- neutral
- somewhat-negative
- somewhat-positive

A.2.12 Representative Sample for Cockamamie

Text. judokas

Labels.

- humorous
- not-humorous

A.2.13 Representative Sample for EPIC

Text. message: "Because the last generation was Gen. Z and there's no where to go past that, so we gotta start from the beginning."
reply: "but we should have just named the first generation "Alpha" instead of doing it now"
Labels.
- Ironic
- Non-ironic

A.3 Task Instructions

In this section, we show the task instructions prepended to prompts in training and inference for each dataset.

A.3.1 Instruction for HateXplain

Please review the following text and indicate if it has the presence of hate speech or is offensive in nature. Respond 'hatespeech' if the text contains hate speech, 'offensive' if the text does not contain hate speech but is offensive and 'normal' if the text is neither offensive nor contains hate speech.

A.3.2 Instruction for MeasuringHateSpeech

On a scale of 0-2 rate the level of hatespeech contained in the following text. Respond with '0' if the text does not contain hate speech, '1' if the text contains some hate speech, and '2' if the text contains a lot of hate speech.

A.3.3 Instruction for GabHate

Please review the following text and indicate if it has the presence of hate speech. Respond 'Hateful' if the text contains hate speech and 'Non-hateful' if the text does not contain hate speech.

A.3.4 Instruction for TweetEval

Please review the following text and indicate if it has the presence of hate speech. Respond 'Hateful' if the text contains hate speech and 'Non-hateful' if the text does not contain hate speech.

A.3.5 Instruction for Unhealthy Conversations

Please review the following text and indicated if it is 'healthy' or 'unhealthy'. Respond 'healthy' if the text is healthy and 'unhealthy' if the text can be considered hostile, antagonistic, condescending, dismissive or an unfair generalization.

A.3.6 Instruction for WikiDetox Aggression

Please review the following text and indicate if it has the presence of malicious remark to a person or group. Respond 'Aggressive' if the text contains a personal attack and 'Normal' if the text does not contain a personal attack.

A.3.7 Instruction for GoEmotion

Please analyze the following text and assign one or more appropriate emotion labels. Emotion labels include happiness, sadness, anger, surprise, joy, fear, disgust. You can select one or multiple emotion labels that best capture the emotional content of the text. Respond with the emotion labels separated by a comma.

A.3.8 Instruction for StudEmo

Please analyze the following text and assign one or more appropriate emotion labels. Emotion labels include joy, trust, anticipation, surprise, fear, sadness, disgust, anger, valence, and arousal. You can select one or multiple emotion labels that best capture the emotional content of the text. Respond with the emotion labels separated by a comma.

A.3.9 Instruction for Subjective Discourse (response)

Please analyze the following text and indicate how the witness responded to the question. Respond with 'answer' if they answered the question reasonably, 'cant-answer-lying' if they could not answer and are lying, 'cant-answer-sincere' if they could not answer but are honest about it, 'shift-dodge' if they shifted the topic with the intent of dodging the question, 'answer_overans-sway' if they over answered the question with the intention of swaying or 'shift-correct' if they shifted the topic with the intention of clarifying the question.

A.3.10 Instruction for Subjective Discourse (question sentiment)

Please analyze the following text and rate your sentiment towards the questioners. Sentiment labels include 'somewhat-positive', 'positive', 'very-positive', 'somewhat-negative', 'very-negative', 'neutral' and 'negative'. Respond with the sentiment label that best captures your sentiment towards the questioners.

A.3.11 Instruction for Subjective Discourse (response sentiment)

Please analyze the following text and rate your sentiment towards the witness. Sentiment labels include 'somewhat-positive', 'positive', 'very-positive', 'somewhat-negative', 'very-negative',

'neutral' and 'negative'. Respond with the sentiment label that best captures your sentiment towards the witness.

A.3.12 Instruction for Cockamamie

Please rate whether the following text is funny or not funny. Respond 'yes' if you think the text is funny and 'no' if you think the text is not funny.

A.3.13 Instruction for EPIC

Irony is a figurative language device that conveys that opposite of literal meaning, profiling intentionally a secondary or extended meaning. Please review the following message and reply and indicate if it has the presence of irony. Respond 'Ironic' if the reply if you think the reply is ironic and 'Non-ironic' if you think the reply is not ironic.