Inducing Human-like Biases in Moral Reasoning Language Models

Artem Karpov * Independent
Seong Hah Cho * Independent
Austin Meek * University of Delaware
Raymond Koopmanschap Independent
Lucy Farnik University of Bristol
Bogdan-Ionut Cirstea † Independent

Abstract

In this work, we study the alignment (BrainScore) of large language models (LLMs) fine-tuned for moral reasoning on behavioral data and/or brain data of humans performing the same task. We also explore if fine-tuning several LLMs on the fMRI data of humans performing moral reasoning can improve the BrainScore. We fine-tune several LLMs (BERT, RoBERTa, DeBERTa) on moral reasoning behavioral data from the ETHICS benchmark [Hendrycks et al., 2020], on the moral reasoning fMRI data from Koster-Hale et al. [2013], or on both. We study both the accuracy on the ETHICS benchmark and the BrainScores between model activations and fMRI data. While larger models generally performed better on both metrics, BrainScores did not significantly improve after fine-tuning.

1 Introduction

Recently, multiple papers have shown surprising similarities between the internal representations in biological brains and those in artificial neural networks, in multiple domains and for multiple tasks; see e.g. Huh et al. [2024] for a review and for some potential theoretical explanations of why one might expect this, including with increasingly powerful machine learning models. However, to the best of our knowledge, no previous work has analyzed potential analogous representational similarities in the domain of moral reasoning, nor whether the degree of similarity might be increased by using human neural data (e.g. fMRI).

In this work, we undertake the first such measurement (BrainScore) of the similarity of the internal representations of biological brains (measured through fMRI) and large language models (LLMs), in the domain of moral reasoning (in a task which is also partially relevant to Theory of Mind). We also study whether fine-tuning the LLMs on a train set of the corresponding neural data helps with improving the BrainScore on a separate test set. While our attempts here proved unsuccessful, we think this is an important problem to study and that increasing said alignment might be important for the AI alignment problem [Christian, 2020].

We next discuss related works in section 2, our methodology in section 3, the results we obtained in section 4, and finally in section 5 conclude and discuss some potential future work.

*Equal Contribution. Authors listed alphabetically, and contributions listed in A.
†Corresponding author: cirstea.bogdanionut@gmail.com

Preprint.

2 Related Works

There is a growing interest in brain-model alignment work, here we only provide a brief overview. For a more detailed survey, see Sucholutsky et al. [2023], especially section 4.3.2, and Schrimpf et al. [2020] for a systematic approach to collecting and scoring many models. Earlier work on fine-tuning transformers to predict fMRI data has found that adding MEG data also helps [Schwartz et al., 2019]. Other work [Aw and Toneva, 2023] focused on fine-tuning models on the much larger Booksum dataset [Kryściński et al., 2022], which they found increased alignment. Dapello et al. [2022] used rhesus macaque neural data and showed improved alignment with human neural data and greater adversarial robustness.

To the best of our knowledge, we are the first to attempt increasing brain-model alignment on moral reasoning neuroimaging data.

3 Methodology

3.1 Benchmark and Dataset

To quantitatively measure the moral reasoning performance of the fine-tuned LLMs, we used the common sense category of the ETHICS benchmark [Hendrycks et al., 2020], which consists of multiple choice questions rather than free form responses. To predict an answer, we use a linear transformation layer (a CLS head or just a head) attached to predictions (logits) for a classification token, "[CLS]", of a base model.

We used the fMRI dataset, 'Moral judgments of intentional and accidental moral violations across Harm and Purity domains', from Koster-Hale et al. [2013]. Human subjects were given a series of scenarios describing moral, immoral, and neutral actions across a wide variety of scenarios, and then answered on a 1-4 scale how moral or immoral each action was. Koster-Hale et al. [2013] was approved by an IRB and subjects were paid and gave written, informed consent.

3.2 Data Pre-processing and Analysis

We used a pre-processed version of the dataset from Thomas et al. [2023], specifically the version fit with the DiFuMo atlas [Dadi et al., 2020] with a dimensionality of 1,024 regions of interest (the maximum number of dimensions DiFuMo provides).

Because of the high granularity of our chosen atlas, we used NeuroSynth [Yarkoni et al., 2011], a tool for meta-analysis conducted over thousands of fMRI studies to isolate regions consistently activated during experiments, to map activations to specific themes. We conducted our analyses on four regions related to Theory of Mind, moral reasoning, language, and vision. We used vision as the control group as we expected scores there not to increase (see Appendix A). We visualized the relationship between the fMRI and LLM activations on the cortical surface (see Appendix A) using the Coefficient of Determination (CoD). The CoD scores were then negative log transformed and the weighted average of the parcel scores were plotted at each vertex, since the DiFuMo atlas is probabilistic with overlapping boundaries.

3.3 Models and Fine-tuning Procedure

We focused on encoder models (BERT-based) due to computational constraints and since this was a classification task. Additionally, encoder models originally showed better results on the ETHICS dataset [Hendrycks et al., 2020]. Overall we used four models, BERT-base-cased and BERT-large cased (108 and 333 million parameters) [Devlin et al., 2019], RoBERTa-large (355 million parameters) [Liu et al., 2019], and DeBERTa-v2-xlarge (864 million parameters) [He et al., 2021].

For fine-tuning, we used the HuggingFace library [Wolf et al., 2019] to train additional heads of dimensionality 1,024 (to match the DiFuMo atlas) on top of the classification token, "[CLS]". We also train heads to predict the ETHICS benchmark [Hendrycks et al., 2020]. We report fine-tuning on ETHICS only and with the addition of fMRI data in Table 1. In total, we ran 450 fine-tuning runs, totaling 292 hours of training for 1,082 different models (not all shown in the results section).

Model	On Ethics only	Runs	CS Hard Set, % (95 CI)		CS Test Set, % (95 CI)	
		count	mean	max	mean	max
BERT-base		35	47.3 (0.0, 53.9)	55.5	57.0 (48.8, 70.5)	73.7
BERT-base	Yes	7	52.3 (50.0, 55.3)	55.4	58.3 (50.0, 71.0)	71.7
BERT-large		28	53.6 (49.4, 59.0)	**61.8**	62.0 (48.5, 78.7)	85.4
BERT-large	Yes	16	52.5 (48.2, 58.8)	58.8	59.2 (43.9, 78.8)	79.3
RoBERTa-large		4	66.1 (51.5, 72.4)	**72.5**	80.6 (53.0, 91.4)	**91.4**
RoBERTa-large	Yes	18	65.7 (49.8, 73.8)	**74.1**	79.0 (49.7, 91.6)	**91.8**
DeBERTa-v2-xlarge		9	51.8 (45.9, 67.4)	70.8	52.9 (49.9, 66.6)	70.0
DeBERTa-v2-xlarge	Yes	3	59.5 (49.8, 77.4)	78.8	64.1 (49.9, 90.2)	92.3

Table 1: Results of fine-tuning four different models on the Commonsense split of the ETHICS dataset [Hendrycks et al., 2020]. Bolded values are those higher than reported by the original authors. Values are for models fine-tuned on ETHICS only if stated or otherwise fine-tuned on both ETHICS and fMRI data. Parentheses indicate a two standard deviation confidence interval.

3.4 Brain Scores

Our 'brain-score' metric is based on similar metrics found within the broader NeuroAI literature, such as in Schrimpf et al. [2020], Aw and Toneva [2023], *inter alia*. We use the Pearson's correlation coefficient (PCC) to measure the correlation between predicted brain activity and actual brain activity. For some moral scenario given to a subject, we sample the fMRI data at several time points, taking the hemodynamic lag into account. This data has been fit to the DiFuMo atlas, resulting in 1,024 ROIs at the time points sampled. We do this over a collection of similar examples, and then fit a regression model to predict this brain activity. The PCC between the predicted response and the actual held out brain activity gives us the brain-score metric. We can also do this on a layer-by-layer basis and aggregrate over all layers to provide a single brain-score for the whole model, which we provide in Table 2.

4 Results

Our results indicate that improving brain-model alignment on moral reasoning by fine-tuning on relevant fMRI data does not consistently improve accuracy on ETHICS [Hendrycks et al., 2020]. While our fine-tuning procedures do improve accuracy on the Commonsense split of ETHICS (bolded values in Table 1 are higher than those reported by Hendrycks et al. [2020]), we could not improve accuracy by fine-tuning on the fMRI data only or on a combination of fMRI and ETHICS, compared to fine-tuning purely on ETHICS.

To thoroughly test these results, we also used a variety of sampling methods to pull from the fMRI data, as shown in Table 3 and Figure 2 in Appendix B. AVG indicates an average of all time points, LAST indicates the time point at the hemodynamic lag before the last time point, MIDDLE indicates the middle point, and SENTENCES indicates four points in a scenario, which match the end of the four sentences read by the subject. We find that LAST tends to produce the best accuracy.

We generally find that larger models are more performant overall, a finding also reported by Hendrycks et al. [2020]. This additionally holds with the brain-score metric, which measures brain-model alignment. However, we were also unable to significantly improve our brain-score metric beyond the pre-trained models, as shown in Table 2. This finding is also consistent in layer-wise scores across each of the three models; see Appendix A for further brain-score details (including region specific information, such as Theory of Mind ROIs) and cortical mappings.

5 Conclusion and Future Work

While we were unable to significantly increase brain alignment on moral reasoning through fine-tuning methods, we do believe that our results can be of use for downstream work. Firstly, we believe that our work is ample evidence for the importance of gathering more data on moral reasoning and for more niche tasks in general if future researchers want to increase brain-model alignment

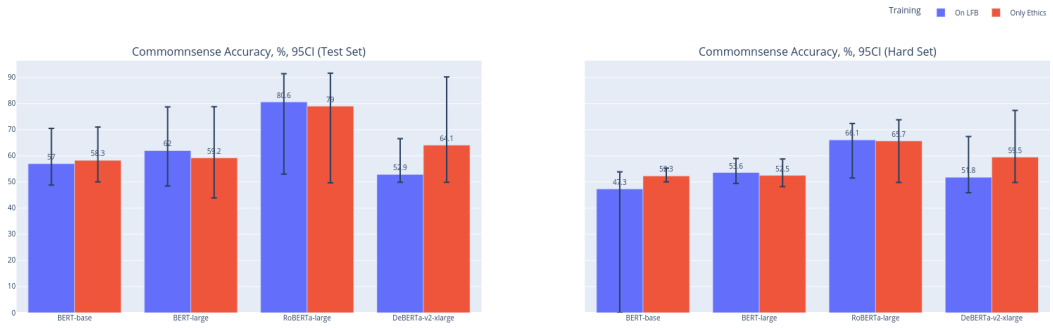

Figure 1: Accuracy values for the Commonsense split of the ETHICS dataset Hendrycks et al. [2020]. See Table 1 for a tabular depiction of the data.

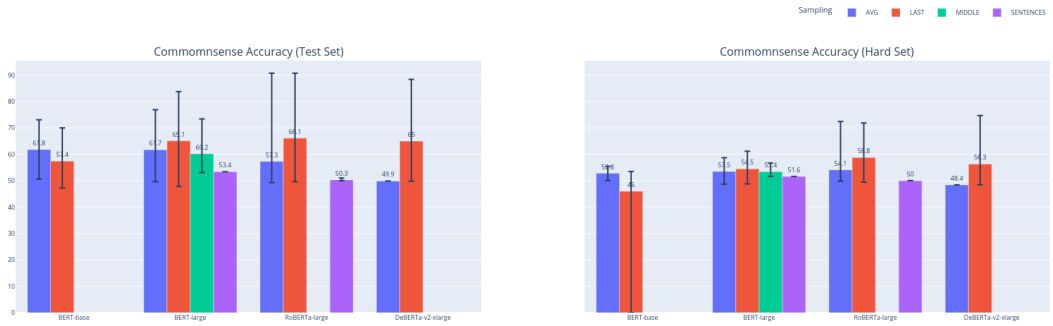

Figure 2: Graphical depiction of the different sampling methods' effect on accuracy on the Commonsense split of ETHICS Hendrycks et al. [2020].

Model	Sampling	Fine-tuning	Brain Score Mean	Brain Score St. Dev.
BERT-large-cased	LAST	No fine-tuning	0.217	0.096
BERT-large-cased	LAST	ETHICS and fMRI	0.213	0.095
RoBERTa-large	LAST	No fine-tuning	0.173	0.117
RoBERTa-large	LAST	ETHICS	0.145	0.097
RoBERTa-large	LAST	ETHICS and fMRI	0.156	0.112
RoBERTa-large	LAST	ETHICS then fMRI	0.144	0.113
DeBERTa-v2-xlarge	LAST	No fine-tuning	0.271	0.094
DeBERTa-v2-xlarge	LAST	ETHICS	0.266	0.095
DeBERTa-v2-xlarge	LAST	ETHICS and fMRI	0.273	0.096
DeBERTa-v2-xlarge	LAST	ETHICS then fMRI	0.264	0.097
DeBERTa-v2-xlarge	LAST	fMRI then ETHICS	0.237	0.097

Table 2: Brain scores across models and different fine-tuning methods. We were unable to significantly increase brain-model correlation using any of the fine-tuning methods.

in specific domains. Secondly, we make our code available.[3] We believe that further work along the neuroconnectionist research agenda [Doerig et al., 2023] will be useful generally, and hope that the preliminary evidence we provide here will help update others' research models on the ability to increase alignment in specific domains.

Acknowledgments and Disclosure of Funding

This work was done with the support from AI Safety Camp, especially Remmelt Ellen and Linda Linsefors. Thanks to Linda Linserfors, Paul Bricman, Koen Holtman and Jeremy Gillen for discussions and feedback which helped significantly improve the proposal draft and to Eleni Angelou for inspiration for the draft format. Bogdan was funded by the Center on Long-Term Risk. Earlier versions of the proposal benefited from Bogdan's previous [postdoc] appointment, funded by the Leverhulme Trust grant RPG-2019-243, and discussions with and feedback from collaborators Fabio Cuzzolin, Christelle Langley, Barbara Sahakian. Artem Karpov was funded by Good Ventures Foundation, the program, "Early-career funding for individuals interested in improving the long-term future" by Open Philanthropy. We thank the team of wandb.ai for giving us free tracking hours to log results from fine tuning, especially the support from Artsiom Skarakhod. Thanks to Hugo Berg for Google Cloud Platform (GCP) compute credits. Thanks to GCP for free access to TPUs. Thanks to people and organizations that generously published tools and libraries we used for free, especially PyTorch and Lightning. Additional thanks to Austin Brockmeier for helpful discussions throughout the project.

References

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment, February 2023. URL http://arxiv.org/abs/2212.10898. arXiv:2212.10898 [cs, q-bio].

Brian Christian. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020.

Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J. Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, 221:117126, November 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117126. URL https://www.sciencedirect.com/science/article/pii/S1053811920306121.

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D. Cox, and James J. DiCarlo. Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. preprint, Neuroscience, July 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.07.01.498495.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Adrien Doerig, Rowan P. Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W. Lindsay, Konrad P. Kording, Talia Konkle, Marcel A. J. van Gerven, Nikolaus Kriegeskorte, and Tim C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24 (7):431–450, July 2023. ISSN 1471-0048. doi: 10.1038/s41583-023-00705-w. URL https://www.nature.com/articles/s41583-023-00705-w. Publisher: Nature Publishing Group.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, October 2021. URL http://arxiv.org/abs/2006.03654. arXiv:2006.03654 [cs].

[3] Code available at: https://github.com/ajmeek/Inducing-human-like-biases-in-moral-reasoning-LLMs

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI With Shared Human Values, February 2020. URL http://arxiv.org/abs/2008.02275. arXiv:2008.02275 [cs].

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Jorie Koster-Hale, Rebecca Saxe, James Dungan, and Liane L. Young. Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110 (14):5648–5653, April 2013. doi: 10.1073/pnas.1207992110. URL https://www.pnas.org/doi/10.1073/pnas.1207992110. Publisher: Proceedings of the National Academy of Sciences.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. BookSum: A Collection of Datasets for Long-form Narrative Summarization, December 2022. URL http://arxiv.org/abs/2105.08209. arXiv:2105.08209 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, January 2020. URL https://www.biorxiv.org/content/10.1101/407007v2. Pages: 407007 Section: New Results.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models, October 2019. URL http://arxiv.org/abs/1911.03268. arXiv:1911.03268 [cs, q-bio].

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, November 2023. URL http://arxiv.org/abs/2310.13018. arXiv:2310.13018 [cs, q-bio].

Armin W. Thomas, Christopher Ré, and Russell A. Poldrack. Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data, January 2023. URL http://arxiv.org/abs/2206.11417. arXiv:2206.11417 [q-bio].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, 2019. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Tal Yarkoni, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8 (8):665–670, August 2011. ISSN 1548-7105. doi: 10.1038/nmeth.1635. URL https://www.nature.com/articles/nmeth.1635. Publisher: Nature Publishing Group.

A Author Contributions

Artem Karpov – implementation of the fine tuning, experiments for fine tuning, reports for the fine tuning experiments, most of the data processing, setting up infrastructure for experiments, code reviews, wrote some parts of this paper.

Seong Hah Cho - brain score, data analysis, conceptual work.

Austin Meek - some of the initial work on data processing, work on brain scores, some infra and experiments, code reviews, wrote parts of this paper.

Raymond Koopmanschap – brain score, most of the potential parameter-efficient fine-tuning extensions, fine-tuning.

Lucy Farnik – potential extensions for better transfer learning / dealing with catastrophic forgetting.

Bogdan Ionut Cirstea – most of the conceptual work, supervision, paper editing.

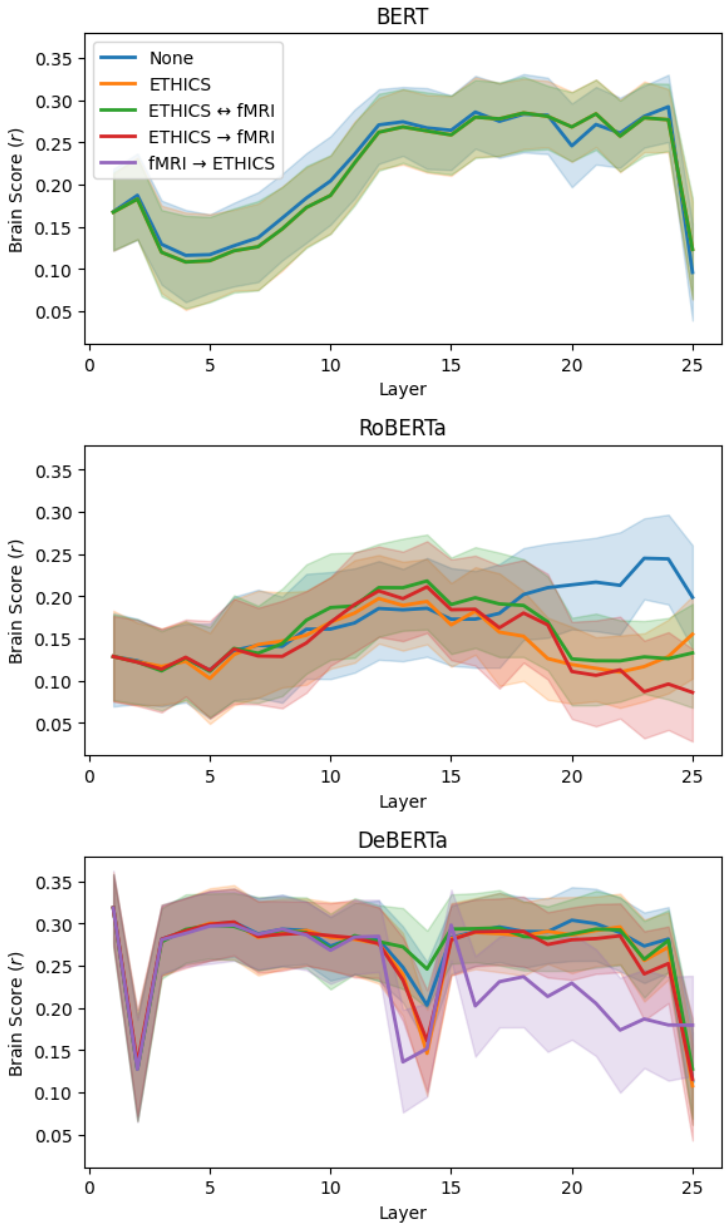

Figure 3: Brain scores across the hidden layers from bert-large-cased, roberta-large, and deberta-v2-xlarge across our different fine-tuning protocols.

B Brain Scores and Cortical Maps

In Figure 3, we plot further brain scores across different layers. Table 2 shares the brain score values averaged across the entire model. Below, in Figures 4 through 12, we plot the cortical maps of different fine-tuning methods on different models, as well as the cortical map of the activations provided by NeuroSynth [Yarkoni et al., 2011] for our four areas of interest: language, moral reasoning, theory of mind, and vision.

We used vision ROIs as a control group. Note that our models were not able to achieve better brain scores than the control group, meaniing that our experiment did not achieve the desired effect. Nevertheless, we believe that releasing the results of our experiments may help to inform future research about the necessity of larger datasets and more effective fine-tuning.

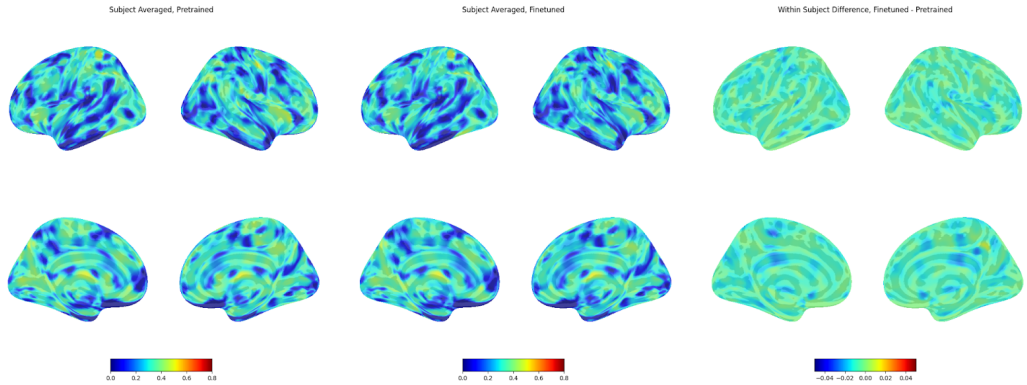

Figure 4: Subject and layer averaged CoD taken from bert-large-cased A) without fine-tuning on ETHICS or the fMRI recordings, B) with fine-tuning on ETHICS and fMRI recordings, and C) their difference.

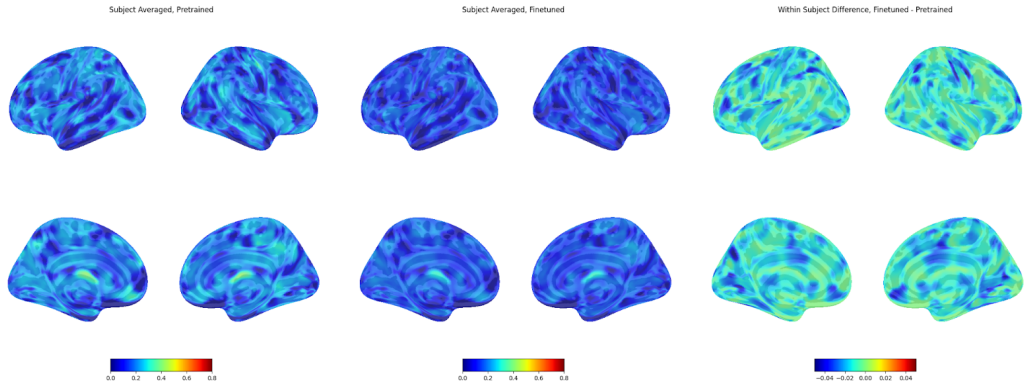

Figure 5: Subject and layer averaged CoD taken from roberta-large A) without fine-tuning on ETHICS or the fMRI recordings, B) with fine-tuning on ETHICS but not the fMRI recordings, and C) their difference.

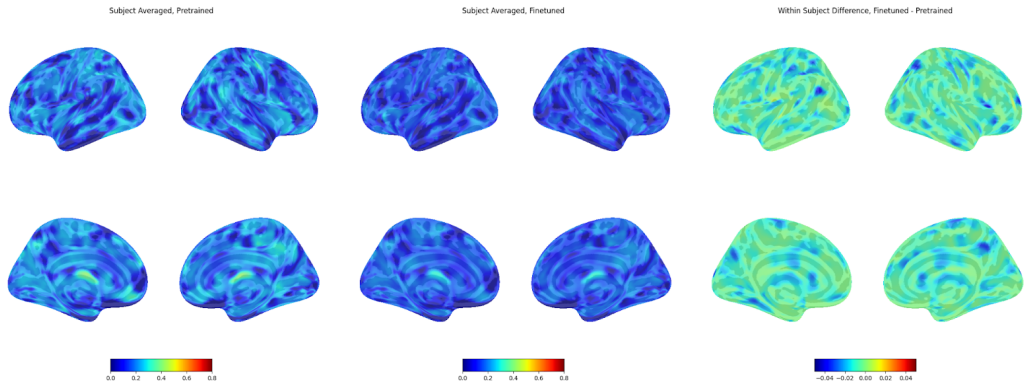

Figure 6: Subject and layer averaged CoD taken from roberta-large A) without fine-tuning on ETHICS and the fMRI recordings, B) with fine-tuning on both ETHICS and the fMRI recordings repeatedly, and C) their difference.

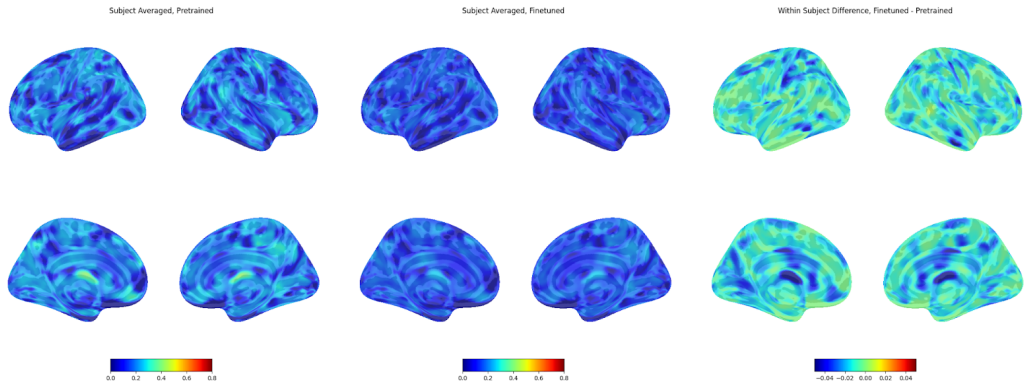

Figure 7: Subject and layer averaged CoD taken from roberta-large A) without fine-tuning on ETHICS and the fMRI recordings, B) with fine-tuning sequentially on ETHICS then on the fMRI recordings, and C) their difference.

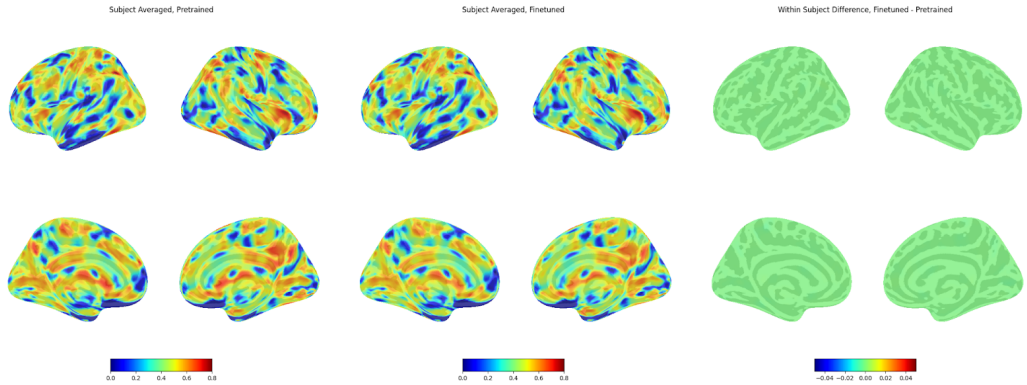

Figure 8: Subject and layer averaged CoD taken from deberta-v2-xlarge A) without fine-tuning on ETHICS and the fMRI recordings, B) with fine-tuning on ETHICS but not the fMRI recordings, and C) their difference.

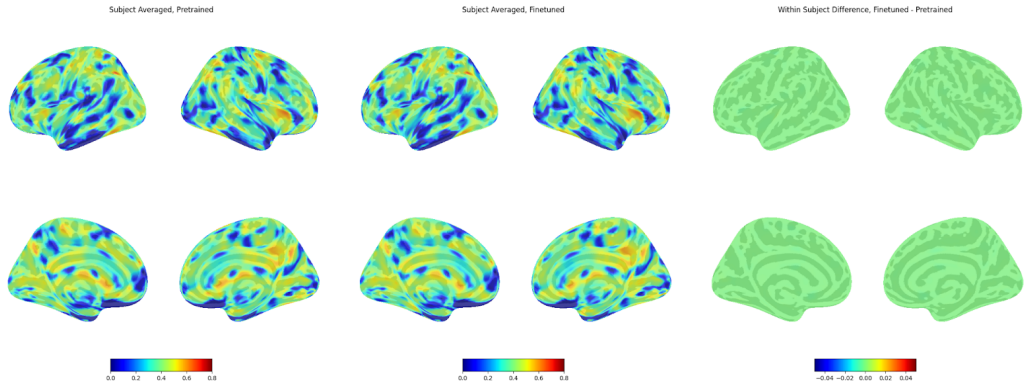

Figure 9: Subject and layer averaged CoD taken from deberta-v2-xlarge A) without fine-tuning on ETHICS and the fMRI recordings, B) with fine-tuning on both ETHICS and the fMRI recordings repeatedly, and C) their difference.

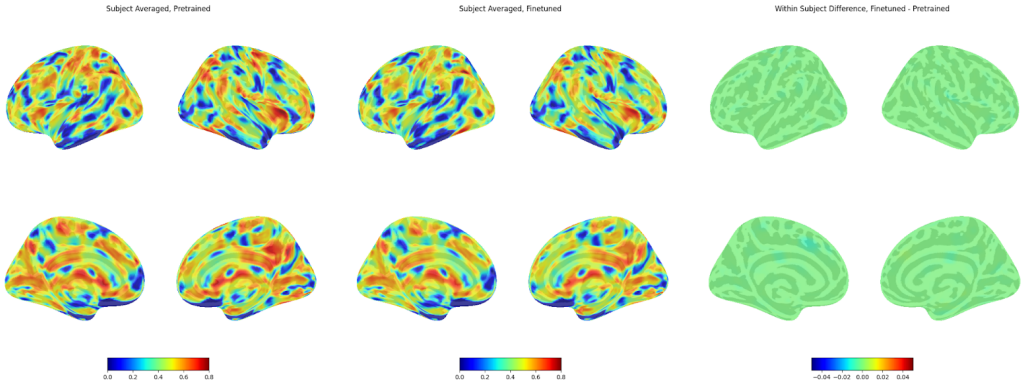

Figure 10: Subject and layer averaged CoD taken from deberta-v2-xlarge A) without fine-tuning on ETHICS and the fMRI recordings, B) with fine-tuning sequentially on ETHICS then on the fMRI recordings, and C) their difference.

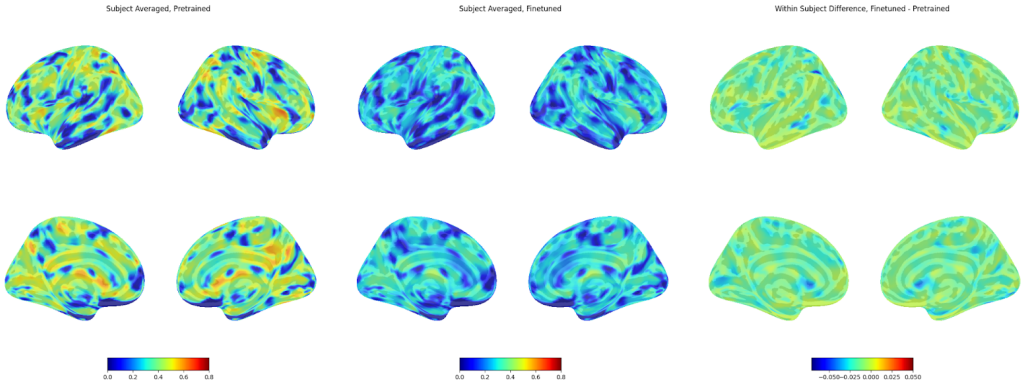

Figure 11: Subject and layer averaged CoD taken from deberta-v2-xlarge A) without fine-tuning on ETHICS and the fMRI recordings, B) with fine-tuning sequentially on fMRI recordings then on ETHICS, and C) their difference.

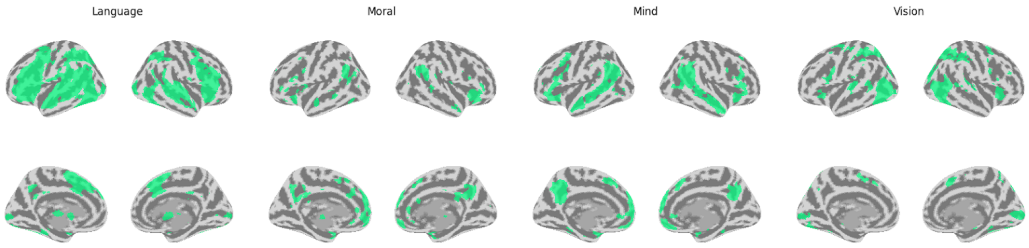

Figure 12: Functional activations taken from NeuroSynth meta-analyses for the terms language, moral, theory of mind, and vision, respectively.

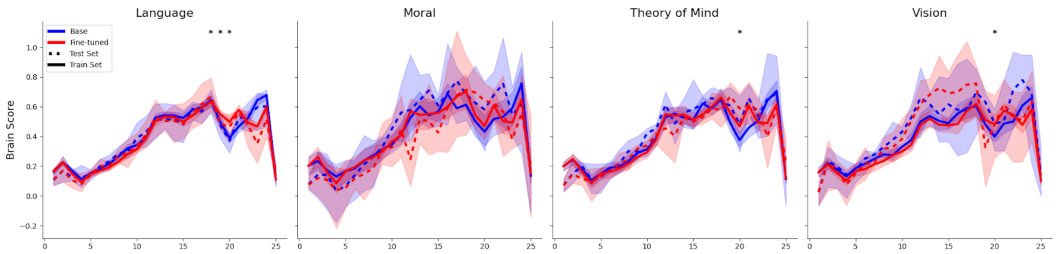

Figure 13: Scores of bert-large-cased across the term-based functional activations from NeuroSynth. The fine-tuning occurred over both ETHICS and the fMRI recordings, repeatedly. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

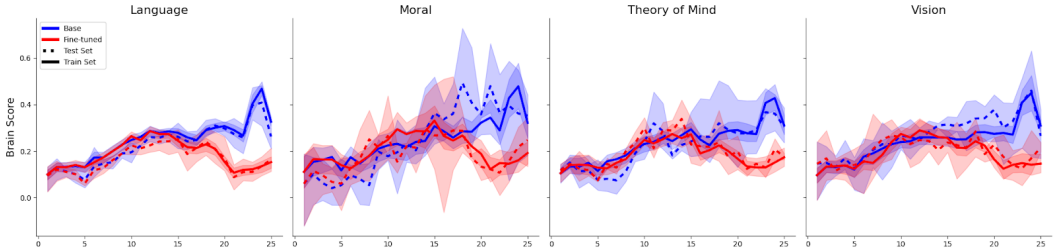

Figure 14: Scores of roberta-large across the term-based functional activations from NeuroSynth. The fine-tuning occurred over ETHICS but not the fMRI recordings. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

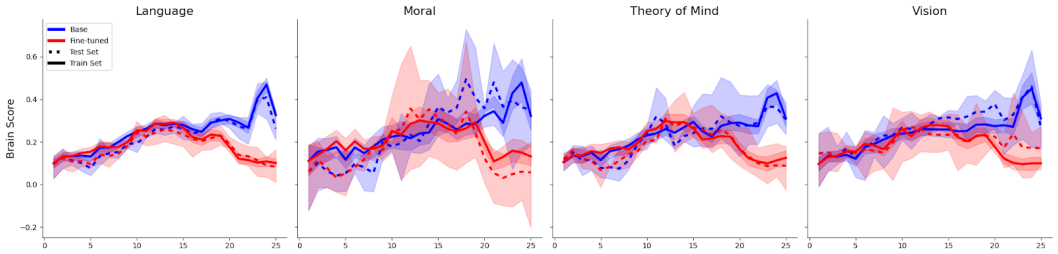

Figure 15: Scores of roberta-large across the term-based functional activations from NeuroSynth. The fine-tuning occurred over both ETHICS and the fMRI recordings, sequentially in that order. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

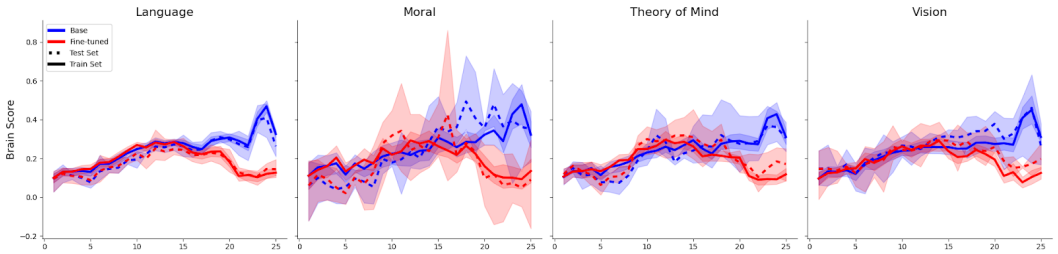

Figure 16: Scores of roberta-large across the term-based functional activations from NeuroSynth. The fine-tuning occurred over both ETHICS and the fMRI recordings, repeatedly. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

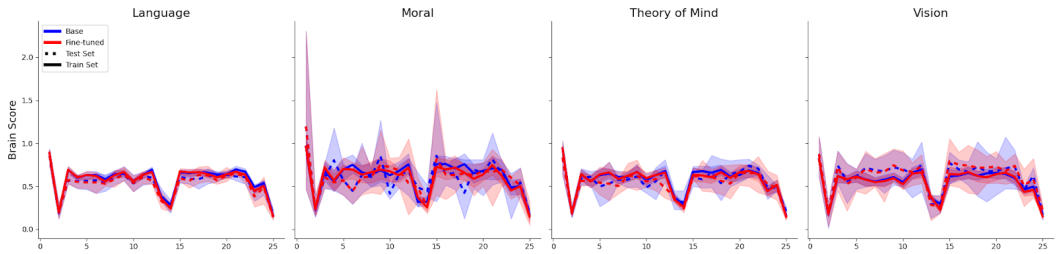

Figure 17: Scores of deberta-v2-xlarge across the term-based functional activations from NeuroSynth. The fine-tuning occurred over ETHICS but not the fMRI recordings. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

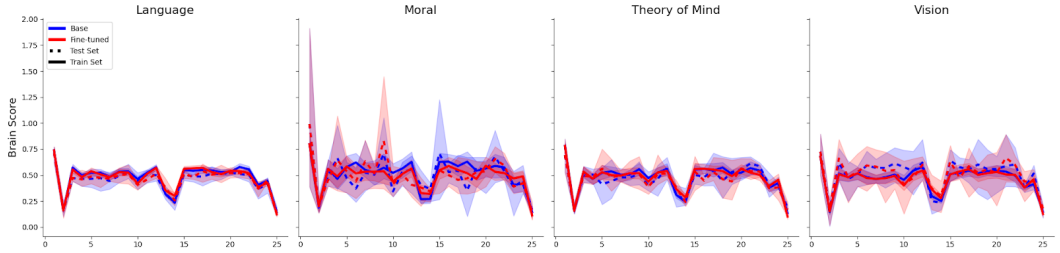

Figure 18: Scores of deberta-v2-xlarge across the term-based functional activations from NeuroSynth. The fine-tuning occurred over both ETHICS and the fMRI recordings, repeatedly. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

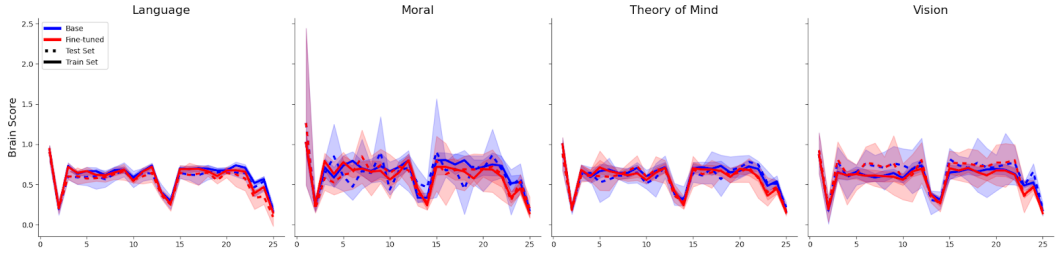

Figure 19: Scores of deberta-v2-xlarge across the term-based functional activations from NeuroSynth. The fine-tuning occurred over both ETHICS and the fMRI recordings, sequentially in that order. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

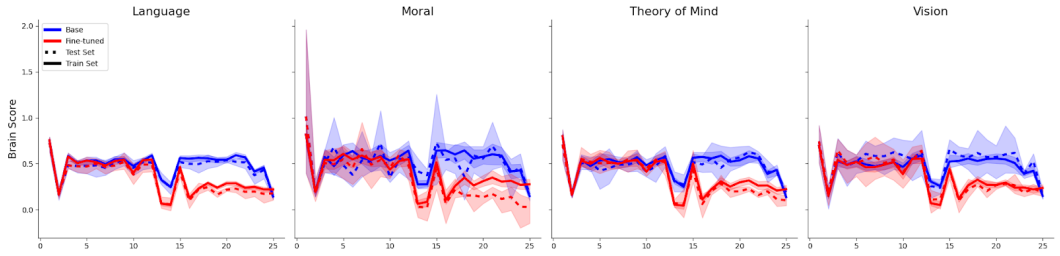

Figure 20: Scores of deberta-v2-xlarge across the term-based functional activations from NeuroSynth. The fine-tuning occurred over both fMRI recordings and ETHICS, sequentially in that order. Asterisks indicate one-tailed Bonferroni corrected significance between the training pre-trained and fine-tuned scores if the fine-tuned scores are greater than the base scores.

Model	Sampling	Runs	CS Hard Set, %		CS Test Set, %	
			mean ± STD	max	mean ± STD	max
BERT-base	AVG	2	**52.8 ± 3.9**	55.5	**61.8 ± 16.7**	73.7
BERT-base	LAST	28	46.0 ± 16.3	53.6	57.4 ± 7.2	70.0
BERT-large	AVG	16	53.5 ± 3.0	59.6	61.7 ± 10.1	77.7
BERT-large	LAST	7	**54.5 ± 4.8**	61.8	**65.1 ± 14.8**	85.4
BERT-large	MIDDLE	3	53.4 ± 3.1	57.0	60.2 ± 12.4	74.5
BERT-large	SENTENCES	1	51.6	51.6	53.4	53.4
RoBERTa-large	AVG	11	54.1 ± 9.0	72.5	57.3 ± 16.4	90.9
RoBERTa-large	LAST	7	**58.8 ± 11.2**	72.2	**66.1 ± 20.3**	91.4
RoBERTa-large	SENTENCES	3	50.0 ± 0.1	50.1	50.3 ± 0.6	51.0
DeBERTa-v2-xlarge	AVG	1	48.4	48.4	49.9	49.9
DeBERTa-v2-xlarge	LAST	4	**56.3 ± 13.6**	76.6	**65.0 ± 19.1**	89.9

Table 3: Comparison of different sampling methods for fine-tuning on the fMRI dataset. Bolded values are the best accuracies per model.

C Additional Experiment Details

In Table 3 we provide a tabular depiction of the effect that different sampling methods had on our fine-tuning experiments and accuracy on the Commonsense split of the ETHICS dataset [Hendrycks et al., 2020]. See Figure 2 for a graphical depiction.