Towards Measuring Representational Similarity of Large Language Models

Max Klabunde **Mehdi Ben Amor** **Michael Granitzer** **Florian Lemmerich**
University of Passau
`firstname.lastname@uni-passau.de`

Abstract

Understanding the similarity of the numerous released large language models (LLMs) has many uses, e.g., simplifying model selection, detecting illegal model reuse, and advancing our understanding of what makes LLMs perform well. In this work, we measure the similarity of representations of a set of LLMs with 7B parameters. Our results suggest that some LLMs are substantially different from others.We identify challenges of using representational similarity measures that suggest the need of careful study of similarity scores to avoid false conclusions.

1 Introduction

Numerous large language models (LLMs) with remarkable natural language understanding and reasoning capabilities have been released in recent months (Yang et al., 2023; Zhao et al., 2023). However, a comprehensive understanding of the differences between these models beyond architectures and benchmark performances is yet to be established. This is partly due to the inherent challenges in LLMs' explainability given their scale, their high demand for computational resources, and the rising trend of proprietary models.

We argue that a thorough understanding of similarities and differences of LLMs is highly desirable: it may help identify factors that make models perform well, clear up the generalizability of studies of individual LLMs, simplify model selection, enhance our ability to ensure alignment of model behavior with human goals, improve ensembling, benchmark models without labeled data, identify (potentially illegal) model (re)use, and may aid certification of models, which could be required by future regulation of AI.

LLM similarity can be studied from multiple perspectives, including functional similarity, i.e., whether they produce similar outputs, representational similarity, i.e., whether they have similar internal representations, whether they have similar reliance on specific training data, or whether they were trained in a similar manner. Methods for these perspectives have been proposed in prior work, but often focus on non-sequence models (Klabunde et al., 2023) or do not scale to the size of LLMs (Shah et al., 2023). In this work, we focus on representational similarity in the last layer as it implies functional similarity, because the final layer has limited options to diverge functionally. Additionally, it allows studying similarity of *how* outputs are generated.

Similarity of language models was studied to some extent (Wu et al., 2020; Ethayarajh, 2019), but these works do not explore similarity of decoder-only models on the scale of recent LLMs, and instead focus on smaller BERT-style models. As LLMs have developed at break-neck speed, analysis of the similarity of LLMs is generally limited. Moreover, many novel tools to study the similarity of representations have emerged relatively recently.

In this paper, we aim to make the first steps towards understanding similarity of LLMs in more detail:

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1. After discussing several options to compare LLMs, we outline how representational similarity measures can be applied to LLMs (Sec. 2).
2. We present first empirical results regarding the representational similarity of a set of 7B parameter models, offering a preliminary view into LLM similarity for commonsense reasoning (Winogrande) and code generation (HumanEval) (Sec. 3).
3. We identify challenges of gaining a reliable picture of similarity of LLM representations (Sec. 3).

Our code and data is publicly available (see Appendix D).

Related Work. Several works study representations of language models and make implicit comparisons: how contextual they are (Ethayarajh, 2019), what interpretable concepts can be decoded from them (Liu et al., 2019), or how models can communicate via representations (Moschella et al., 2023). Wu et al. (2020); Abnar et al. (2019) explicitly compare representations between models. However, these works have in common that they do not study the current generation of LLMs, and instead focus on smaller models with different architectures like BERT (Devlin et al., 2019) or ELMo (Peters et al., 2018). Similarity of these models was also studied from a functional perspective (McCoy et al., 2020). As an exception, Gurnee et al. (2023) probe the recent Pythia models (Biderman et al., 2023). Concurrent work (Yousefi et al., 2023) studies representations of modern LLMs. Performance of LLMs is compared in many benchmarks (e.g., Srivastava et al., 2023).

2 Comparing Large Language Models

In this section, we discuss different ways to study similarity of LLMs and explain representational similarity in more detail, which we focus on in our experiments.

Multitude of Comparison Approaches. There are multiple different approaches to comparing LLMs and measuring their similarity. Five approaches are functional similarity, representational similarity, weight similarity, similarity of training data attribution, and procedural similarity.

Functional similarity aims to compare the outputs of models (Klabunde et al., 2023). A common approach is to compare performance, where performance similarity equates to model similarity. However, performance alone only gives a partial view of functional similarity: benchmarks represent only a sample of data and models may differ on individual predictions or subgroups of the data, which may impact other functional aspects such as fairness. Hence, more facets of model function need to be compared for a thorough understanding of functional similarity.

Representational similarity compares representations of different layers or models under consideration of their symmetries, i.e., transformations of representations that keep them equivalent such as changing neuron order (Klabunde et al., 2023). Studying *weight similarity* is a related perspective requiring consideration of symmetries (Wang et al., 2022b). These approaches can identify cases where two models can be functionally similar but produce the same output differently.

Training data attribution (e.g., Shah et al., 2023; Grosse et al., 2023) is an approach that identifies the most relevant samples of the training data with respect to influencing the model towards a specific output. From this perspective, models are similar if they have the same relevant training samples for a specific prediction.

Finally, *procedural similarity* is one step removed from the model itself. It takes the perspective that models are similar if the way they are produced is similar. This includes similarity of datasets used for training, hyperparameters, and architecture (e.g., Zhao et al., 2023).

In this work, we focus on one perspective: representational similarity. In the following, we explain how representational similarity measures can be applied to LLMs.

Representational Similarity. Let $f^{(l)}$ be the model that consists of the first l layers of the model f. Then, given N inputs, their D-dimensional representations in layer l are given as $\boldsymbol{R} := \boldsymbol{R}^{(l)} = f^{(l)}(\boldsymbol{X}) \in \mathbb{R}^{N \times D}$, where $f^{(l)}$ is applied row-wise on the inputs $\boldsymbol{X}$. We denote the representations $\boldsymbol{R}$ is compared to as $\boldsymbol{R'} = f'^{(l')}(\boldsymbol{X})$. A representational similarity measure $m(\boldsymbol{R}, \boldsymbol{R'})$ then assigns a similarity (or a dissimilarity) score to two representations. Although representations may not be identical, they might be seen as equivalent, e.g., if $\boldsymbol{R} = -\boldsymbol{R'}$. Hence, a measure respecting these symmetries has certain *invariances*. A measure is invariant to a set of transformations Φ if

$m(\phi(\boldsymbol{R}), \psi(\boldsymbol{R'})) = m(\boldsymbol{R}, \boldsymbol{R'})\ \forall \phi, \psi \in \Phi$. We refer to the survey by Klabunde et al. (2023) for a more detailed introduction.

In this work, we study the representations of the last layer before the final classifier under the invariances to orthogonal transformations (OT), which include rotations and reflections, isotropic scaling (IS), and translation (TR). We justify this selection with the fact that these representations do not have a privileged basis (Elhage et al., 2021), i.e., there is no reason for the basis dimensions (neurons) to have special meaning. This is because the representations could be arbitrarily rotated by applying the same transformation to the weight matrices that add information to the stream. The arbitrary rotation validates OT invariance; similarly, IS invariance is useful. A translation of the representations can affect the outcome of the classification layer for models without a bias, but could also be arbitrarily added by the biases of previous layers. Hence, we study similarity with and without TR invariance. The measures we use for OT and IS invariance experiments are Orthogonal Procrustes (Ding et al., 2021), Aligned Cosine Similarity (Hamilton et al., 2016), the norm of the difference of pairwise similarity matrices (Yin and Shen, 2018), and Jaccard similarity (Schumacher et al., 2021; Wang et al., 2022a). For the experiment with OT, IS, and TR invariance, we additionally use RSA (Kriegeskorte et al., 2008) and CKA (Kornblith et al., 2019). To achieve the desired invariances, we preprocess the representations in some cases by normalizing their scale or centering the columns. Details are in Appendix A.

All of these measures make two assumptions that are generally violated with LLMs: (i) the representation of inputs is deterministic, and (ii) all rows of the representations correspond exactly. With LLMs, the representation of a part of an input sequence depends on representations of earlier parts. These parts are usually sampled if new text is generated, which contradicts (i). Further, differing tokenization can lead to a different number of tokens for the same input text, and thus to a different number of rows in the representation matrix, which violates (ii). We can sidestep these problems. First, we only study the representations of fixed input prompts, which avoids the problem of non-determinism of text generation. Second, we only compare the representations of the final token in the last layer to avoid the issue of differing tokenization. Since these representations are used for the next token prediction, we argue that they have similar meaning across models. Other solutions to this issue based on aligning differing tokenizations were proposed (Liu et al., 2019; Clark et al., 2019), which enable more fine-grained comparisons at increased computational cost.

3 Experiment

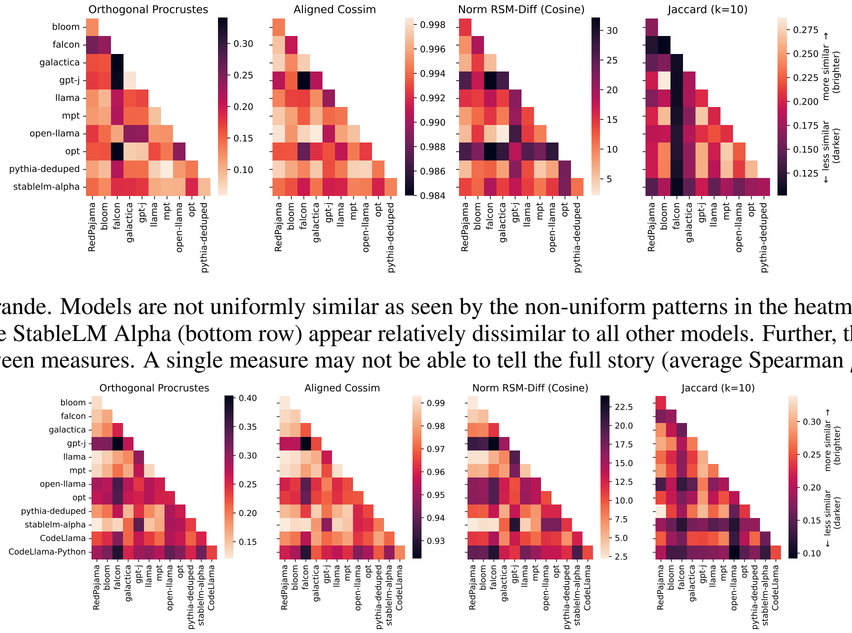
Data. As a first step, we use two datasets from different domains. Winogrande (Sakaguchi et al., 2020) is a benchmark aimed at measuring commonsense reasoning abilities by asking a model to fill in a blank in a sentence with binary options. We use the Winogrande validation set. Further, we use HumanEval (Chen et al., 2021), a code generation benchmark. Here, prompts consist of a comment that describes the functionality of code that should be generated. We always feed data in without additional examples (zero-shot). We show the prompt styles in Appendix B.

Models. We use a set of 11 freely available LLMs with roughly 7B parameters: RedPajama (Together.ai, 2023), Bloom (BigScience Workshop et al., 2023), Falcon (Penedo et al., 2023), Galactica (Taylor et al., 2022), GPT-J (Wang and Komatsuzaki, 2021), Llama (Touvron et al., 2023), MPT (MosaicML, 2023), OpenLlama (Geng and Liu, 2023), OPT (Zhang et al., 2022), Pythia (Biderman et al., 2023), and StableLM Alpha (StabilityAI, 2023). All models are base models without instruction finetunining. For the code data, we add CodeLlama and CodeLlama-Python (Rozière et al., 2023), which are specifically trained on code. Except for Llama, we use the weights published on Hugging Face.

3.1 Results

In Figure 1, we report representational similarity with OT and IS invariance. Results are similar with OT, TR, and IS invariance, which we show in Appendix C due to space constraints.

Significant differences between models. On both datasets, some models have significant differences. On Winogrande, for example, Falcon stands out for a relatively low similarity by Orthogonal Procrustes and Jaccard similarity. Similarly, StableLM Alpha seems to be relatively dissimilar to all other models. On HumanEval, OpenLlama and OPT stand out as dissimilar.



(a) Winogrande. Models are not uniformly similar as seen by the non-uniform patterns in the heatmaps. Some models like StableLM Alpha (bottom row) appear relatively dissimilar to all other models. Further, the patterns differ between measures. A single measure may not be able to tell the full story (average Spearman $\rho = 0.35$).

(b) HumanEval. As not all models have similar amounts of code in their training data, similarity patterns are expectedly different to Winogrande. Of the code-specific models, only the Python variant seems generally dissimilar to the other models. Compared to (a), correlation between measures is higher (average $\rho = 0.65$).

Figure 1: Representational similarity on Winogrande (top) and HumanEval (bottom) with OT and IS invariance. Bright colors show the most similar models, dark colors the most dissimilar ones.

Significant differences between similarity measures. Again looking at Falcon on Winogrande, this model stands out as dissimilar only with two of the four similarity measures. For both Aligned Cosine Similarity and Norm RSM-Diff, differences to other models are less pronounced. For these measures, GPT-J and OPT seem more dissimilar instead. These differences between measures occur despite them sharing the same invariances. It seems that while these measures share the same (high-level) view on representational similarity, finer-grained differences have substantial influence. Although Ding et al. (2021) report a similar result for CKA and Orthogonal Procrustes, our results show this is a more wide-spread issue. Few dimensions with high mean and variance may mask differences in all but these few dimensions for cosine similarity-based measures (Timkey and van Schijndel, 2021).

Similarity is application-dependent. Similarity for one task does not imply similarity for another task: while GPT-J has a similar Orthogonal Procrustes score to the other models on Winogrande, it stands out as dissimilar on HumanEval. This difference is expected to some degree as not all models had similar amounts of code in their training data, but still highlights that data dependency is an important factor when making claims about similarity of LLMs.
Additionally, the measures with discrepancies are not the same across datasets: while Orthogonal Procrustes and Norm RSM-Diff differ on Winogrande, they show highly similar patterns for HumanEval. The average Spearman correlation between heatmaps on Winogrande versus HumanEval is only 0.34.

Difficulty of interpretation. Scores of measures like Orthogonal Procrustes and Norm RSM-Diff that do not have a clear upper bound are difficult to interpret. In the absence of an interpretable scale, it is unclear whether uniform scores imply all models are equally similar or equally dissimilar.

4 Conclusions

We demonstrate measuring representational similarity of LLMs using a set of 7B parameter models. Representations do not seem to be universal, which may limit generality of study of any single LLM, but boost abilities to detect specific models. We identify several challenges of using representational similarity measures for measuring LLM similarity: discrepancies between measures, task-dependency, and interpretation. These challenges provide interesting avenues for future work.

Acknowledgments and Disclosure of Funding

This work is partially supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany. The Federal Office for Agriculture and Food (BLE) provides coordinating support for artificial intelligence (AI) in agriculture as funding organisation, grant number **2820KI012**.

References

Abnar, S., Beinborn, L., Choenni, R., and Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

BigScience Workshop, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., et al. (2023). Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. In Linzen, T., Chrupała, G., Belinkov, Y., and Hupkes, D., editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. (2001). On kernel-target alignment. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ding, F., Denain, J.-S., and Steinhardt, J. (2021). Grounding representation similarity through statistical testing. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1556–1568. Curran Associates, Inc.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Geng, X. and Liu, H. (2023). OpenLLaMA: An Open Reproduction of LLaMA.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. In Jain, S., Simon, H. U., and Tomita, E., editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 63–77, Berlin, Heidelberg. Springer.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., Hubinger, E., Lukošiūtė, K., Nguyen, K., Joseph, N., McCandlish, S., Kaplan, J., and Bowman, S. R. (2023). Studying large language model generalization with influence functions. *ArXiv preprint*, abs/2308.03296.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. (2023). Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint*, abs/2305.01610.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Klabunde, M., Schumacher, T., Strohmaier, M., and Lemmerich, F. (2023). Similarity of Neural Network Models: A Survey of Functional and Representational Measures. *ArXiv preprint*, abs/2305.06329.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

McCoy, R. T., Min, J., and Linzen, T. (2020). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In Alishahi, A., Belinkov, Y., Chrupała, G., Hupkes, D., Pinter, Y., and Sajjad, H., editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

MosaicML (2023). Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolà, E. (2023). Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *ArXiv preprint*, abs/2306.01116.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Ellen, X., Adi, Y., Liu, J., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C. C., Grattafiori, A., Xiong, W., Défossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., and Synnaeve, G. (2023). Code Llama: Open Foundation Models for Code. *arXiv preprint*, abs/2308.12950.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. (2020). Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Schumacher, T., Wolf, H., Ritzert, M., Lemmerich, F., Grohe, M., and Strohmaier, M. (2021). The Effects of Randomness on the Stability of Node Embeddings. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Communications in Computer and Information Science, pages 197–215. Springer International Publishing.

Shah, H., Park, S. M., Ilyas, A., and Madry, A. (2023). ModelDiff: A framework for comparing learning algorithms. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30646–30688. PMLR.

Shahbazi, M., Shirali, A., Aghajan, H., and Nili, H. (2021). Using distance on the riemannian manifold to compare representations in brain and in models. *NeuroImage*, 239:118271.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

StabilityAI (2023). StableLM Alpha.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A Large Language Model for Science. *ArXiv preprint*, abs/2211.09085.

Timkey, W. and van Schijndel, M. (2021). All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Together.ai (2023). RedPajama 7B now available, instruct model outperforms all open 7B models on HELM benchmarks.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv preprint*, abs/2302.13971.

Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Wang, C., Rao, W., Guo, W., Wang, P., Liu, J., and Guan, X. (2022a). Towards understanding the instability of network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):927–941.

Wang, G., Wang, G., Liang, W., and Lai, J. (2022b). Understanding Weight Similarity of Neural Networks via Chain Normalization Rule and Hypothesis-Training-Testing. *ArXiv preprint*, abs/2208.04369.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wu, J., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2020). Similarity analysis of contextual word representation models. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655, Online. Association for Computational Linguistics.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., and Hu, X. (2023). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ArXiv preprint*, abs/2304.13712.

Yin, Z. and Shen, Y. (2018). On the dimensionality of word embedding. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Yousefi, S., Betthauser, L., Hasanbeig, H., Saran, A., Millière, R., and Momennejad, I. (2023). In-Context Learning in Large Language Models: A Neuroscience-inspired Analysis of Representations.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models. *ArXiv preprint*, abs/2205.01068.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A Survey of Large Language Models. *ArXiv preprint*, abs/2303.18223.

A Representational Similarity

Here, we give an overview of the similarity measures we use and their invariances as well as preprocessing that is applied to the representations in some cases.

A.1 Invariances

The following three invariances are relevant for our work:

- Invariance to orthogonal transformation (OT) of a representational similarity measure means that $m(\boldsymbol{R}, \boldsymbol{R'}) = m(\boldsymbol{RQ}, \boldsymbol{R'Q'})$ for any $\boldsymbol{Q}, \boldsymbol{Q'}$ from the orthogonal group $\mathcal{O}(D)$.
- Invariance to isotropic scaling (IS) means $m(\boldsymbol{R}, \boldsymbol{R'}) = m(a\boldsymbol{R}, b\boldsymbol{R'})$ for any $a, b \in \mathbb{R}^+$.
- Invariance to translation (TR) means $m(\boldsymbol{R}, \boldsymbol{R'}) = m(\boldsymbol{R} + \boldsymbol{1}_N \boldsymbol{c}^\mathsf{T}, \boldsymbol{R'} + \boldsymbol{1}_N \boldsymbol{d}^\mathsf{T})$ for any $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^D$ with $\boldsymbol{1}_N$ being a vector of N ones.

A.2 Preprocessing

Preprocessing of representations allows us to augment the natural invariances of a similarity measures by eliminating differences that the measure is not invariant to. Centering columns, i.e., setting the mean of neurons to zero, adds invariance to translation as it removes the same for all representations. The centered representation is defined as $\widetilde{\boldsymbol{R}} = \boldsymbol{HR}$ with $\boldsymbol{H} = \boldsymbol{I}_N - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^\mathsf{T}$. To eliminate scale differences of representations, we can normalize them by their Frobenius norm: $\widetilde{\boldsymbol{R}} = \boldsymbol{R}/\|\boldsymbol{R}\|_F$. This adds IS invariance.

Finally, some similarity measures require that representations have equal dimensionality. In those cases, we zero-pad the representation with lower dimension.

A.3 Representational Similarity Measures

Orthogonal Procrustes. This measure aims to find an orthogonal transformation such that representations are optimally aligned in terms of minimizing the norm of their difference. Formally:

$$m_{\mathrm{OP}}(\boldsymbol{R}, \boldsymbol{R'}) = \min_{\boldsymbol{Q} \in \mathcal{O}(D)} \|\boldsymbol{RQ} - \boldsymbol{R'}\|_F, \tag{1}$$

where $\mathcal{O}(D)$ is the orthogonal group of D-dimensional matrices and $\|\cdot\|_F$ is the Frobenius norm. Orthogonal Procrustes is invariant to orthogonal transformations only, but, for our experiments, we add IS invariance by normalizing the representations. For the experiment with OT, IS, and TR invariance, we first center the representations and then normalize them.

Aligned Cosine Similarity. Aligned Cosine Similarity (Hamilton et al., 2016) is similar to Orthogonal Procrustes in that first an optimal orthogonal alignment of the representation is computed. Then, the cosine similarity $\mathrm{cossim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\mathsf{T}\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}$ is computed between corresponding rows of the representations and aggregated:

$$m_{\mathrm{ACS}}(\boldsymbol{R}, \boldsymbol{R'}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{cossim}((\boldsymbol{RQ}^*)_i, \boldsymbol{R'}_i) \tag{2}$$

with $\boldsymbol{Q}^* = \arg\min_{\boldsymbol{Q} \in \mathcal{O}(D)} \|\boldsymbol{RQ} - \boldsymbol{R'}\|_F$ being the optimal alignment. Aligned Cosine Similarity is invariant to orthogonal transformations and isotropic scaling. For the experiment with OT, IS, and TR invariance, we first center the representations.

Norm RSM-Difference. This measure uses so-called *representational similarity matrices* (RSMs), that capture the pairwise similarities within one representation matrix—different ways to compute those similarities are possible. The RSMs of both representations are then compared by taking the norm of their difference. Formally, let $s : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$ be a similarity (distance) function between vectors such as cosine similarity. The choice of s determines the invariances of this measure: cosine similarity is invariant to isotropic scaling and orthogonal transformation, whereas Euclidean distance is invariant to orthogonal transformations and translations. The matrix of pairwise similarities (the RSM) $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ between the N representations of $\boldsymbol{R}$ is then defined elementwise as $\boldsymbol{S}_{i,j} = s(\boldsymbol{R}_i, \boldsymbol{R}_j)$. The similarity between two representations $\boldsymbol{R}, \boldsymbol{R'}$ is computed as (Shahbazi et al., 2021; Yin and Shen, 2018):

$$m_{\mathrm{Norm}}(\boldsymbol{R}, \boldsymbol{R'}) = \|\boldsymbol{S} - \boldsymbol{S'}\|_F. \tag{3}$$

For the experiment with OT and IS invariance, we use cosine similarity as the similarity function. For the other experiment, we normalize the representations and use Euclidean distance as the similarity function.

Jaccard Similarity. Jaccard Similarity measures the similarity of the nearest neighbors in the representation space (Schumacher et al., 2021; Wang et al., 2022a). Let $\mathcal{N}^k_{\boldsymbol{R}}(i)$ be the set of k nearest neighbors of representation $\boldsymbol{R}_i$ with respect to a similarity function s. Then, representational similarity is computed as follows:

$$m_{\mathrm{Jac}}(\boldsymbol{R}, \boldsymbol{R'}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\left|\mathcal{N}^k_{\boldsymbol{R}}(i) \cap \mathcal{N}^k_{\boldsymbol{R'}}(i)\right|}{\left|\mathcal{N}^k_{\boldsymbol{R}}(i) \cup \mathcal{N}^k_{\boldsymbol{R'}}(i)\right|}. \tag{4}$$

Again, invariances depend on the choice of the similarity function s. We always use cosine similarity. To add TR invariance, we center the representations first.

Representational Similarity Analysis. Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) uses RSMs similar to Norm RSM-Difference. Given two similarity functions, s_{inner} and s_{outer}, first RSMs are computed with s_{inner}. Then the RSMs are compared with s_{outer} by first vectorizing the lower triangle of the RSMs (because they are symmetric) and then applying the similarity function:

$$m_{\mathrm{RSA}}(\boldsymbol{R}, \boldsymbol{R'}) = s_{\mathrm{outer}}(\mathrm{vec}(\boldsymbol{S}), \mathrm{vec}(\boldsymbol{S'})). \tag{5}$$

Again, the invariances are decided by the similarity functions. For the experiment with OT and IS invariance, we use Pearson correlation as the inner similarity function and Spearman correlation as the outer one. For the other experiment, we normalize the representations, then use Euclidean distance as the inner similarity function, and finally use Spearman correlation as the outer similarity function.

Centered Kernel Alignment. Centered Kernel Alignment (CKA) (Cortes et al., 2012; Cristianini et al., 2001; Kornblith et al., 2019) is a measure that uses the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), which can be used to test for statistical independence of two sets of random variables. HSIC operates on RSMs $\boldsymbol{S}, \boldsymbol{S'}$ and is defined as $\mathrm{HSIC}(\boldsymbol{S}, \boldsymbol{S'}) = \frac{1}{(N-1)^2} \operatorname{tr}(\boldsymbol{S}\boldsymbol{H}\boldsymbol{S'}\boldsymbol{H})$ with $\boldsymbol{H} = \boldsymbol{I}_N - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^\mathsf{T}$ a centering matrix. HSIC is divided by a normalization term to achieve IS invariance in addition to OT and TR invariance:

$$m_{\mathrm{CKA}}(\boldsymbol{R}, \boldsymbol{R'}) = \frac{\mathrm{HSIC}(\boldsymbol{S}, \boldsymbol{S'})}{\sqrt{\mathrm{HSIC}(\boldsymbol{S}, \boldsymbol{S})\,\mathrm{HSIC}(\boldsymbol{S'}, \boldsymbol{S'})}}. \tag{6}$$

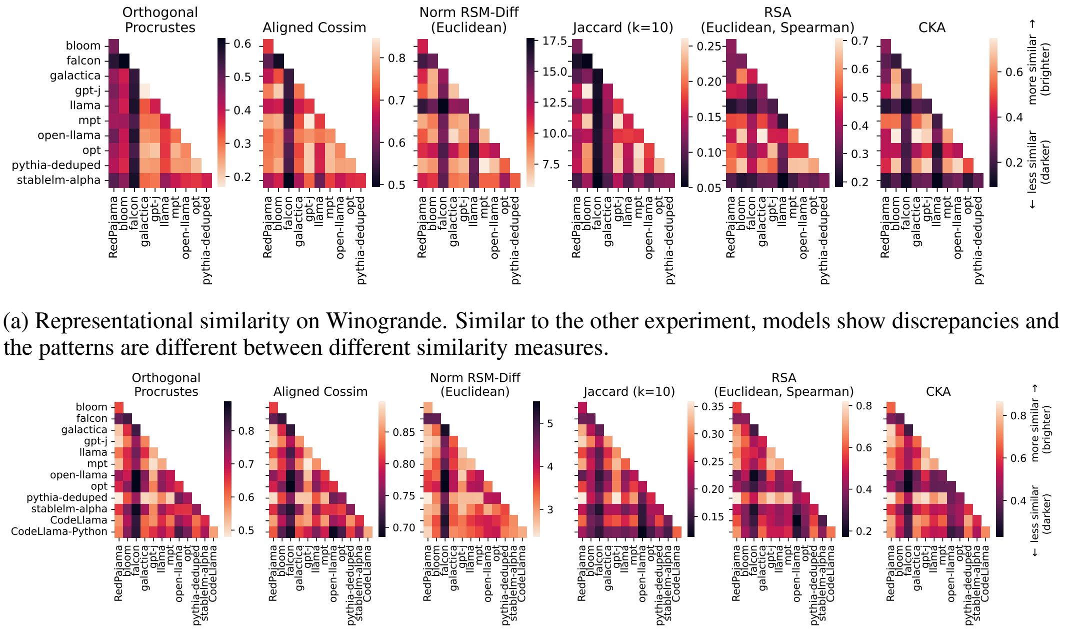
CKA requires that representations are preprocessed to have mean-centered columns. The RSMs are typically computed with the linear kernel, i.e., $\boldsymbol{S} = \boldsymbol{R}\boldsymbol{R}^\mathsf{T}$, although other options are possible. We only use this linear version of CKA.

B Prompt Style

For Winogrande, we prompt the model in the following style:

```
Fill in the _ in the below sentence:
Sentence: Sarah was a much better surgeon than Maria so _ always got the easier cases.
Option 1: Sarah
Option 2: Maria
Does _ in the sentence above refer to Option 1 or 2?
Answer: Option
```

The sentence and the options differ from input to input. For HumanEval, we follow the prompting style used in the original paper for code completion:



(a) Representational similarity on Winogrande. Similar to the other experiment, models show discrepancies and the patterns are different between different similarity measures.

(b) Representational similarity on HumanEval.

Figure 2: Representational similarity on Winogrande (top) and HumanEval (bottom) with OT, IS, and TR invariance. Bright colors show the most similar models, dark colors the most dissimilar ones.

```
def max_element(l: list):
    """Return maximum element in the list.
    >>> max_element([1, 2, 3])
    3
    >>> max_element([5, 3, -5, 2, -3, 3, 9, 0, 123, 1, -10])
    123
    """
```

C Additional Experiment Results

Additional results for OT, IS, and TR invariance are shown in Figure 2.

D Code and Data Availability

Code is available at https://github.com/mklabunde/llm_repsim. Data is available at https://doi.org/10.5281/zenodo.8411089.

E Limitations

Our experiments have several limitations.

First, the datasets we use are limited in size. The number of samples is lower than the number of dimensions, which may allow overfitting of alignment-based measures like Orthogonal Procrustes and Aligned Cosine Similarity. Thus, their scores may be overestimated compared to when applied on larger sets of representations. In future work, additional data—not necessarily from benchmarks—can be used to provide additional evidence.

Second, we only studied models of a specific size. It is unclear, to what extent these patterns generalize to other models.

Further, we use a specific prompting format, which may influence model similarity. Importance of prompts for model outputs was demonstrated in prior work (Kojima et al., 2022; Wei et al., 2022), but has not been studied in detail for representations.