SOCIALITE-LLAMA: An Instruction-Tuned Model for Social Scientific Tasks

Gourab Dey[1*]**, Adithya V Ganesan**[1*]**, Yash Kumar Lal**[1*]**,**
Manal Shah[2]**, Shreyashee Sinha**[3]**, Matthew Matero**[4]
Salvatore Giorgi[5,6]**, Vivek Kulkarni**[7]**, H. Andrew Schwartz**[1]
[1]Stony Brook University, [2]JP Morgan Chase, [3]Bloomberg
[4]Preferabli, [5]University of Pennsylvania, [6]National Institute of Health, [7]Grammarly
*{gdey,avirinchipur,ylal}@cs.stonybrook.edu

Abstract

Social science NLP tasks, such as emotion or humor detection, are required to capture the semantics along with the implicit pragmatics from text, often with limited amounts of training data. Instruction tuning has been shown to improve the many capabilities of large language models (LLMs) such as commonsense reasoning, reading comprehension, and computer programming. However, little is known about the effectiveness of instruction tuning on the social domain where implicit pragmatic cues are often needed to be captured. We explore the use of instruction tuning for social science NLP tasks and introduce SOCIALITE-LLAMA— an open-source, instruction-tuned LLAMA2. On a suite of 20 social science tasks, SOCIALITE-LLAMA improves upon the performance of LLAMA2 as well as matches or improves upon the performance of a state-of-the-art, multi-task finetuned model on a majority of them. Further, SOCIALITE-LLAMA also leads to improvement on 5 out of 6 related social tasks as compared to LLAMA2, suggesting instruction tuning can lead to generalized social understanding. All resources including our code, model and dataset can be found through bit.ly/socialitellama.

1 Introduction

Instruction finetuned large language models (LLMs) have demonstrated impressive performance on many standard NLP tasks (Wei et al., 2022a; Chung et al., 2022), but these models tuned on non-social tasks seem to have poor social pragmatics (Ziems et al., 2023; Choi et al., 2023; Havaldar et al., 2023; V Ganesan et al., 2023). The fact that instruction tuning has been successful with limited amounts of data (Gupta et al., 2023) and that instruction tuned models have the ability to generalize to new tasks in both few- and zero-shot settings (Wei et al., 2022a), suggests that LLMs

could become more socially capable by instruction tuning them on a wide variety of social NLP tasks.

Here, we introduce SOCIALITE-LLAMA, a LLAMA2 7B-based LLM (Touvron et al., 2023) instruction tuned on a suite of social scientific classification tasks spanning 5 broad categories for which we hand-craft instructions. We evaluate its performance on both zero- and few-shot settings on seen and related social tasks, demonstrating that our model significantly outperforms prior open models. Our results support prior research highlighting the effectiveness of instruction tuning when applying LLMs to a new domain.

Social and psychological factors have been shown important and beneficial to model in past interdisciplinary NLP studies (Lynn et al., 2017; Flek, 2020; Hovy and Yang, 2021). Modeling human factors and social context can not only improve performance on primarily non-social NLP tasks (Lynn et al., 2017; Flek, 2020; Hovy and Yang, 2021), but can also prove to be beneficial for a number of psychological and social scientific tasks (Garten et al., 2019; Matero et al., 2021). However, a major limitation of prior models has been that they have been task-specific and do not generalize well to new tasks. Many such models also are trained from scratch and typically need an extensive amount of human-annotated training data. With the development of LLMs that exhibit the capability to learn from instructions, we posit that this new capability can be leveraged to address these two major limitations.

Our **contributions** include: (1) we develop and systematically evaluate SOCIALITE-LLAMA– an instruction-tuned language model for social science tasks – across 20 seen and 6 related social scientific tasks, (2) we show SOCIALITE-LLAMA consistently improved over LLAMA2 in all seen tasks and the improvement generalized to 5 of 6 related tasks; In fact, it matched the performance of a state-of-the-art multi-task tuned DEBERTA

---
*Equal contribution

on a majority of seen tasks, (3) we suggest that the benefits of few-shot examples (over zero-shot) become negligible on tasks seen during instruction tuning as opposed to related tasks where few-shot still provided a benefit, and (4) we release SOCIALITE-LLAMA[1] as well as its instructions corpora, SOCIALITEINSTRUCTIONS[2], as open-source resources for the community.

2 Related Work

Language is inherently social — the underlying meaning is constructed through social interactions (Wittgenstein, 1953; Clark and Schober, 1992; Hovy and Spruit, 2016). Understanding communication requires reasoning about the social implications drawn from that message (Halliday, 2004). Prior work has sought to build language models for social scientific tasks, which we discuss next.

LMs for Social Scientific NLP Social science NLP models are usually built by fine-tuning (Sap et al., 2020; Matero et al., 2019; V Ganesan et al., 2021) for specific tasks or pretraining language models on the the domain of language that captures the social factors. BERTweet (Nguyen et al., 2020), the first public large-scale pre-trained language model on English tweets, was trained using the BERT architecture with the RoBERTa pre-training procedure (Liu et al., 2019) for this purpose. DeLucia et al. (2022) built Bernice, a multilingual model for social science NLP tasks using the RoBERTa architecture and pretrained from scratch on 2.5 billion tweets. It outperforms a variety of models adapted to social media data as well as strong multilingual baselines. Despite the strength of such existing task-specific models, there exists no base instruction-following LLM tailored to the domain of social science tasks.

Instruction-Tuning Instruction tuning, in general, refers to the practice of finetuning pre-trained language models to better understand and respond to a wide variety of human requests that are expressed in natural language (Wei et al., 2022a; Mishra et al., 2022). The success of instruction tuning requires two key components: 1) a powerful pre-trained language model like LLAMA2 (Touvron et al., 2023), and 2) an instruction dataset that is diverse and representative enough to adapt the LLM to potential downstream usage. However,

existing instruction-tuning datasets (Longpre et al., 2023; Sanh et al., 2022; Wang et al., 2023, 2022) are general-purpose and do not contain a significant amount of social science tasks. Consequently, models built using this data have limited usability for social scientific tasks (Choi et al., 2023; Ziems et al., 2023). To address this gap, we curate an instruction-tuning dataset aimed at modeling social scientific knowledge and use it to train our model.

3 Datasets and Experiments

We use a diverse collection of English social science NLP tasks to create SOCIALITEINSTRUCTIONS by (re-)framing each task into a binary or multi-class classification problem. For each task, we hand-craft instructions for the model to follow. These instructions were composed with task descriptive which included examples in some cases (e.g. BIASEDIMPLICATION from Table 9), followed by the classification label choices it has to pick from. The instructions and the dataset sizes for all the tasks have been tabulated in Appendix A.

3.1 Training Tasks

We draw on datasets mentioned in Choi et al. (2023) for compiling our instruction following dataset. Our train set spans 20 different datasets across 5 broad categories of tasks: Humor, Offensiveness, Sentiment and Emotion, Trustworthiness, and other social factors. To ensure broader applicability, we frame the non-classification datasets as classification tasks. We refer to these as seen tasks going forward.

Humor We use SemEval 2021 Task 7 data (Meaney et al., 2021) to capture humor. We use the binary humor detection task as is (HUMOR). Additionally, we convert the humor rating task into a binary classification problem collapsing ground truth labels higher than 3 into high humor, and the others into low humor (HUMOURRATING).

Offensiveness We use SOCIAL BIAS FRAMES (Sap et al., 2020) as a benchmark to detect offense and bias directed towards people and groups. SOCIAL BIAS FRAMES comprises of 4 binary classification tasks - OFFENSIVE to discern whether a given text exhibits rudeness, disrespect, or toxicity, SEXIST to determine if a text contains lewd or sexual references which can be considered as offensive, INTENTTOOFFEND to capture whether the perceived motivation of the author was to indeed offend, and BIASEDIMPLICATION to identify any

forms of prejudice or group-based discrimination within the text.

Sentiment and Emotion For emotion classification, we use SemEval 2018 Task 1 data (Mohammad et al., 2018). EMOTION involves classifying a tweet into anger, joy, optimism or sadness displayed by the author. We also include Emobank (Buechel and Hahn, 2017) to infer the valence, arousal and dominance levels that a text would invoke in a reader. We convert these regression tasks to binary classification - by transforming a score greater than 4 to 'high' and less than 4 to 'low', for each of the three datasets respectively (VALENCECLS, AROUSALCLS, DOMINANCECLS). We use SemEval 2017 Task 4 data for sentiment classification (Rosenthal et al., 2017) to identify the overall sentiment of a text as positive, negative or neutral. Further, we also include SAMESIDESTANCE (Körner et al., 2021), the task of classifying whether two different texts are on the same side of an argument.

Trustworthiness We use HYPERBOLE (Zhang and Wan, 2022) to detect hyperbolic language present in a text, and SUBJECTIVEBIAS (Pryzant et al., 2020), to detect which sentence exhibits subjective bias among two pieces of text.

Other Social Factors We also include other discursive and rhetorical type tasks — empathy scored on multi-item scales (EMPATHYEXPLORATIONS) (Sharma et al., 2020), self-rated empathy (EMPATHYSELFRATED) (Buechel et al., 2018), distress (DISTRESSSELFRATED) (Buechel et al., 2018), figurative speech detection (FLUTE) (Chakrabarty et al., 2022), politeness (POLITENESSHAYATI) (Hayati et al., 2021) and intimacy (INTIMACY) (Pei and Jurgens, 2020).

3.2 Evaluation Tasks

For evaluation, we use the same twenty seen datasets as well as six datasets for related social tasks. The process we used to choose the 'seen tasks' for instruction tuning was towards the goal of training on a very broad category of tasks and consequently, any task would be related to these fundamental categories seen during training. Hence for evaluation we choose six tasks, half of which are more directly related to seen tasks than the other. The datasets used for evaluation contained no data overlap with the training tasks' data. Among the related tasks' datasets, we first select three datasets directly related to our training tasks

— HATESPEECH (Basile et al., 2019) to detect hate speech in a tweet, POLITENESSSTANFORD (Fu et al., 2020) to judge whether a text is polite or impolite, and AGREEDISAGREE (Varadarajan et al., 2022) to classify whether two texts agree, disagree or N/A on a particular topic. To strengthen our evaluation, we also evaluate on three other tasks which are less related to the seen set — IRONY (Van Hee et al., 2018) to classify if a tweet is ironic, OPTIMISM (Ruan et al., 2016) to categorize a text as optimistic, pessimistic or neutral, and COMPLAINTS (Preoţiuc-Pietro et al., 2019) to judge if a text is a complaint or not. We refer to these six as related social tasks.

3.3 Task Selection Criterion

The above tasks were chosen from SOCKET (a) to be representative of each social category and (b) if the document lengths weren't too long, i.e., utterance level tasks (Ziems et al., 2023).

Our initial experiments suggested that skew in sample sizes from different tasks affected the overall performance. Particularly, using all 35k examples from SOCIAL BIAS FRAMES for all the four tasks limited the performance on POLITENESSHAYATI and SAMESIDESTANCE. Hence, we picked datasets from SOCKET based on 2 additional factors: recency and minimum number of 8k examples. This is why we chose Sap et al. (2020) over other toxicity datasets (further down sampled 8k examples for each task), and Rosenthal et al. (2017) over Socher et al. (2013) for sentiment task.

3.4 Training Procedure

We transform each of the 20 training datasets using hand-crafted instructions. To reduce the skew caused by the disproportionate number of examples present in a few tasks, we limit the number of training examples from OFFENSIVE, SEXIST, INTENTTOOFFEND, BIASEDIMPLICATION, SENTIMENT and SUBJECTIVEBIAS to 8k and use the original number of training data points for the rest. This resulted in ~108k data points. We employ Low-Rank Adaptation (Hu et al., 2022), with rank set to 8 applied on the key and query projection matrices reducing the number of trainable parameters to 4.1 million, and use AdamW (Loshchilov and Hutter, 2019) with a learning rate of $\eta = 1e-4$ and a batch size of 64 for a maximum of 7 epochs with early stopping. We expand the context length to 3k using mixed precision training, gradient checkpointing and DeepSpeed Zero (Rajbhandari et al.,

TASK	SOCIALITE-LLAMA		LLAMA2	
	Few-shot	Zero-shot	Few-shot	Zero-shot
SEEN TASKS				
SENTIMENT	72.69	**73.43**	46.40	18.28
EMOTION	79.68	**81.08**	66.84	23.92
VALENCECLS	82.34	**83.84**	44.52	31.59
AROUSALCLS	60.69	**64.22**	52.69	34.42
DOMINANCECLS	60.88	**66.63**	43.05	19.33
EMPATHYEXPLORATIONS	30.10	**56.99**	29.43	5.89
EMPATHYSELFRATED	63.59	**63.97**	58.99	8.75
DISTRESSSELFRATED	68.45	**70.96**	42.36	9.80
FLUTE	80.77	**96.27**	55.13	4.81
HYPERBOLE	65.62	**69.23**	59.20	25.43
SAMESIDESTANCE	**88.19**	**88.19**	82.10	25.00
HUMOR	90.48	**95.29**	44.44	38.8
HUMOURRATING	**52.81**	48.66	43.51	34.84
POLITENESSHAYATI	**89.36**	84.54	83.58	24.34
INTIMACY	26.25	**33.21**	13.24	3.17
SUBJECTIVEBIAS	52.04	**87.13**	51.18	40.98
OFFENSIVE	80.47	**83.11**	71.78	35.76
SEXIST	72.61	**73.85**	52.81	6.25
INTENTTOOFFEND	74.17	**75.65**	32.09	21.34
BIASEDIMPLICATION	79.66	**85.72**	56.41	26.16
RELATED TASKS				
HATESPEECH	42.13	39.63	**68.66**	31.34
IRONY	42.99	**59.04**	47.11	32.24
POLITENESSSTANFORD	63.22	**64.89**	58.64	1.2
OPTIMISM	**62.24**	36.96	56.61	44.49
COMPLAINTS	79.35	**85.37**	58.23	22.8
AGREEDISAGREE	**39.54**	19.3	28.1	10.01

Table 1: Evaluation of SOCIALITE-LLAMA and LLAMA2 on 20 seen tasks and 6 unseen tasks using macro-F1 scores. We report few-shot performance with $k = 5$ for SOCIALITE-LLAMA on seen tasks and $k = 15$ on unseen tasks. For LLAMA2, we use $k = 15$ for all tasks since it is not trained for social scientific tasks. k refers to the number of examples used in the few-shot setting. All differences between SOCIALITE-LLAMA and LLAMA2 are statistically significant ($p < 0.001$). **Bold** indicates the performance with the best setting for that task.

2020) Stage 2 which enables finetuning on 2 48GB A6000 GPUs in ~21 hours and train using Huggingface (Wolf et al., 2020). Training SOCIALITE-LLAMA on Lambda[3] would cost $33.60, making it very cost-effective.

4 Results

We analyze the effect of instruction tuning LLAMA2 on 20 seen and 6 related social scientific tasks and report macro F1 scores in Table 1.
Instruction tuning improves social understanding. We show that SOCIALITE-LLAMA outper-

---

[3]https://lambdalabs.com

forms LLAMA2 on all seen tasks in both zero- and few-shot setting. In fact, SOCIALITE-LLAMA used in a zero-shot setting is even better than LLAMA2 in few-shot setting. Surprisingly, we find that the benefit of few-shot learning over zero-shot is much less on seen tasks as opposed to the related tasks with the instruction tuned model.

Particularly, performance of our model is better for all tasks that are proximate to personal factors or how one feels, such as the affective dimensions, sentiment, emotion and empathy/distress. We also see consistent improvements over LLAMA2 on offensiveness related tasks. Despite the tremendous

TASK	SOCIALITE-LLAMA	DEBERTA (multi-task FT)
SENTIMENT	**73.43**	69.00
EMOTION	**81.08**	80.00
DISTRESSSELFRATED	**70.96**	65.00
EMPATHYSELFRATED	**63.97**	59.00
HYPERBOLE	**69.23**	69.00
SAMESIDESTANCE	**88.19**	76.00
HUMOR	**95.29**	91.00
POLITENESSHAYATI	84.54	89.00
INTIMACY	33.21	46.00
SUBJECTIVEBIAS	87.13	96.00
OFFENSIVE	**83.11**	83.00
INTENTTOOFFEND	**75.65**	74.00
SEXIST	73.85	79.00
BIASEDIMPLICATION	85.72	87.00

Table 2: We compare SOCIALITE-LLAMA with a multi-task finetuned (FT) DEBERTA reported in Choi et al. (2023). SOCIALITE-LLAMA matches or outperforms it on 10 out of 14 tasks that both models have been trained on, even though our training data size is much lower. **Bold** macro F1 represents improvement or equivalent performance of SOCIALITE-LLAMA with DEBERTA (statistically significant results performed using bootstrapped resampling; $p < .05$). Because DEBERTA is not generative, it is not possible to apply to tasks for which it was not finetuned.

general-purpose capabilities of LLAMA2, it appears to lack social understanding from text, and can be improved using instruction tuning on social scientific tasks.

Instruction tuning helps generalization. In related social science tasks, which are composed of new instructions and language samples, we find that SOCIALITE-LLAMA is better on 4 out of 6 under few-shot setting and 5 out of 6 overall. These 6 related tasks were on a varying degree of relatedness to the ones in the seen tasks. SOCIALITE-LLAMA shows significant improvement in performance over LLAMA2 on the related social science tasks. Through instruction tuning, our model generalizes social aspects of language to perform well on new related social tasks and task categories.

To assess the generalization emerging from the commonalities present between the task, we applied the instructions from the seen tasks' set on a related task and compared its performance with the related tasks' instruction (which was unseen to SOCIALITE-LLAMA). A non-author NLP expert suggested 4 seen tasks (OFFENSIVE, SEXIST, INTENTTOOFFEND and BIASEDIMPLICATION) that they thought were very similar to the HATESPEECH task. Of those, instructions from OFFENSIVE and

INTENTTOOFFEND gave improvement over zero-shot HATESPEECH performance, while the other two led to worse performance. This suggests it's possible to do better on zero-shot by using a related task prompt from instruction tuning but it can also hurt performance.

SOCIALITE-LLAMA is state-of-the-art overall on SOCKET. We compare SOCIALITE-LLAMA with a state-of-the-art multi-task finetuned DE-BERTA model presented in Choi et al. (2023) for a subset of the seen tasks since their DEBERTA model can't be applied on the related social tasks. Despite being trained on significantly less data than DEBERTA model (2.1 million data points and 58 tasks as compared to 108k data points and 20 tasks), Table 2 shows that our model is better or equivalent in performance on 10 out 14 tasks. This result highlights a large reduction in the performance separation between LLMs and smaller task finetuned language models in Social Science tasks established by previous works (Ziems et al., 2023; Choi et al., 2023).

5 Conclusion

We introduce SOCIALITE-LLAMA, an instruction tuned LLAMA2 7B for social science applications. SOCIALITE-LLAMA is trained using SOCIALITE-INSTRUCTIONS, a diverse collection of 20 social scientific tasks and tested on those plus an additional 6 related social tasks. It consistently performs better than LLAMA2 on all the seen tasks and all but 1 related task, demonstrating its strong generalization abilities. Instruction tuning on diverse social scientific tasks helps our model achieve state-of-the-art overall. Our results provide valuable insights to improve the social understanding of LLMs.

Limitations

The improvements brought through instruction tuning of LLAMA2 on social scientific tasks highlights the prevailing neglect of these tasks and the importance of doing this. However, there are some important limitations. For one, the tests for generalization to unseen tasks were limited to those that are semantically related to seen tasks (many tasks in SOCKET are related to each other). Future work would need to evaluate this model and process for generalizing to more distant tasks as well as across other points in time.

This work was informed by a number of works

in the NLP and deep learning literature, was largely facilitated by various open sourced datasets (Choi et al., 2023) and code (Wolf et al., 2020). However, the number of tasks in our work is not close to the number of tasks found in the prior instruction tuning literature (Wang et al., 2022; Sanh et al., 2022). This work would likely benefit more tasks within the realm of social science NLP.

Owing to limited access to computational resources, we could only focus on classification tasks on the smallest available LLAMA2 model and Low-Rank Adaptation as training strategy. However, literature on scaling models (Brown et al., 2020; Wei et al., 2022b) suggests that these performances likely go up with increasing the model size, datasets, instructions, which would make it more capable for regression and generation tasks with increased input lengths. Inclusion of few-shot examples in the training set could also improve the few-shot generalization capability of the model, which could not be performed with the resources available.

Ethics Statement

Large language models (LLMs) like GPT-4 (OpenAI, 2023) have been deployed to tens of millions of consumers in different forms (Heaven, 2023; Hu, 2023). While a number of works have already shown its limited social understanding (Ziems et al., 2023; Choi et al., 2023; Havaldar et al., 2023; V Ganesan et al., 2023), their rapid adoption and proliferation necessitates improving their social sensibility to make it reliable and useful. This work is aimed towards building a socially sensible language model for the practitioners of social science NLP research.

We hope our work improves NLP's benefit for social scientific pursuits and caution against the use of such models for non-scientific pursuits like targeting of individuals without their consent or awareness. This work has been built for research and can be re-distributed as per the policy laid out by Meta for LLAMA2[4]. We strongly believe that these models have to be tested for failure modes and harmful biases, and should be further be adjusted through appropriate methods (Ouyang et al., 2022) before deploying it into the hands of users.

---

[4] https://ai.meta.com/llama/license/

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, 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.

Herbert H. Clark and Michael F. Schober. 1992. Asking questions and influencing answers.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, Online. Association for Computational Linguistics.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. *Cognitive Science*, 43(1):e12701.

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. Instruction tuned models are quick learners.

Michael AK Halliday. 2004. Introduction: How big is a language? on the power of language. *The language of science*, 5:19–32.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Heaven. 2023. Chatgpt is everywhere. here's where it came from. *MIT Technology Review*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Krystal Hu. 2023. Chatgpt sets record for fastest-growing user base - analyst note. *Reuters*.

Erik Körner, Gregor Wiedemann, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. 2021. On classifying whether two texts are on the same side of an argument. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10130–10138, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multilevel dual-context language and bert. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 39–44.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H Andrew Schwartz. 2021. Melt: Message-level transformer with masked document representations as pre-training for stance detection. *arXiv preprint arXiv:2109.08113*.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *arXiv preprint arXiv:1911.09709*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Xianzhi Ruan, Steven Wilson, and Rada Mihalcea. 2016. Finding optimists and pessimists on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 320–325, Berlin, Germany. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 5477–5490, Online. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Schwartz. 2023. Systematic evaluation of GPT-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Vasudha Varadarajan, Nikita Soni, Weixi Wang, Christian Luhmann, H. Andrew Schwartz, and Naoya Inoue. 2022. Detecting dissonant stance in social media: The role of topic exposure. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 151–156, Abu Dhabi, UAE. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Ludwig Wittgenstein. 1953. *Philosophical investigations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yunxiang Zhang and Xiaojun Wan. 2022. MOVER: Mask, over-generate and rank for hyperbole generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6018–6030, Seattle, United States. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large

Language Models Transform Computational Social Science? *Computational Linguistics*, pages 1–53.

A Appendix

We first provide statistics about SOCIALITEIN-STRUCTIONS which has been used to instruction tune SOCIALITE-LLAMA. Table 3 includes information about amount of data used per task as well as the number of labels used.

Next, we provide descriptions of the instructions and prompt templates used for each task in SOCIALITEINSTRUCTIONS. Please note that this may contain examples of potentially dangerous and harmful text.

Dataset	Train Set	Validation Set	Test Set	Num_classes
EMOTION	3257	374	1421	4
FLUTE	6780	754	1498	4
EMPATHYEXPLORATIONS	2220	247	617	3
HUMOR	8000	1000	1000	2
OFFENSIVE	8000	4666	4691	2
SEXIST	7999	4666	4691	2
INTENTTOOFFEND	7999	4666	4691	2
BIASEDIMPLICATION	7999	4666	4691	2
POLITENESSHAYATI	256	32	32	2
HYPERBOLE	2580	323	323	2
SAMESIDESTANCE	140	18	17	2
SENTIMENT	8000	2000	12284	3
INTIMACY	1797	225	225	6
SUBJECTIVEBIAS	8000	9379	9379	2
VALENCECLS	9002	510	550	2
AROUSALCLS	9002	510	550	2
DOMINANCECLS	9002	510	550	2
EMPATHYSELFRATED	1487	186	186	2
DISTRESSSELFRATED	1487	186	186	2
HUMOURRATING	4932	632	615	2
HATESPEECH	-	-	2970	2
IRONY	-	-	784	2
POLITENESSSTANFORD	-	-	567	2
OPTIMISM	-	-	1495	3
COMPLAINTS	-	-	345	2
AGREEDISAGREE	-	-	4760	3

Table 3: Training, validation and test set statistics SOCIALITEINSTRUCTIONS. '-' denotes that the dataset was not used in training SOCIALITE-LLAMA but we create instructions for its test set. Overall, SOCIALITEINSTRUCTIONS contains ~202k data points.

Dataset	Instruction Format Example
HUMOR	Instruction: Upon receiving a piece of text, your task is to analyze and determine whether the intention of the text was to be humorous. You are instructed to look at the text and identify the structure of the joke, e.g. setup and punchline, or the content of the joke, e.g. absurdity, in order to determine if the intention of the text was to be humorous. If you think the intention of the text was to be humorous, classify it as 'humorous', else classfy it as 'non-humorous'. Input: TENNESSEE: We're the best state. Nobody even comes close. *Elevennessee walks into the room* TENNESSEE: Oh shit... Output: humorous
HUMOURRATING	Instruction: Upon receiving a piece of text, your task is to assess its comedic quality and categorize it as either 'low humor' or 'high humor'. Input: How many men does it take to open a can of beer? None. It should be open by the time she brings it to the couch. Output: low humor

Table 4: Instruction prompts and output examples for humour task

Dataset	Instruction Format Example
SENTIMENT	Instruction: Evaluate the sentiment conveyed in the input text and determine whether it is positive, negative, or neutral. This sentiment assessment should encompass the overall sentiment of the event described within the context of the topic mentioned in the text. Your options for classification are confined to positive, negative or neutral. Input: Few more hours to iPhone 6s launch and im still using the 4th generation Output: positive
EMOTION	Instruction: Analyze the following sentence and determine the predominant emotion it displays. Your options for classification are confined to anger, joy, optimism, or sadness. Please select one emotion from the given alternatives that you believe best epitomizes the emotional context of the sentence. Input: Worry is a down payment on a problem you may never have'. Joyce Meyer. #motivation #leadership #worry Output: Optimism
VALENCECLS	Instruction: Analyze the provided text using the Valence-Arousal-Dominance model for emotional assessment. Your task is to classify the valence level it would likely elicit in an average reader, where 'Low Valence' indicates a low level of pleasant feelings and 'High Valence' indicates a high level of pleasant feelings. Remember, the valence scale is used to measure the degree of pleasure or displeasure a person may feel towards something. Your options for classification are confined to 'Low Valence' or 'High Valence'. Input: Remember what she said in my last letter? Output: High Valence
AROUSALCLS	Instruction: Analyze the provided text using the Valence-Arousal-Dominance (VAD) emotional model. Your task is to classify the arousal level it might trigger in an average reader. Arousal, in this context, refers to the degree of energy or lethargy the text might induce. 'Low Arousal' indicates a low arousal level, suggesting the text is likely to make the reader feel calm or lethargic. Conversely, 'High Arousal' indicates a high arousal level, suggesting the tweet is likely to energize or excite the reader. Your options for classification are confined to 'Low Arousal' or 'High Arousal'. Input: Remember what she said in my last letter? Output: High Arousal
DOMINANCECLS	Instruction: Please analyze the provided text using the Valence-Arousal-Dominance (VAD) model for emotional response. Specifically, we're interested in the Dominance aspect of this model. This involves assessing the level of control or dominance the text might make an average reader feel, versus feelings of being controlled or submissive.\nPlease classify this dominance level as 'Low Dominance' or 'High dominance'. 'Low Dominance' indicates that the text is likely to evoke a low level of dominance or control in the reader, making them feel more submissive or controlled. Conversely, 'High Dominance' suggests that the text would make the reader feel highly dominant or in control.\nYour options for classification are confined to 'Low Dominance' or 'High Dominance'. Input: Remember what she said in my last letter? Output: High Dominance
SAMESIDESTANCE	Instruction: You are provided with two pieces of text sourced from an online debate forum. Your task is to analyze and categorize these texts based on their argumentative stance. Determine whether both texts are arguing in favor of the same viewpoint or if they are opposing each other. Your options for classification are confined to 'same side' or 'not same side'. Input: Legalizing gay marriage will not destroy man/woman relationships. [SEP] Why should gay marriage destroy man/woman relationships? It would just give gays the option to marry. Output: same side

Table 5: Instruction prompts and output examples for sentiment and emotion task

Dataset	Instruction Format Example
SUBJECTIVEBIAS	Instruction: Given two pieces of text, your objective is to detect subjective bias, which manifests when language that should remain neutral and impartial is influenced by feelings, opinions, or personal preferences, whether intentionally or unintentionally. If you find bias in the first sentence, indicate 'first sentence' as the output; otherwise, specify 'second sentence'. Input: the term finds widespread usage among members of the educational establishment who see students as tools of social change. [SEP] the term finds widespread usage among members of the educational establishment who see students as agents of social change. Output: first sentence
HYPERBOLE	Instruction: Upon receiving a piece of text, your task is to analyze and determine whether it contains hyperbolic language, which is an exaggerated statement or claim not meant to be taken literally, or if it does not. Your options for classification are confined to 'hyperbole' or 'not hyperbole'. Input: He looks great and yet he must be pushing sixty by now. Output: not hyperbole

Table 6: Instruction prompts and output examples for trustworthiness task

Dataset	Instruction Format Example
EMPATHYEXPLORATIONS	Instruction: Evaluate the degree of inquiry exhibited in the counselor's response provided below, categorizing it as either "Strong Exploration", "Weak Exploration" or "No Exploration". We define 'exploration' as instances where a mental health counselor displays keen interest in a patient by asking about experiences that haven't been explicitly mentioned. Input: Patient: I'm finally on medication and found a therapist who makes me feel hopeful for the first time in years. It has been a while since I've felt sincere hope. Counselor's response: im envious of the fact that you want hope. Output: No Exploration
EMPATHYSELFRATED	Instruction: Please carefully peruse the subsequent text, which is a personal account penned by an individual expressing their emotions and reflections after reading a news article. This account is directed towards their friends. After reading, your task is to accurately classify the level of empathetic concern demonstrated by the author. Your options for classification are 'low empathy' which indicates low empathetic concern or 'high empathy' which signifies a high degree of empathetic concern. Input: This sounds like a horrible accident. I can't even imagine what that family is going through. The ones that were involved in the accident but survived probably saw some horrible things, and those kids who witnessed their mother die. As a parent myself that's a terrifying thought I would never want my son to go through something like that. Output: high empathy
DISTRESSSELFRATED	Instruction: Please carefully peruse the subsequent text, which is a personal account written by an individual to their friends. This account details their emotional reactions and cognitive responses upon reading a specific news article. Your task is to accurately classify the level of personal distress experienced by the author. Your options for classification are 'low distress' or 'high distress'. Input: This sounds like a horrible accident. I can't even imagine what that family is going through. The ones that were involved in the accident but survived probably saw some horrible things, and those kids who witnessed their mother die. As a parent myself that's a terrifying thought I would never want my son to go through something like that. Output: low distress
FLUTE	Instruction: Please follow these steps: 1. First, you'll be presented with a premise and a hypothesis in the input section. 2. Your task is to determine and categorize the type of figurative language utilized in the hypothesis. 3. Finally, based on your assessment, respond with a single answer that most accurately represents the figurative language detected in the hypothesis. Choose from these four classifications: Idiom, Metaphor, Sarcasm, or Simile. Input: Premise: he was flying high happy and joy. Hypothesis: he was feeling down hurt and sad. Output: Metaphor

Table 7: Instruction prompts and output examples for other social factors

Dataset	Instruction Format Example
POLITENESSHAYATI	Instruction: Upon receiving a piece of text, your task is to analyze and determine whether the language used within it is courteous and respectful, indicating politeness, or if it contains disrespectful or rude elements, indicating impoliteness. Your options for classification are confined to 'impolite' or 'polite'. Input: @selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!! Output: impolite
INTIMACY	Instruction: Assess the degree of intimacy expressed in the input text, taking into account the social context within the text. Your options for classification are confined to 'very intimate', 'intimate', 'somewhat intimate', 'not very intimate', 'not intimate' or 'not intimate at all'. Input: What are the most mediocre animals in the animal kingdom? Output: not intimate

Table 8: Instruction prompts and output examples for other social factors

Dataset	Instruction Format Example
OFFENSIVE	Instruction: Evaluate the given text for any offensive content, which includes rudeness, disrespect, or toxicity. This assessment should consider if the text could potentially offend anyone, based on previous studies indicating a higher recall rate. Identify any disrespectful, inappropriate, or harmful language, phrases, or sentiments. If these elements exist, label the text as 'offensive'. If these elements are absent, mark the text as 'not offensive'. Input: jack and jill went up a hill . <repeat> jill came down , i went up , and blew jack . ooohh ! <allcaps> it do n't rhyme . it just happened that way. Output: not offensive
SEXIST	Instruction: Analyze the provided sentence and evaluate if it contains any elements that could be considered as gender-based discrimination. You are required to categorize the sentence into one of two classifications: 'sexism' if it exhibits gender discrimination, or 'not sexism' if it does not. Input: "Alternative Theory: They had to change the timeline to explain why the building engineer was up there–had he heard shots, he would not have gone to 32nd floor– in order to have #StephenSchuck corroborate #JesusCampos story, it had to have been before shooting started. Output: not sexism
INTENTTOOFFEND	Instruction: Your task is to analyze the given text and identify if it intentionally promotes social biases and stereotypes. If you perceive that the author's motive was to deliberately offend, categorize the text as 'intentional'. If there's no apparent intent to offend, classify it as 'not intentional'. Remember, the focus is on the author's intent to offend, not just the offensiveness of the content. Input: should i start discourse with eileen and hope it makes them stop telling my friends im a pedophile cus i'm fucking not? Output: not intentional
BIASEDIMPLICATION	Instruction: Your task is to scrutinize a given text and identify any signs of prejudice. You need to differentiate between statements that merely insult an individual and those that imply group discrimination, such as comparing 'F*ck you' (an individual insult) with 'F*ck you, f*ggot' (a group implication). Based on your analysis, categorize the text as either 'biased' or 'not biased'. Input: Obviously everyone you disagree with on twitter is a manipulation bot & everyone you agree with is real, right? :p Output: biased

Table 9: Instruction prompts and output examples for offensive task

Dataset	Instruction Format Example
HATESPEECH	Instruction: Analyze the given text for hateful contents against women or immigrants. Your options for classification are confined to 'hate speech' or 'not hate speech'. Input: Send home migrants not in need of protection, Peter Dutton tells UN, HEY DUTTON HOW ABOUT THE ONES THAT HAVE STAYED AND NOT LEFT THE COUNTRY WHEN THEY SHOULD OVERSTAYERS ? WHY DONT YOU GO AND ROUND ALL THEM UP ? Output: hate speech
IRONY	Instruction: Classify the input sentence as ironic or not ironic. Here ironic can be defined as the use of words or situations to convey a meaning that is opposite to what is expected. Your options for classification are confined to 'ironic' or 'not ironic'. Input: Now I remember why I buy books online @user #servicewithasmile Output: ironic
POLITENESSSTANFORD	Instruction: Analyze the provided text, considering its tone and language, and categorize it as either polite or impolite. Your options for classification are confined to 'impolite' or 'polite'. Input: Hi, I made on comment on the image page about this being different than Sbarro3 image. Please... comment? Output: polite
OPTIMISM	Instruction: Analyze the sentiment of the provided text and classify it as 'optimistic', 'pessimistic' or 'neutral'. Input: fuck you fuck you fuck you Output: pessimistic
COMPLAINTS	Instruction: Given an input text, identify if it contains a complaint or not. Complaining is a basic speech act used to express a negative mismatch between reality and expectations towards a state of affairs, product, organization or event. Your options for classification are confined to 'complaint' or 'not complaint'. Input: @SamsungSupport Can someone please help me? I've already sent a DM. Output: complaint
AGREEDISAGREE	Instruction: You are provided with two pieces of text and your task is to analyze and categorize these texts based on their argumentative stance. Determine whether both texts are arguing in favor of the same viewpoint, if they are opposing each other or if they are talking about two different topics. Your options for classification are confined to 'agree', 'disagree' or 'N/A'. Input: President Trump also intentionally lied to hide the truth from the American people and make himself look better. [SEP] President Trump told public lies every single day for his first 40 days. Output: agree

Table 10: Instruction prompts and output examples for related social tasks tasks