Characterizing Human and Zero-Shot GPT-3.5 Object-Similarity Judgments

D. Estes McKnight
Dept. Computing Science
University of Alberta
Alberta Machine Intelligence Inst.
research@demcknight.com

Alona Fyshe
Depts. Comp. Sci. and Psychology
University of Alberta
Alberta Machine Intelligence Inst.
alona@ualberta.ca

Abstract

Recent advancements in large language models' (LLMs) capabilities have yielded few-shot, human-comparable performance on a range of tasks. At the same time, researchers expend significant effort and resources gathering human annotations. At some point, LLMs may be able to perform some simple annotation tasks, but studies of LLM annotation accuracy and behavior are sparse. In this paper, we characterize OpenAI's GPT-3.5's judgment on a behavioral task for implicit object categorization. We characterize the embedding spaces of models trained on human vs. GPT responses and give similarities and differences between them, finding many similar dimensions. We also find that despite these similar dimensions, augmenting humans' responses with GPT ones drives model divergence across the sizes of datasets tested.

1 Introduction

Large language models (LLMs) are capable of accomplishing a variety of language-oriented tasks in zero- or few-shot settings (Brown et al., 2020). Examples include common natural-language understanding and processing (NLU/P) tasks such as sentiment analysis and classification (Brown et al., 2020), language translation (Hendy et al., 2023), and named entity recognition (Ji, 2023); but also applied domains such as text tagging (Gilardi et al., 2023), multimodal tagging (Li et al., 2023), and text sample augmentation (Dai et al., 2023).

Current LLM performance indicates we may be able to use pre-trained high-resource LLMs to augment human annotations for tasks where data is sparse or compute resources are low (Møller et al., 2023). However, we do not currently know for which domains it is appropriate to augment human data with LLM-generated responses. This uncertainty stems from a poor understanding of how LLM and human annotation responses systematically differ. Thus, characterizing the ways in which world knowledge manifests itself in the generations of LLMs is crucial for incorporating LLMs into annotation workflows.[1]

The domain of object-similarity judgment is a useful base-case for exploring the similarities and substitutability of LLM for human responses. On a human level, object-similarity judgment informs how we interact with objects (Desmarais et al., 2007), organize our world (Smith, 1981) and acquire new concepts from a young age (Markman and Hutchinson, 1984). Meanwhile, many corpus-based computational models, including deep transformer models that leverage corpora such as ChatGPT, leverage lexical co-occurrence relations to derive semantic meaning (i.e. the distributional hypothesis). Despite differences in process, these models' representations display correspondences with human judgment (Torabi Asr et al., 2018; Chandrasekaran and Mago, 2022).

In this paper, we collect GPT-3.5 responses to an object similarity task introduced by Hebart et al. (2020). We reformat their image-based paradigm as a chat-completion task for GPT.[2] Like Hebart et al., we also train a sparse embedding model that can predict object-similarity judgments. We annotate the dimensions of the embedding model to provide an interpretable characterization of the reasoning behind such judgments. Finally, we compare the GPT- and human-derived characterizations and embeddings. We also simulate the effects of GPT response replacement and augmentation. To do this, we train models on different mixtures and proportions of human and GPT responses, then compare their embedding spaces to baseline human-derived ones.

[1] There is evidence that LLMs may already be incorporated into annotation workflows without researcher knowledge, as crowdworkers are already using LLMs to speed up their annotation tasks (Veselovsky et al., 2023).

[2] At the time of our experimentation, the multimodal GPT-4 was not widely available.

2 Methodology

The Odd-One-Out (OOO) Task To obtain object-similarity responses from GPT, we used the ***odd-one-out*** (OOO) task, wherein participants indicate the least similar amongst three objects. For example, we might ask, "Which of these concepts is the odd one out: apple, banana, car?" and expect factors such as edibility to affect the response. The OOO task is well-established in the field of psychology for eliciting concept-relational preferences (Mirman et al., 2017; Valenti and Firestone, 2019).

Human OOO Responses Hebart et al. (2020) used an image-based OOO task to collect millions of object-similarity judgements. They did this in two rounds, first collecting 1.46M responses (Hebart et al., 2020), then creating a larger, 5M response dataset Hebart et al. (2023).[3] We used these two datasets to create two disjoint OOO response sets of equal size (1.46M). We refer to the first of these datasets as the ***full human dataset*** and the second as the ***baseline dataset***.

GPT OOO responses We then created a parallel GPT-only dataset with answers to the OOO questions from the full human dataset. We reformatted the original prompt from (Hebart et al., 2020) to create a text completion task suitable for GPT. We referred to these GPT prompts and answers as the ***full GPT dataset***.

For cost and task-efficacy reasons, we used OpenAI's GPT (GPT-3.5-Turbo-0613). Preliminary analysis revealed that smaller models (Falcon-7B, Alpaca-7B, Vicuna-7B) had difficulty answering odd-one-out questions in a coherent manner with simple prompting. Larger models, (e.g. Falcon-40B), produced coherent responses, but not at the scale afforded by GPT's API.

Transformer models such as GPT incorporate word position for next-word prediction, and GPT demonstrated a strong positional preference (see Appendix C). While humans situationally exhibit ordered preferences, we found a roughly uniform distribution for this task (see Appendix C). Thus, to collect position-neutral responses, we permuted the order of the three objects in the prompts to create six total questions (3!). We then used relative majority voting across the six questions to compute GPT's odd-one-out choice, breaking ties randomly.

[3]These datasets were collected before GPT existed and thus are free of GPT-derived responses.

	metallic	food-related	$\cdots$	cylindrical
aardvark	$a_{1,1}$	$a_{1,2}$	$\cdots$	$a_{1,49}$
abacus	$a_{2,1}$	$a_{2,2}$	$\cdots$	$a_{2,49}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
zucchini	$a_{1854,1}$	$a_{1854,2}$	$\cdots$	$a_{1854,49}$

Learned object-similarity embeddings

Figure 1: An example embedding space with words as rows and *characterizing dimensions* as columns.

See Supplementary Materials for API calls and a formatted table of all responses.

Human–GPT Datasets We aimed to study the effect of replacing only some human responses with GPT responses. Thus, we created <1.46M count ***partial human response sets*** by taking proportions $[0.125, 0.25, 0.375, 0.5, 0.625, 0.75,$ and $0.875]$ of the 1.46M full human-only response set. We then create a 1.46M-count ***mixed GPT–human response set*** for each partial human set by considering each unused human response and including the corresponding GPT response.

2.1 Model Details

We use the similarity-prediction model designed by Hebart et al. (Hebart et al., 2020), which comprises a shallow neural network consisting of a single 90×1854 embedding layer. Each object i has a corresponding vector v_i. In a triplet with objects i, j, and k we compute $z_i = v_j \cdot v_k$ (and do likewise for z_j and z_k), then use it to estimate the probability of object i being the odd one out:

$$\mathrm{P}(i \text{ odd one out}) = \sigma(\mathbf{z})_k = \frac{e^{z_k}}{e^{z_i} + e^{z_j} + e^{z_k}} \quad (1)$$

Model Training To train each model, we used a cross-entropy loss with an ℓ^1-norm penalty on the embedding to encourage sparsity. Hebart et al. (2020) found that training sparse models in this manner resulted in an embedding space with interpretable dimensions. We refer to these dimensions as ***characterizing dimensions***. We show an example embedding matrix in Figure 1 wherein the rows are the vector representations of the object-concepts of the THINGS dataset, and the columns are characterizing dimensions.

Using a set of odd-one-out responses $\mathcal{S}$, we took the average cross-entropy loss, $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} H(q,p)|s$. Here, $H(q,p)|s$ is the

cross-entropy of the model prediction probability p for the odd-one-out question s relative to the entry q in the actual one-hot response vector. We incorporate an ℓ^1-norm penalty on the embedding space to encourage sparsity, weighted by a hyperparameter λ. Elaborated loss details are given in Appendix D.

For training, we assumed concavity of validation accuracy on the choice of λ and performed a two-tiered four-fold grid-search over 90–10 train–test dataset splits: we started with $\lambda = 0.0064$ and took steps of 0.0016 to find a coarse maximum, then took steps of 0.0004 around that coarse maximum to establish a finer maximum. We trained for a fixed 1000 epochs for each model, mirroring the setup of Zheng et al. (2019) to ensure convergence. Further specifics are given in Appendix E.

We trained ten models each on the full human, full GPT, and baseline human sets and four each on the partial human and mixed human–GPT datasets to produce ***full human, full GPT, partial human, mixed human–GPT, and baseline models***.

2.2 All-GPT Model Characterization

To better understand the basis for GPT responses to OOO questions, we manually annotated each dimension of the full GPT embedding space as in Hebart et al. (2020). Annotators were presented with images of objects at pre-determined intervals along a dimension's range (e.g., Appendix F). Six respondents gave up to three descriptors for each dimension. We iteratively generated aggregate labels for each annotation until none were ungrouped, then chose the aggregate labels that covered the most participants. We call this the ***labelled GPT model***, and we compare it to a previous ***labelled human model*** produced with the full human dataset from Hebart et al. (2020).

The labels for the nine dimensions with the highest means are given in Figure 2, while those for the 39 dimensions with max value above 0.1 are given in Appendix G; see Supplementary Materials for raw responses and coding.

Labelled Correlations We computed the correlations of each of these GPT-derived dimensions with dimensions from the labelled human model. The correlations of the top 9 dimensions (by column mean) from each labelled model are shown in Figure 2; the full 39-by-49 correlation matrix, as well as correlation matrices ordered by maximal correlation matching, appear in Appendix H.

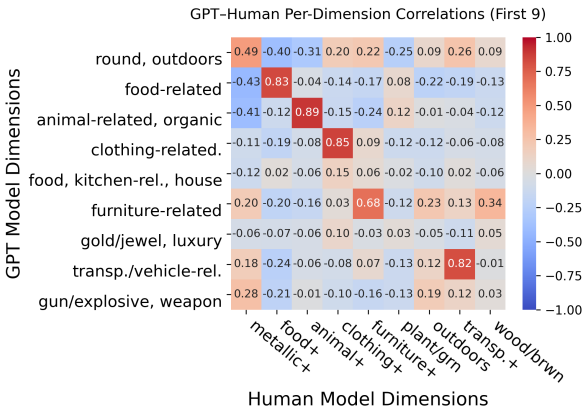


Figure 2: Correlation heatmap between the first 9 dimensions of the labelled GPT model and the labelled human model (all-dimension version located in Appendix H). Note the strong correlations between similarly labelled dimensions. The similar labels indicate agreement concerning what the dimensions convey in their scores for concepts, while the correlations indicate that the dimensions have statistical agreement.

We also performed PCA and UMAP (McInnes et al., 2018) on the labelled dimensions, which are displayed in Appendix J.

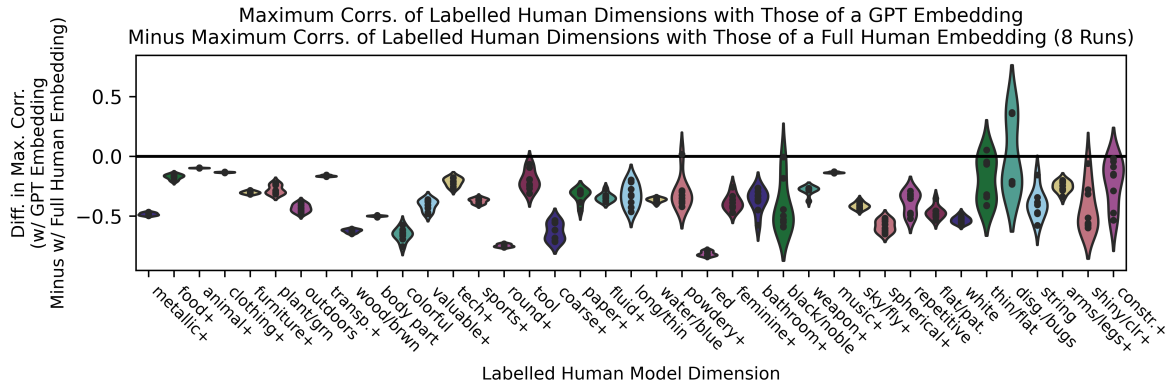
2.3 GPT–Human Response Substitutability

To determine the impact of augmenting human responses with GPT responses, we compared embedding spaces trained on datasets with varying amounts of each. For this comparison, we used representational similarity analysis (RSA) (Kriegeskorte, 2008) with a linear kernel.

Given two embeddings $\mathbf{X}_1$ and $\mathbf{X}_2$, we obtained their respective Gram matrices $\text{sim}(\mathbf{X}) = \mathbf{X}^\top\mathbf{X}$. These are the representational similarity matrices, or RSMs, of each space. Then, we calculated the Pearson correlation between the upper triangle of each RSM. The result is the ***RSA correlation***, and we report an ***RSA score***, the average RSA correlation of a model with the baseline human models.

GPT Response Substitution Given a full human dataset, if we replace some of the human responses with GPT responses, how does that affect the RSA score? Here, we are comparing the purple pluses with the large red circle in Figure 4. To examine the effects of mixing GPT completion-driven responses into a human dataset, we computed the RSA scores of the mixed human–GPT embeddings. These results are given in Figure 4. A table of these values can be found in Appendix K. Even though the datasets were larger, the mixed GPT–human embeddings each have lower RSA scores

Figure 3: Differences: maximal correlations of the labelled human characterizing dimensions with any dimension of a full GPT model and with any dimension of a full human model (the full GPT correlations minus the full human correlations) over 8 such models of each. For the correlations in isolation, see Appendix I.



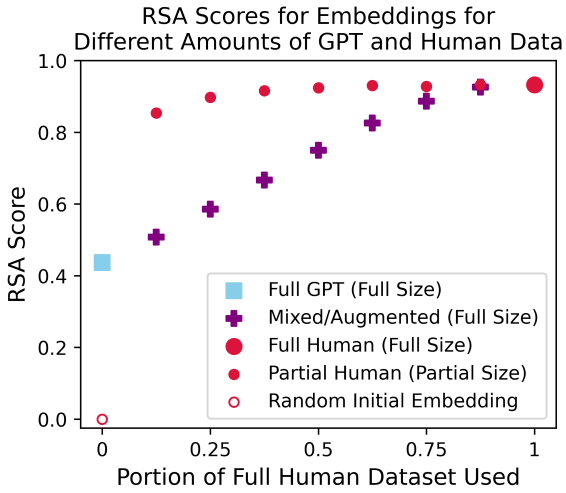


Figure 4: Average RSA scores for full GPT (blue), mixed GPT-human (purple), and full human (large red) models. Also plotted are the scores for the smaller, partial human (small red) models. The x-axis is the proportion of the original human dataset in each model's training set. The RSA score for a no-data, random embedding (small hollow red) is given for comparison. A table of these results is located in Appendix K.

than the corresponding partial human embeddings. The scores trend downward in a sigmoid fashion as the proportion of human data decreases, with the most noticeable effects happening after .25 of the human data has been replaced.

GPT Response Augmentation Next, we compared models trained on the same amount of human data, but with differing amounts of GPT augmentation. In contrast to the previous paragraph, in this situation we are comparing models trained on datasets of differing size. Comparing these models tells us whether adding GPT data hindered, facilitated, or neutrally impacted the final model's ability to represent human similarity judgment. To make this comparison, consider the small red circles and the corresponding purple plusses in Figure 4. We found that for all tested ratios, augmenting with GPT data results in lower RSA scores even though the resulting dataset size has increased.

Individual Dimension Capturing Finally, we explored the correspondence of dimensions from the labelled all-human model to those of the full GPT embeddings. To do this, we started with a full GPT embedding and full human embedding. For each labelled human dimension, we found the dimension of maximal correlation in the full GPT embedding and the dimension of maximal correlation in the full human embedding. These correlations signify the full GPT and full human embeddings' ability to reproduce each labelled human dimension. We then subtracted the full human correlations from the full GPT correlations to determine how much worse one was at capturing the labelled human dimensions (more negative corresponds with the full GPT embedding doing worse). We did this 8 times; the results are given in Figure 3. The correlations themselves, as well as graphs for the same process but with labelled GPT dimensions, are given in Appendix I.

3 Conclusions

Our work illustrates GPT's judgment in an odd-one-out similarity task, provides 39 judgment-characterizing dimensions with human annotations, and compares those dimensions with those derived from a human-only model. Notably, many GPT (and human) dimensions have similar, shared-word-or-synonym labelling, such as food-related (food-

related/eating-related/kitchen-related) and animal-related/organic (animal-related/organic). We compared the labelled GPT and human dimensions, and we found that over half of the labelled GPT dimensions had correlations above 0.5 with a similarly labelled human one (for individual results, see Figure 2 or Figure 12). However, while using GPT responses did produce many characterizing dimensions similar to those derived from human responses, substituting in GPT responses still resulted in worse approximations of human decision-making under RSA, as demonstrated by Figure 4. Some of this is likely attributable to modality differences between the image and text questions, as some of the dimensions least captured by the model are color-oriented, such as "wood/brownish", "red", and "colorful", as shown in Figure 3 and Appendix I.

Surprisingly, even when we used relatively little human data, adding GPT responses did not improve the trained model's RSA score. As shown by Figure 4, there must be a point below which this improvement appears, as the full GPT RSA score is above 0.4, a randomly initialized embedding has an RSA score of 0, and there's very little information present in, e.g., one triplet. However, that point was below the lowest proportion we tested, as shown by gaps between the RSA scores of the partial human models and the GPT-augmented mixed human–GPT models in Figure 4. This ostensibly contradicts previous studies showing LLMs having human-comparable performance on a wide variety of tasks, but *human-level performance is different than human behavior*. Partial human RSA scores largely saturated by the time 0.375x the full human dataset was used, indicating that for RSA purposes, the sizes of the partial human datasets we considered may have been larger than needed. Nonetheless, it seems misguided to augment with GPT responses unless human data is considerably scarcer than we tested, as the mixed human–GPT datasets have considerably worse RSA scores than those observed on much smaller partial human datasets. Were our task extremely low-resource, the 0.42 RSA score achieved by the full GPT model might be useful, however.

Encouragingly, when swapping human responses with those from GPT, the RSA scores appeared fairly robust to replacement, as shown by the full human and partial human models' RSA scores in Figure 4. When we replaced 25% of the data (0.75 on the Human-Data-Portion x-axis) the RSA score dropped by less than 10%. There was at most a $\sim$60% reduction when 100% was replaced (0 on the Human-Data-Portion x-axis). This is important to consider for future crowdsourced odd-one-out experiments, such as the Hebart dataset, because many crowdworkers have begun using GPT for their tasks.

In conclusion, our work characterizes GPT object-similarity judgments, enhancing our understanding of how LLMs and humans behave similarly or differently. Notably, despite a modality difference, GPT responses produced embeddings with labels mirroring or closely resembling those from human responses. Our findings also indicate utility in using LLM completions for extremely low-resource environments as a proxy for human judgment. However, these findings suggest little benefit from *augmenting* human responses for any sizeable number, especially when crowdsourcing human data is feasible. Our findings also warrant caution when otherwise human-looking GPT responses might become part of collected data, but offer hope for the odd-one-out task, as the embeddings proved fairly robust to lower levels of response replacement.

4 Future Work

Our choice of LLM for our experiment was constrained by the sizes of (effective) current models, computing resources, and modality. As image-capable and more powerful models appear, future work should repeat our experiments using them.

Future work may also examine whether the choice of dataset affects the characterizing dimensions produced for an LLM. The THINGS dataset is a set of concrete objects, and lacks more expansive concepts like scenes, environments, actions, or emotions. GPT gives us a budget-effective way of considering whether the introduction of such concepts might change what characterizing dimensions appear.

Limitations

Our work uses text-only prompts, while the human experiment uses images. The objects of the THINGS dataset were chosen to be highly imageable, but this nonetheless almost certainly played a role in shaping what GPT found salient in the object-comparison task. At time of writing, GPT-4's vision API had not seen full release.

Our prompts presented GPT with objects in an ordered fashion that it heavily utilized (see Appendix C). To remedy this, we used aggregate responses on permuted prompts. However, humans may have used the ordering of questions (or responses from previous questions) in ways our setup did not account for.

We used OpenAI's GPT-3.5. It is possible certain aspects of our characterization are specific to it. In particular, we anecdotally observed that smaller models had difficulty completing the odd-one-out task as far as we could understand; other models likely exhibit more or less similar behavior to humans as well.

During the survey, multiple respondents mentioned that the percentile structure made it difficult to discern continuous meaning across the entire dimension scale. This may be because the dimensions only hold palpable information at higher levels. Regardless, the common strategy employed was to look at the top and bottom objects rather than the ones in the middle. Our percentiles were chosen to align with previous work, but nonetheless, other methods may elucidate more nuances than our prompt and coding schema did.

Finally, GPT-3.5 is largely trained on English text and corpora. This has cultural and linguistic implications, and future work may wish to consider examining models trained specifically on data from other languages or specific communities.

Our work serves as one data point for understanding LLMs. This should be sufficient for giving insight into related work, but (especially given the quickly-arriving ubiquity of LLMs and potential for harm; see Ethics), it is not in isolation nearly sufficient for determining whether LLMs should be used in real-world applications.

Ethics

Risks

Our model illuminates GPT's behavior in a direct odd-one-out task, and some of the characterizing dimensions have strong correlation with previously obtained dimensions that characterize human object-similarity judgment. There is a potential to misinterpret this as meaning GPT uses these dimensions in the same way humans do or that these dimensions apply to all tasks GPT performs.

Resources

Response-collection was performed using OpenAI's `GPT-3.5-Turbo-0613` endpoint. The 4,385,040 responses took one week for OpenAI's systems to process at a total cost of $722 USD. Training was done with NVIDIA P100 GPUs on Digital Research Alliance (Compute Canada) clusters, taking about 16 hours per model.

Licensing and Artifacts

Our GPT odd-one-out response dataset and model are available under a CC-BY version 4 licence in Supplementary Materials. The intended use of our dataset is general-purpose, so long as it is not harmful.

We use the THINGS images dataset (Hebart et al., 2019) under the terms of the CC BY 4.0 under which it was released (`https://osf.io/qyd6u`). We use the THINGS odd-one-out dataset (Hebart et al., 2023) under the terms of the CC-BY-4.0 license under which it was released (`https://osf.io/5wcte`). Its intended use is to further research (as per the Things Initiative's website (Hebart et al., 2019)).

We use Pandas (pandas development team (2020); Wes McKinney (2010)) under its BSD 3 licence. We use Scikit-Learn (Pedregosa et al., 2011) under another BSD 3 licence. We use SciPy (Virtanen et al., 2020) under the terms of a similar licence. We use Matplotlib (Hunter, 2007) under a BSD-like licence. Finally, we also use PyTorch (Paszke et al., 2019). We satisfy the licensing terms of it, along with the previous software packages, by not redistributing the source code. These software packages' intended use is scientific and general-purpose application, and we satisfy both those criteria.

We also use representational similarity analysis (RSA) (Kriegeskorte, 2008) and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018). Kriegeskorte and McInnes both likely intended others to use their algorithms for general research.

We use ChatGPT-3.5 and ChatGPT-4 for some code generation under OpenAI's commercial terms. At no point do we provide sensitive or copyrighted information to it.

Response Collection

All respondents were members of the same research team. However, as responses were collected using respondents' choices of identifying keywords

(initials were suggested), that identification was removed from any public release. This minimal information was necessary because respondents were informed they could have their responses deleted, should they desire. All respondents were part of the research team; no formal recruitment was done. For the same reason, no compensation was given. Respondents knew ahead of time what this project was for, but details were given in the instructions as well.

The instructions given can be found in Supplementary Materials.

All responses were from graduate students and postdocs at a leading university. The respondents' countries of origin were diverse (only two respondents were from the same country), and all were fluent in English, although for half, it was not a first language.

Supplementary Materials

All supplementary materials, including code, datasets, and grid-search results, are available at `https://osf.io/7vz2h/`.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences and Humanities Research Council (SSHRC), the Digital Research Alliance of Canada (`alliancecan.ca`), and the Canadian Institute for Advanced Research (CIFAR). Alona Fyshe holds a Canada CIFAR AI Chair.

References

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.

Dhivya Chandrasekaran and Vijay Mago. 2022. Evolution of Semantic Similarity—A Survey. *ACM Computing Surveys*, 54(2):1–37.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation. ArXiv:2302.13007 [cs].

Geneviève Desmarais, Maria Cristina Pensa, Mike J. Dixon, and Eric A. Roy. 2007. The importance of object similarity in the production and identification of actions associated with objects. *Journal of the International Neuropsychological Society*, 13(6):1021–1034.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. ArXiv:2303.15056 [cs].

Martin Hebart, Oliver Contier, and Lina Teichmann. 2023. Things-odd-one-out. *Open Science Framework*.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. ArXiv:2302.09210 [cs].

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Bin Ji. 2023. VicunaNER: Zero/Few-shot Named Entity Recognition using Vicuna. ArXiv:2305.03253 [cs].

Nikolaus Kriegeskorte. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.

Chen Li, Yixiao Ge, Jiayong Mao, Dian Li, and Ying Shan. 2023. TagGPT: Large Language Models are Zero-shot Multimodal Taggers. ArXiv:2304.03022 [cs].

Ellen M. Markman and Jean E. Hutchinson. 1984. Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, 16(1):1–27.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Daniel Mirman, Jon-Frederick Landrigan, and Allison E. Britt. 2017. Taxonomic and thematic semantic systems. *Psychological Bulletin*, 143(5):499–520.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. ArXiv:2304.13861 [physics].

The pandas development team. 2020. pandas-dev/pandas: Pandas.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Linda B Smith. 1981. Importance of the Overall Similarity of Objects for Adults' and Children's Classifications. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4):811–824.

Fatemeh Torabi Asr, Robert Zinkov, and Michael Jones. 2018. Querying Word Embeddings for Similarity and Relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 675–684, New Orleans, Louisiana. Association for Computational Linguistics.

J.J. Valenti and Chaz Firestone. 2019. Finding the "odd one out": Memory color effects and the logic of appearance. *Cognition*, 191:103934.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. ArXiv:2306.07899 [cs].

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. 2019. Revealing interpretable object representations from human behavior. ArXiv:1901.02915 [stat.ML].

A Response-Order Counts

For a set of triplets, each object is either ordered first, second, or third in their presentation to a respondent. Below are the holistic choice rates for each in the odd-one-out task for for GPT (Figure 5), for humans (Figure 6), and for GPT aggregated (Figure 7).

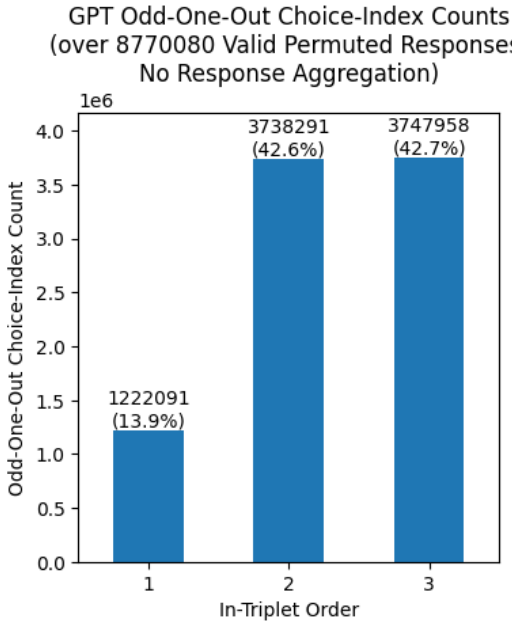


Figure 5: Counts of order-within-triplet responses for raw GPT calls. For example, given a prompt asking about 'apple', 'banana', and 'car', in that order, and a response of 'car', this would be a response with an index of 3. These are unbalanced, so we resort to permuting them; see section 2, Human–GPT Datasets for details of this.

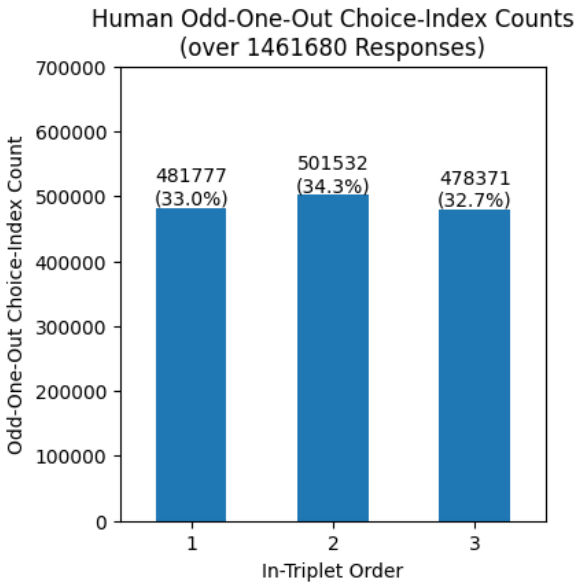


Figure 6: Counts of order-within-triplet responses for adult respondents on the dataset. For example, given a prompt asking about 'apple', 'banana', and 'car', in that order, and a response of 'car', this would be a response with an index of 3. These responses are from (Hebart et al., 2020).

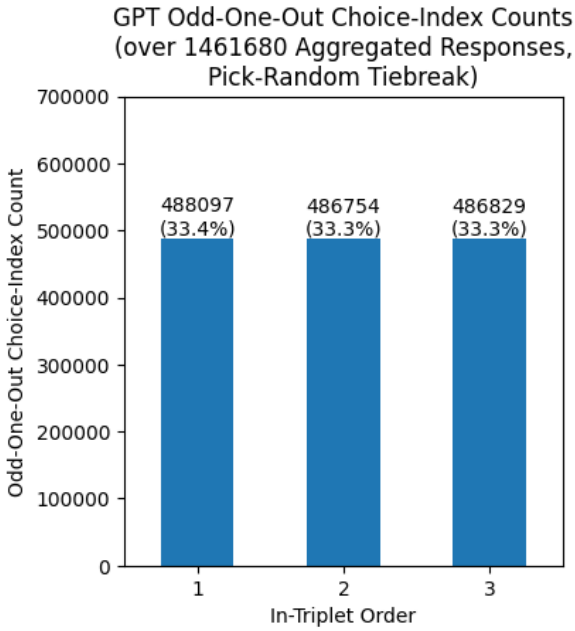


Figure 7: Counts of order-within-triplet responses for aggregated GPT calls. For example, given 'apple', 'banana', and 'car', if the relative majority vote was 'banana', this would be a response of index 2. In the case of tiebreaks, in actuality the earliest tiebreaking indexed response was chosen; this is easier to reproduce and works out to be equivalent to choosing randomly due to the orders of the objects within the questions being completely random. See section 2, Human–GPT Datasets for permutation details.

B Odd-One-Out Prompt

The prompts we provided to GPT were of the following form:

```
<|im_start|>system
Which of the objects are more similar to
    each other? Say the object that
    doesn't match. Format your choice as
    [[object]]<|im_end|>
<|im_start|>user
{object1}, {object2}, {object3}.<|im_end
    |>
```

This was intended to be as close to the language used by (Hebart et al., 2020) as possible. Their instruction example is as follows:

```
The three pictures show {object1}, {
    object2}, and {object3}. Which are
    more similar to each other? Click on
    the picture that doesn't match.
```

C Permuted Response Distribution

For a given set of three objects, GPT may answer differently when the objects' order is permuted in the prompt. The rates of agreement of these individual permutations with the accepted aggregate response are given in Figure 8.

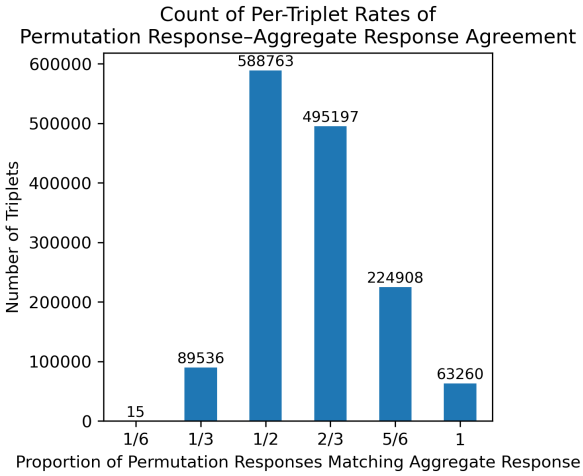


Figure 8: Distribution of the rate of agreement of model permutation responses with the aggregate model response (see section 2, Human–GPT Datasets for permuting details). 1.00 denotes that all 6 permutations of an odd-one-out triplet resulted in the same response; $\frac{2}{3}$indicate that 4 of 6 permutations resulted in the same response. $\frac{1}{2}$ and $\frac{2}{3}$ indicate possible ties, which were broken by choosing the first response at a tying index. Due to the questions being random ordered, consistently doing this is equivalent to choosing randomly between the options with the most votes.

D Model Loss

The cross-entropy loss used by the model in training is given here.

$$
\begin{aligned}
&H(q,p)_{\substack{\text{object set is } \{i,j,k\},\\ \text{k is the odd-one-out}}} \\
&= \sum_{c\in\{i,j,k\}} q_{c \text{ is the odd-one-out}} \cdot \ln(p_{c \text{ is the odd-one-out}}) \\
&= -\ln\left(p(c_{\text{odd-one-out}})\right) \\
&= -\ln\left(\sigma\left(\mathbf{z}\right)_c\right) = -\ln\frac{e^{z_k}}{e^{z_k}+e^{z_j}+e^{z_i}}
\end{aligned}
$$

where

- H is the cross-entropy loss function
- i, j, k denote the three objects of a triplet, where k is the true odd-one-out
- z_c where $c \in \{i,j,k\}$ and z_c represents the dot product between the vectors of the pair of objects $\{i,j,k\} \setminus \{c\}$
- $\mathbf{z} = \{z_i, z_j, z_k\}$
- σ is the softmax function
- q is the probability of an object being the odd one out (so 100% for the identified odd-one-out, 0% for any other object)
- p is the estimated probability the model gives that a given object is the odd-one-out

For the ℓ^1-norm penalty, we flatten the embedding matrix and take the ℓ^1 norm of the resulting vector. We weight this norm by $\lambda/\text{num\_items}$ and add it to the cross-entropy loss to obtain our full loss.

E Grid Search Specifics

For a given training set, we perform a grid search: we take steps of 0.0016 over the range $\lambda \in \{0.0064..0.0144\}$ to find a maximum, expanding the search radius if necessary. We then perform $(k=4)$-fold cross-validation ($(k=10)$-fold for the full GPT set) in steps of 0.0004 to the adjacent previously-found 0.0016-stepped lambdas to find the optimal lambda in the region around that local maximum. We train on a 90% split for a fixed 1000 epochs for each model, mirroring the setup of Zheng et al. (2019) to ensure convergence. The per-epoch performance and final validation accuracies for the grid-search folds of the full GPT model are given in Figure 9. The final validation accuracies for those λs are given in Figure 10, illustrating the degree of local concavity. All grid-search results, as well as further by-fold stats for the mixed human–GPT and partial human models, are found in Supplementary Materials.

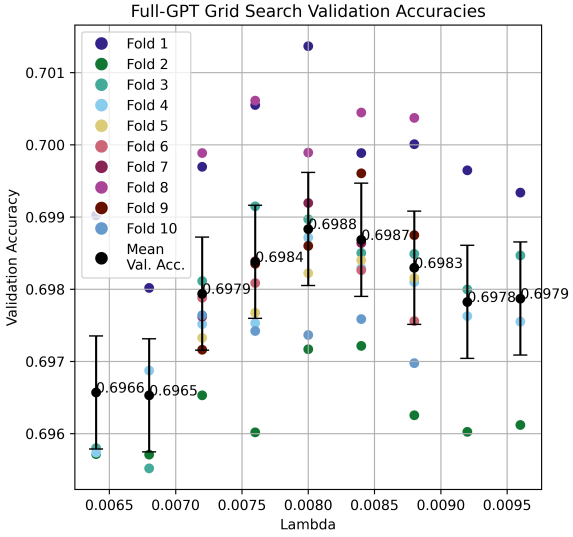


Figure 10: Step 2 of the grid-search for the full GPT model (stepping at intervals of 0.0004). The x-axis gives training lambda values, and the y-axis gives the validation accuracy at 1000 epochs. The error bars assume fold results are normally distributed and give a range of one standard deviation. $\lambda = 0.008$ is the highest performer.

F Dimension Scales

For each dimension, we produced scales with objects whose values spanned the dimension, as in Figure 11.

Namely, we made images as seen in Figure 11. The six images on the left have Dimension 12 values at the 0th, 1st, 5th, 10th, 15th, and 20th percentiles for the dimension. The images at the next tick have dimension values at the 33rd percentile, and thereafter the images at each successive tick are at a percentile 13.333 more. This continues until the last tick, denoting the 100th percentile, where the six top-scoring images are shown.

G Dimension Labels

The aggregated dimension names for the 39 largest dimensions of the labelled GPT model are given in Table 1.

H Correlation Heatmaps

The full heatmap of the correlations between the dimensions of the labelled GPT model and those

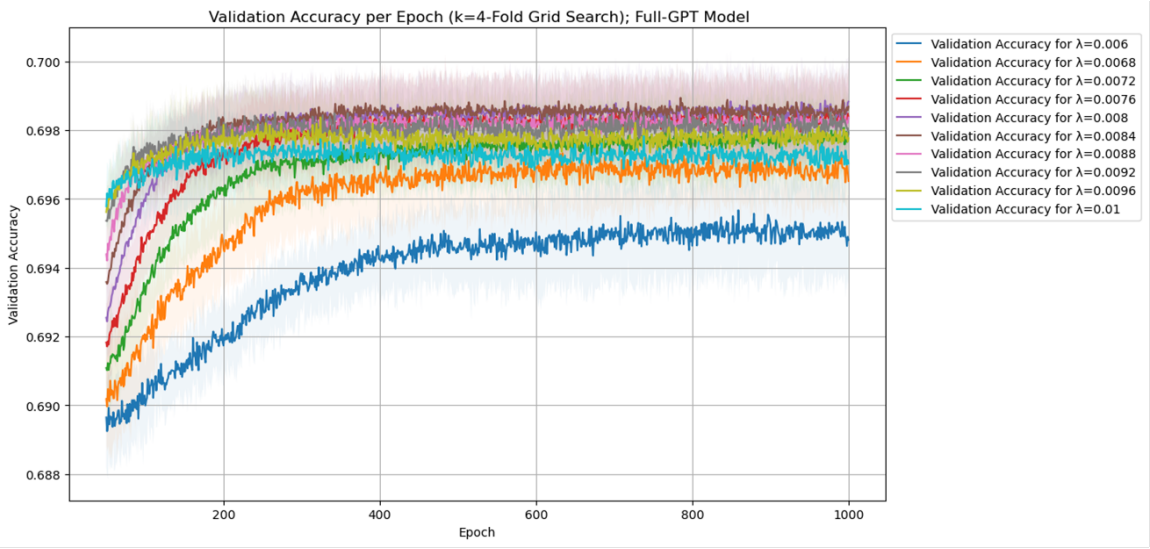


Figure 9: Per-epoch validation accuracies for step 2 of the grid search for the full GPT model (with additional lambdas for illustration). Note the saturation of the validation accuracies before 1000 epochs. Additional 0.0004-intervaled grid-search results are included for context. $\lambda = 0.008$ is the highest-scoring performer.

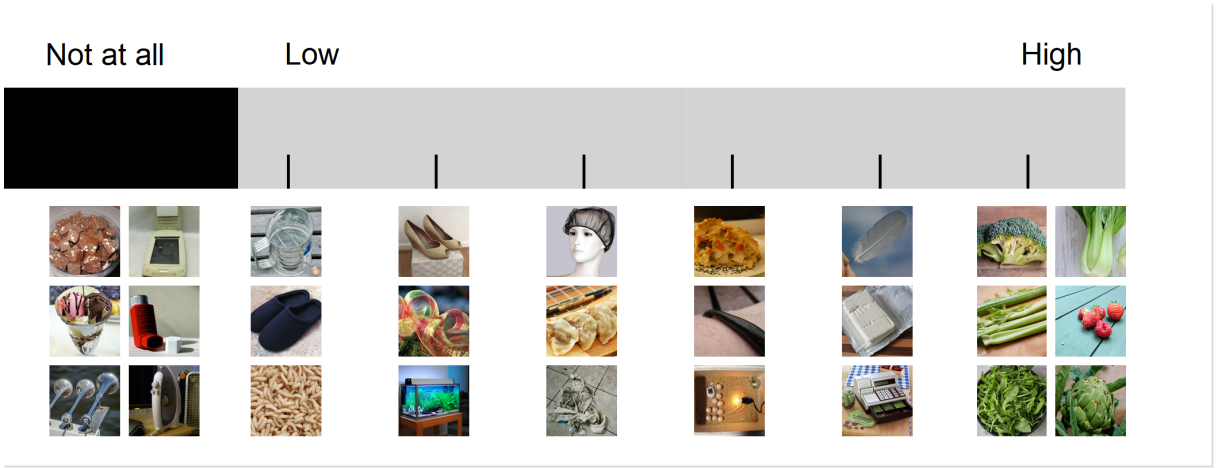


Figure 11: Scale produced for Dimension 12 of the full GPT model for annotations

Dimension Ordering	Aggregate Dimension Label	Dimension Ordering	Aggregate Dimension Label
1	round, outdoors	21	alive/nature/plant-related
2	food-related	22	boats/water-related
3	animal-related, organic	23	box/container-related
4	clothing-related	24	sports-related
5	food, kitchen-related, house	25	small, (flying) insect-related
6	furniture-related	26	music-related
7	gold/jewel, luxury, ostentatious	27	vehicle-related, outdoors
8	transportation/vehicle-related	28	fruit-related
9	gun/explosive, weapon	29	aquatic/sea-related
10	electronics-related	30	crafts, push item through hole
11	(melee) weapon, long/thin	31	wound/rolled, thread-related
12	edible/vegetable-related	32	round, colorful, sports
13	tool-related	33	sanitation, garbage-related
14	(sharp) tools	34	medical (equipment/tools)
15	delicious/sweet liquid/food	35	toy-related
16	(metallic) housing hardware-related	36	vertical, elevated
17	earth/rock-related	37	industrial/mechanical
18	candy/sweet, food	38	paper/literacy-related
19	textiles	39	temperature/temperature-change related
20	container, tableware-related		

Table 1: Aggregate labels for the characterizing dimensions of the labelled GPT model. Labels were obtained via the coding process described in subsection 2.2.

of the labelled human is shown in Figure 12. To illustrate the closest dimensions between the labelled GPT and labelled human embeddings, we performed a bipartite max-correlation-as-weight matching of the labelled human dimensions to the labelled GPT embeddings Figure 13 (and vice-versa in Figure 14).

I Dimension Reproducibility and Overlap

We wished to gauge the reproducibility of the labelled GPT/Human embedding dimensions and determine the extent to which the dimensions of one are reproduced by the other. To determine this, we took our labelled human model and considered each dimension. We ran 8 other full human models and 8 other full GPT models, each time calculating the maximal correlation that the labelled dimension had with any of the new dimensions. This told us (1) how reproducible the labelled human dimensions were, and (2) the extent to which full GPT models captured the labelled human dimensions. The differences between the respective GPT and human correlations then conveyed how much better or worse the typical full human or full GPT embedding was at reproducing the labelled human dimensions. These results are shown in Figure 15.

We also repeated this setup using the labelled GPT model as the basis of comparison. This conveyed the reproducibility of the labelled GPT dimensions and the extent to which full human models captured the labelled GPT dimensions. Similar to before, the differences between the respective GPT and human correlations then indicated how much better or worse the typical full human or full GPT embedding was at reproducing labelled GPT dimensions. These results are shown in Figure 16.

J Dimension UMAP and PCA

We performed Uniform Manifold Approximation and Projection (UMAP) and Principal Component Analysis (PCA) on the labelled human and GPT embeddings for insight into the dimensions' spatial relationships. These are given in Figure 17.

Most of the largest human embedding dimensions have a strong correlation and corresponding label with a dimension in the GPT embedding (and vice-versa). However, the largest-magnitude dimension of each are quite different. These two dimensions have an outsized effect on the choice of principal components, as evidenced by them

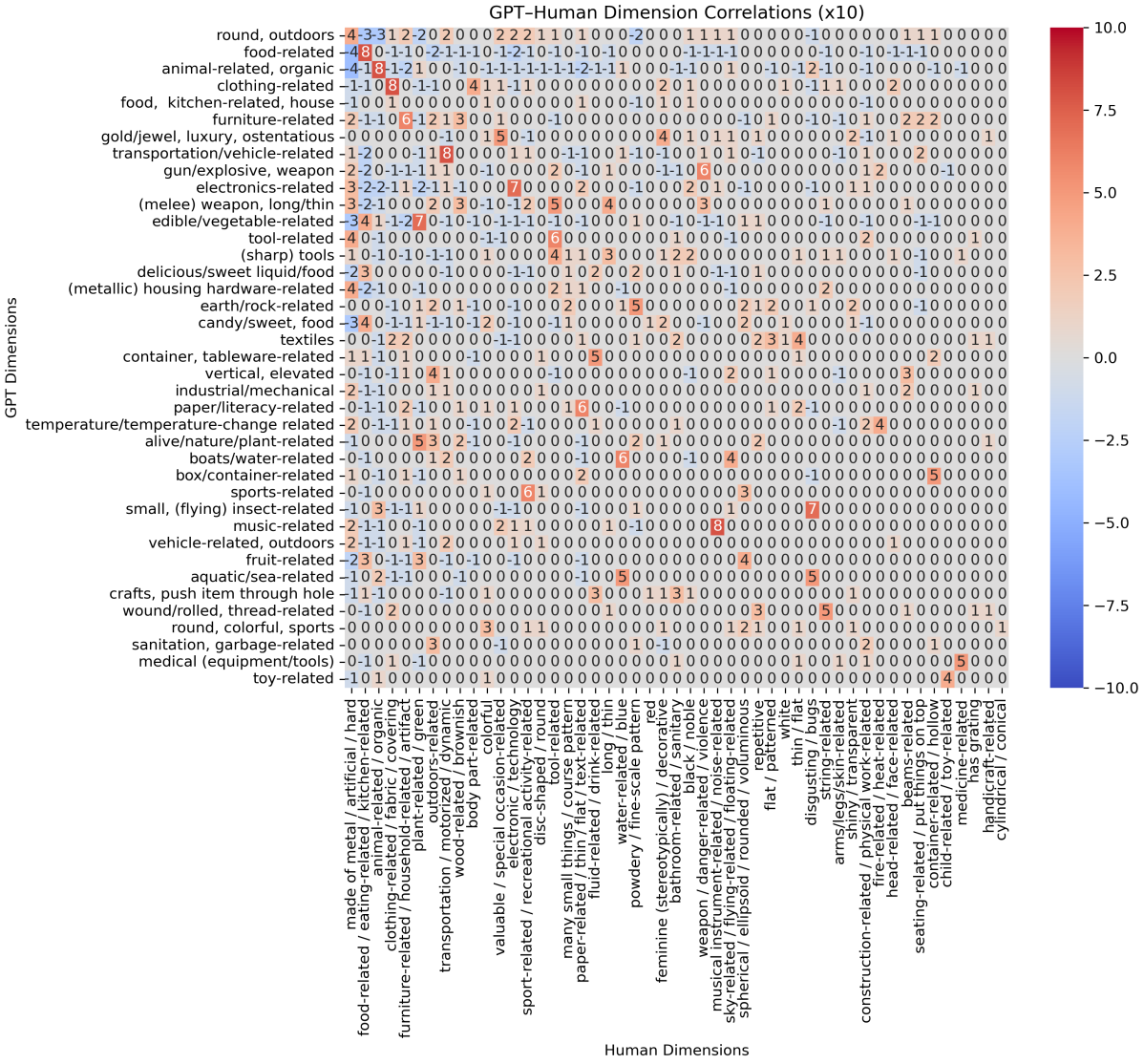


Figure 12: Full correlation heatmap between the dimensions of the labelled GPT model and the labelled human model, with aggregate labels on left. Dimensions are ordered by the mean value over objects. Correlations are multiplied by 10 and rounded to the nearest integer for text-size reasons.

Figure 13: Correlation heatmap between each labelled GPT embedding dimension and the closest labelled human embedding dimension under bipartite max-correlation matching. The GPT dimensions are ordered by their mean value over all objects.

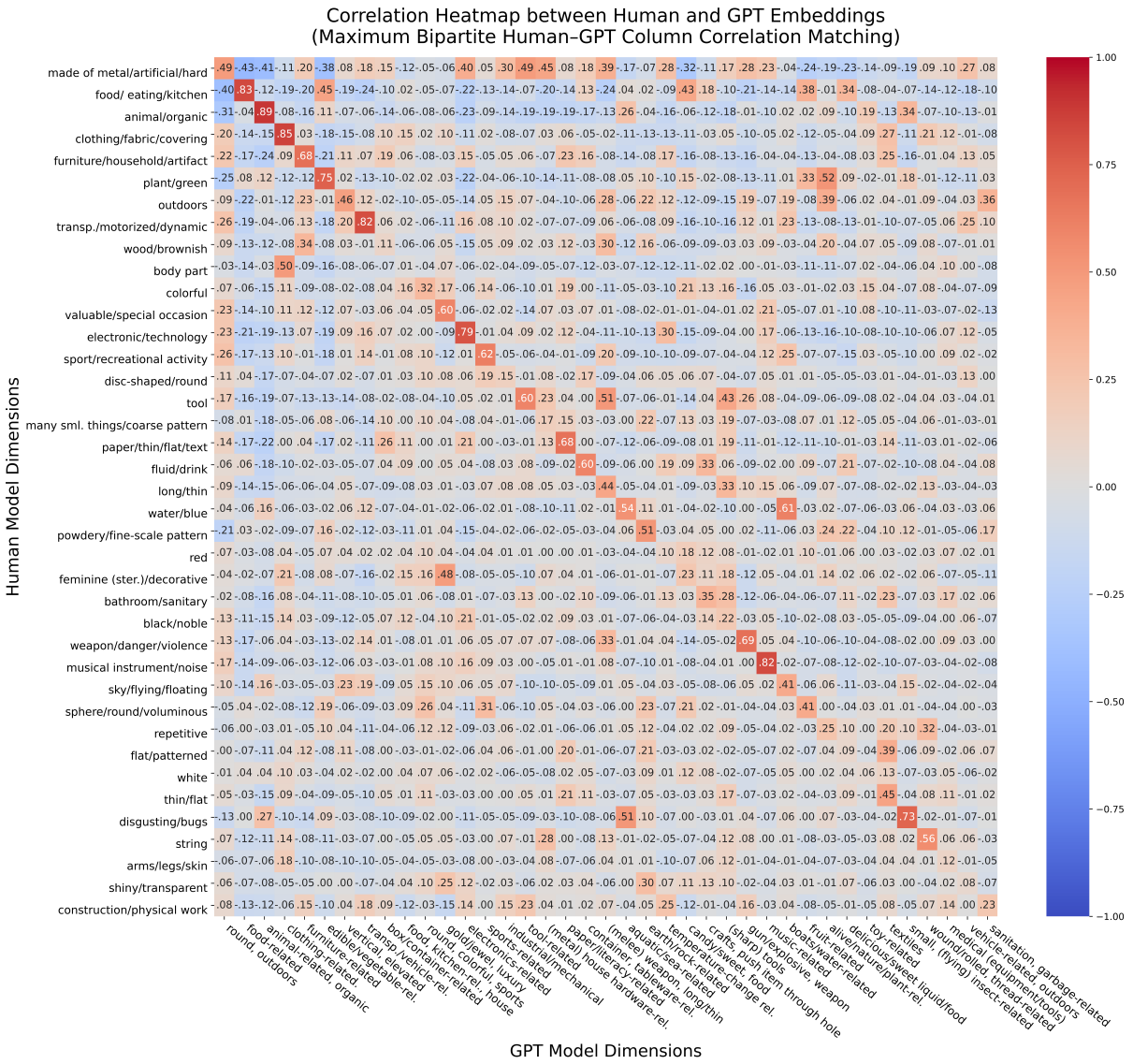


Figure 14: Correlation heatmap between each labelled human embedding dimension and the closest labelled GPT embedding dimension under bipartite max-correlation matching. The human dimensions are ordered by their mean value over all objects.

Figure 15: Maximal correlations of the labelled human characterizing dimensions with any dimension of a full human model and a full GPT model (over 8 such models of each).

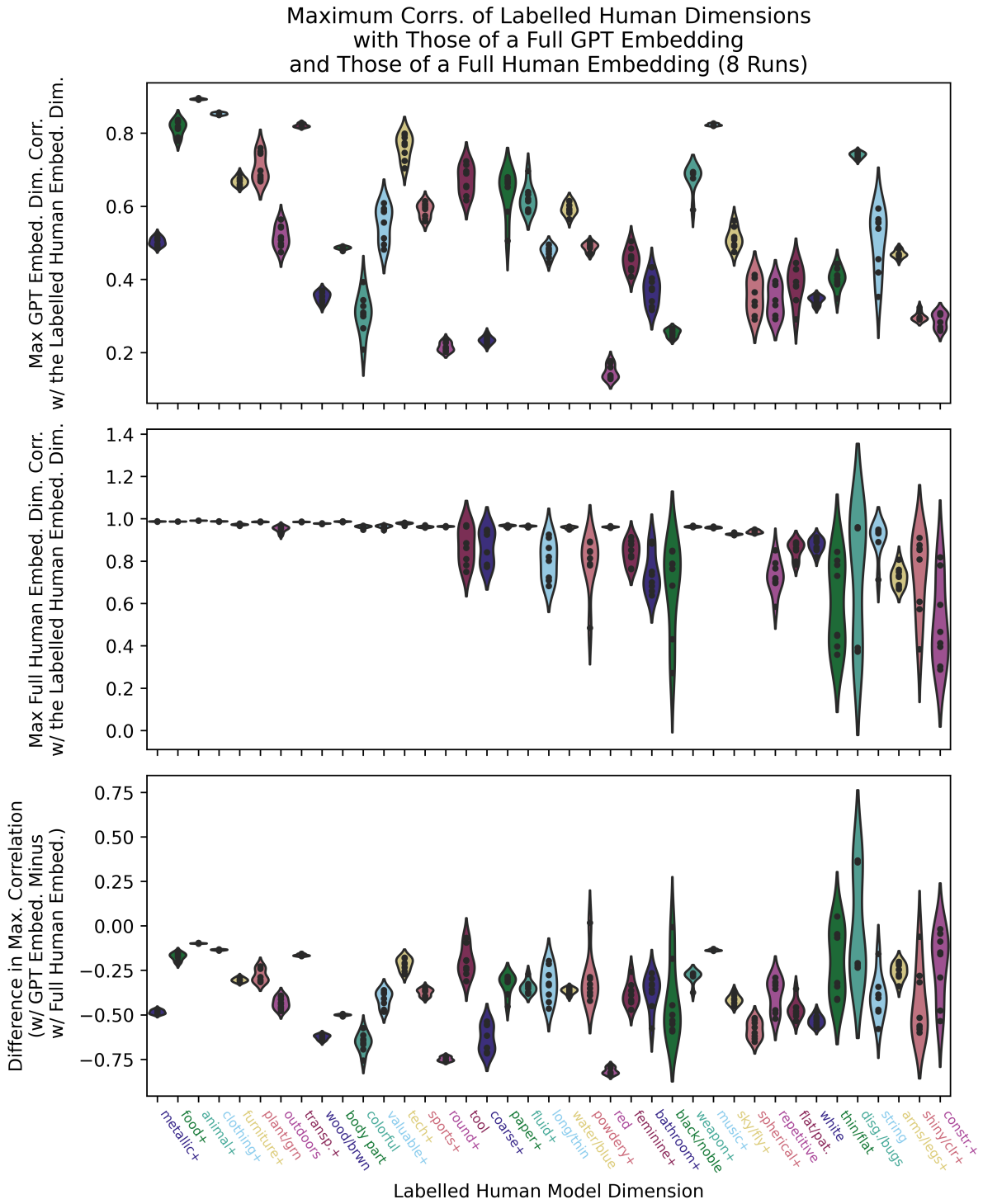
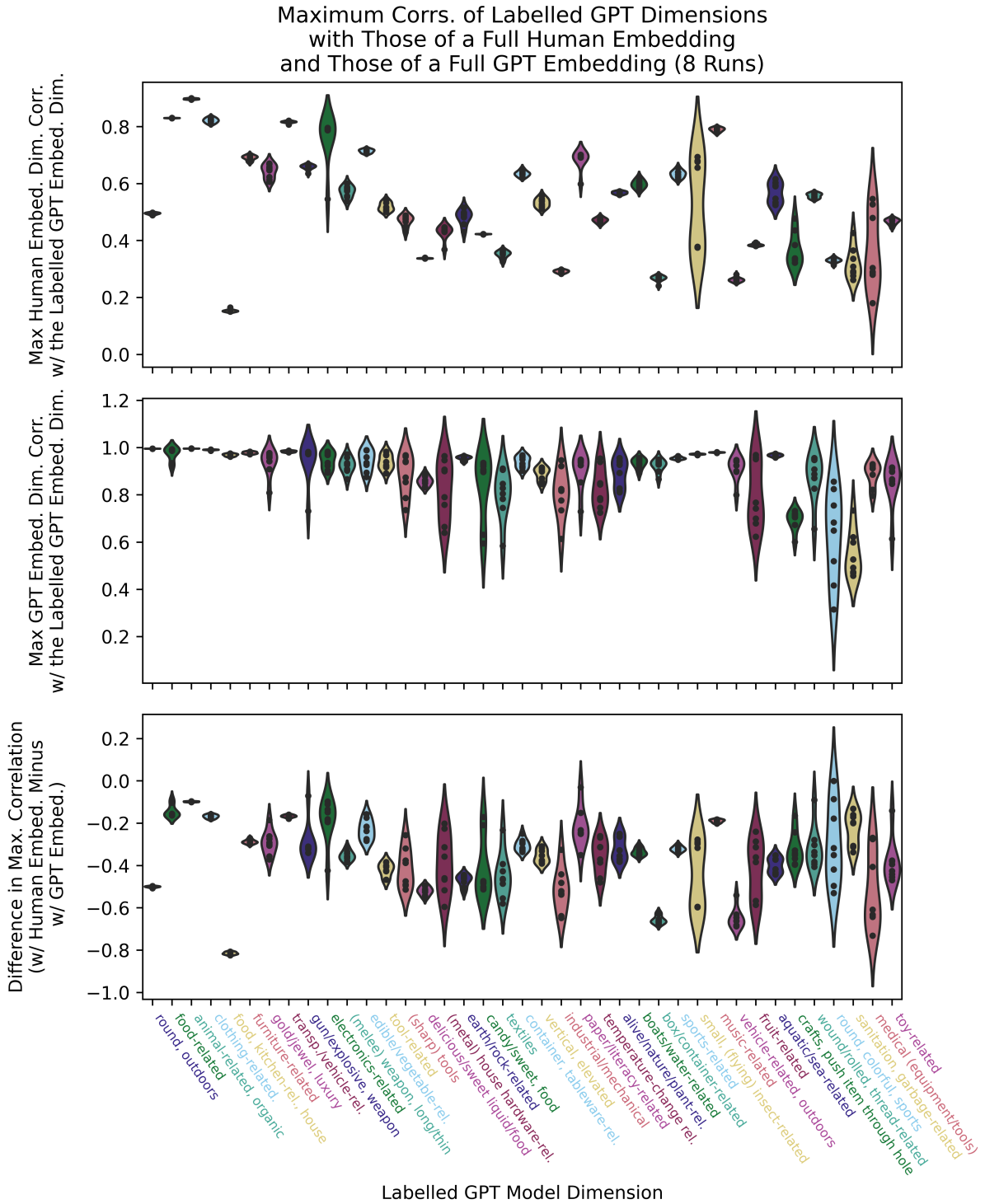


Figure 16: Maximal correlations of the labelled GPT characterizing dimensions with any dimension of a full GPT model and a full human model (over 8 such models of each).



being in clearly distinct clusters (in the case of 'round/outdoors', human dimension 1, it is the only embedding dimension present in the entire right half-plane of the first principal component). Consequently, as the choice of principal components is dominated by these most significant dimensions, the relationships between the rest of the dimensions are less considered.

On the other hand, since UMAP considers the distance between each pair of dimensional vectors when bringing the structure of the projection close to one imposed on the higher-dimensional vectors, the local relationships are better preserved.

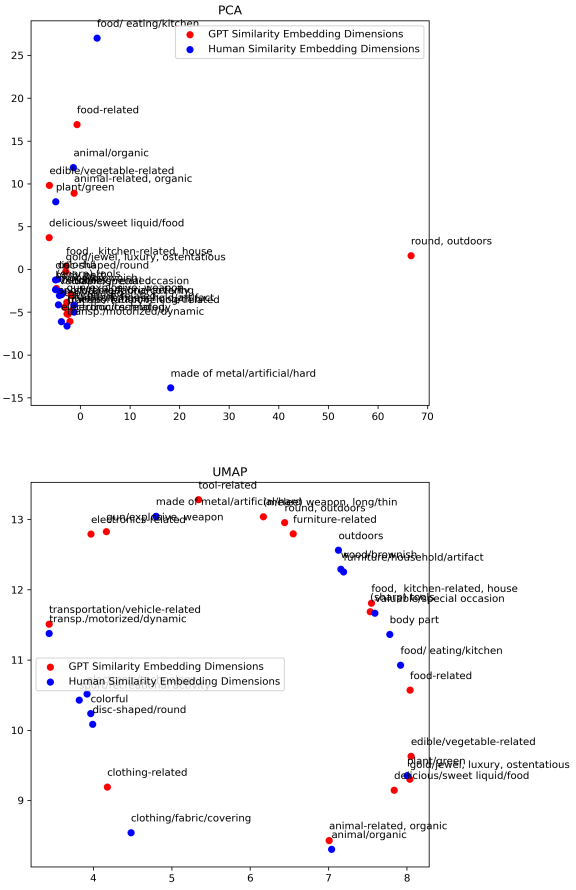


Figure 17: UMAP and PCA performed on the labelled GPT and human embeddings' dimensions.

K Mixed Human–GPT RSA

Table 2 gives the RSA scores used in Figure 4 in tabular form. As such, it holds the average RSA correlations with the baseline human embeddings. The "Dot RSM" column represents using a dot-product kernel to take a representational similarity matrix (RSM) when comparing the various models to the baseline human models, while the "Cos RDM Corr" column represents using cosine similarity to produce representational difference matrices (RDMs) in lieu of those RSMs. For our experiments, we used the dot-product RSMs.

Dataset Type	Lambda	Proportion Human Data	Dot RSM Corr. (RSA Score)	Cos RDM Corr
Random Embedding		0	0	-0.01
Full GPT	0.008	0	0.437	0.438
Partial Human	0.0092	0.125	0.853	0.638
Partial Human	0.0108	0.25	0.897	0.710
Partial Human	0.0128	0.375	0.916	0.752
Partial Human	0.0144	0.5	0.924	0.772
Partial Human	0.0176	0.625	0.930	0.797
Partial Human	0.02	0.75	0.928	0.763
Partial Human	0.024	0.875	0.933	0.808
Mixed	0.0084	0.125	0.507	0.502
Mixed	0.0084	0.25	0.585	0.566
Mixed	0.0084	0.375	0.667	0.613
Mixed	0.0092	0.5	0.750	0.680
Mixed	0.0084	0.625	0.826	0.723
Mixed	0.0088	0.75	0.887	0.774
Mixed	0.0092	0.875	0.926	0.809
Full Human	0.008	1	0.933	0.808
Baseline Human	0.008	1 (separate dataset)	0.978	0.926

Table 2: A table of average RSA scores for different datasets over 4 folds. The lambda values are those produced from our grid-search procedure in Appendix E. For individual folds, see Supplementary Materials.