CAT, RAT, MEOW: ON THE ALIGNMENT OF LANGUAGE MODEL AND HUMAN TERM-SIMILARITY JUDGMENTS

Lorenz Linhardt[1,2,*] **Tom Neuhäuser**[1,2] **Lenka Tětková**[3] **Oliver Eberle**[1,2]

ABSTRACT

Small and mid-sized generative language models have gained increasing attention. Their size and availability make them amenable to be analyzed at a behavioral as well as a representational level, allowing investigations of how these levels interact. We evaluate 32 publicly available language models for their representational and behavioral alignment with human similarity judgments on a word triplet task. This provides a novel evaluation setting to probe semantic associations in language beyond common pairwise comparisons. We find that (1) even the representations of small language models can achieve human-level alignment, (2) instruction-tuned model variants can exhibit substantially increased agreement, (3) the pattern of alignment across layers is highly model dependent, and (4) alignment based on models' behavioral responses is highly dependent on model size, matching their representational alignment only for the largest evaluated models.

1 INTRODUCTION

Large language models (LLMs) have recently seen rapid progress, leading to the creation of numerous benchmarks, with great emphasis being placed on *behavioral* evaluations (e.g. (Srivastava et al., 2023; Liu et al., 2023; Liang et al., 2023)). The demand for computational efficiency, accessibility, and privacy has driven the development of smaller language models (Lu et al., 2024), which was made possible by advances in model distillation (Gu et al., 2024), quantization (Lin et al., 2024), and pruning (Wang et al., 2024). Such models, which, unlike their larger counterparts, are often made publicly available, offer an opportunity to investigate the *representations* underlying language model behavior. To understand language model representations and uncover relevant conceptual directions, methods such as representational similarity (Kriegeskorte et al., 2008; Kornblith et al., 2019; Klabunde et al., 2024), probing and sparse autoencoders (Bricken et al., 2023; Cunningham et al., 2023), manifold analysis (Mamou et al., 2020), feature attribution approaches (Eberle et al., 2020; Kauffmann et al., 2022), and function vectors (Todd et al., 2024) have been explored. Studies comparing human signals to model predictions have revealed that models, even those trained on unrelated tasks like self-supervised prediction, show some representational alignment with human data of visual (Yamins et al., 2014; Zhang et al., 2018; Conwell et al., 2024) and language (Abdou et al., 2021; Huh et al., 2024; Goldstein et al., 2024) processing. However, significant differences in robustness, generalization, and alignment (e.g. (Lapuschkin et al., 2019; Geirhos et al., 2020; Momennejad et al., 2023; Muttenthaler et al., 2024b)) between humans and deep models still persist, highlighting the need for a deeper understanding of these discrepancies.

In this work, we take a step towards understanding the structure of the internal representation spaces of language models. In particular, we use a triplet task from cognitive science that probes which words are con-

[1]Machine Learning Group, Technische Universität Berlin, Berlin, 10623, Germany
[2]BIFOLD - Berlin Institute for the Foundations of Learning and Data, Berlin, 10623, Germany
[3]Section for Cognitive Systems, DTU Compute, Technical University of Denmark, Kongens Lyngby, 2800, Denmark
[*]Correspondence to: l.linhardt@tu-berlin.de

sidered more similar than others, and analyze the agreement of human similarity judgments with language model responses. We evaluate both representation similarities across multiple layers of recent models, as well as the models' behavioral (i.e. generative) responses. We seek to answer the following questions:

(Q1) What is the general level of human alignment regarding similarity judgments, and how is it affected by instruction tuning and model size?

(Q2) How does this alignment change across layers? Can it be localized at a particular layer?

(Q3) Does representational alignment correspond to behavioral (generative) alignment?

We find that (1) the representation spaces of even small models are remarkably aligned with human similarity judgments and model size does not appear to be the main factor determining differences in alignment – this stands in contrast to results on vision models, where alignment with human similarity judgments remains limited (Muttenthaler et al., 2023), (2) the pattern of representational alignment with human similarity judgments across the layers differs between models, (3) representations of instruction-tuned models are generally more aligned than their pretrained counterparts, (4) behavioral model evaluations show a clear correlation of model size and alignment – the level of representational alignment is only reached by the largest models considered.

2 Related Work

Textual semantic similarity tasks have widely been used in cognitive science (Tversky, 1977; Nosofsky, 1986; Hebart et al., 2020) and natural language processing (Agirre et al., 2009; Camacho-Collados et al., 2017; Chandrasekaran & Mago, 2021) to assess semantic similarity judgments in text. This required the collection of pair-wise similarity scores assigned by human raters, resulting in various datasets, e.g. (Agirre et al., 2012; Hill et al., 2015; Muennighoff et al., 2022), typically containing few hundreds of samples. These datasets have so far mostly been used to evaluate or improve predictive capabilities of similarity models on pair-wise retrieval tasks (Thakur et al., 2021; Vasileiou & Eberle, 2024; Jiang et al., 2024).

Triplet tasks, instead of defining similarity via absolute pair-wise comparisons, rely on the relative evaluation of an anchor to potential targets (Hebart et al., 2020). Earlier works have proposed the evaluation of vision model representations on triplet tasks (Attarian et al., 2020; Muttenthaler et al., 2023) to assess how well their representation spaces align with human similarity judgments (e.g. using the THINGS dataset (Hebart et al., 2023)). This has led to techniques to align vision models post-hoc (Muttenthaler et al., 2024b;a). Recently, the triplet evaluation methodology was applied to language models (Hrytsyna & Alves, 2024). The authors translated THINGS images to text by captioning techniques, the choice of which had a significant impact on the results. In contrast, we employ the 3TT dataset (Borghesani et al., 2023), ensuring that the human and model responses are recorded for the same type of stimulus (text) and the effect of stimulus-specific context is minimized, e.g. no image background needs to be considered.

3 Extracting Similarity Judgments From Humans and Models

Human Similarity Judgments. Different experimental designs exist to extract human similarity ratings of concepts. As asking for a numerical similarity score for two items suffers from mismatching scales across different human raters (Hebart et al., 2020), triplet-task designs have emerged as an alternative (e.g. (Fukuzawa et al., 1988; Robilotto & Zaidi, 2004; Li et al., 2016; Hebart et al., 2020; Borghesani et al., 2023)). In the design used in this work, given terms A, B, and an anchor C, raters are asked a variation of: *Which of the terms, A or B, is closer in meaning to C?* The resulting choice is binary (scale-free) and allows to gauge *relative* distances of concepts.

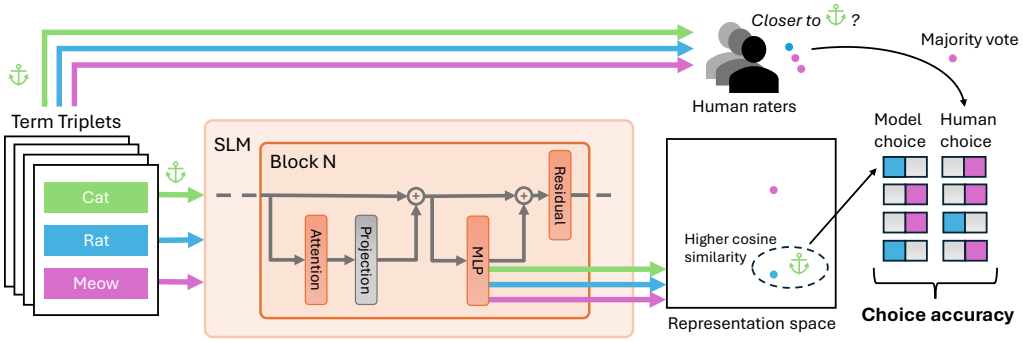

Figure 1: We assess the alignment of language model representations at different layers (attention block, MLP, residual stream) with the human similarity space via a triplet task: Human raters judge which of two terms is more similar to an anchor term. The human choice is compared to the model choice, which is based on representational similarities to the anchor. The fraction of agreeing choices is the "choice accuracy".

We use the three-terms-task (3TT) dataset (Borghesani et al., 2023), containing 10 107 term triplets sampled from 6433 unique words. Of these triplets 2555 have been labeled by a set of 1322 raters, with at least 17 judgments per triplet. A majority response can be calculated for each triplet, as well as an *agreement* score, which is the absolute difference of raters choosing the first and second target term, divided by the total number of judgments. As our primary objective is the comparison of model representations with human judgments, we only use the subset of human-evaluated triplets. In the following, we consider the *majority vote* of all human raters for any triplet as the "human choice". If both choices have an equal number of votes, we omit the triplet from our analysis, leaving $N = 2539$ triplets.

Model Response Extraction. To extract and evaluate model responses based on representations, we follow the pipeline shown in Fig. 1. By "model choice," we denote the term which, when embedded at a given layer, has larger cosine similarity to the anchor term. We consider the fraction of model choices that agree with the human choice as our measure of alignment with the human similarity space, which we name *choice accuracy*. To additionally extract behavioral model responses, we prompt the instruction-tuned models with an adapted version of the prompt used for the creation of the 3TT dataset (Borghesani et al., 2023). Details on the extraction of representational and behavioral responses are summarized in Appx. A.

4 Evaluation of Language Models on the 3TT Dataset

We evaluate a set of 32 language models* from 6 model families (Gemma 2 (Riviere et al., 2024), LLama 3 (Grattafiori et al., 2024), Minitron (Muralidharan et al., 2024), OpenELM (Mehta et al., 2024), Phi (Li et al., 2023; Javaheripi et al., 2023; Abdin et al., 2024), and Qwen 2.5 (Yang et al., 2025)) on the 3TT dataset. 17 of the models are only *pretrained*, and 15 are *instruction tuned* after pretraining.

4.1 (Q1) How Well Aligned are the Representations of Language Models?

To assess whether language models produce similarity choices akin to humans', we investigate: (1) choice accuracy, which serves as a basic indicator of the alignment of representation spaces, and (2) whether the ratio of distances of the two choices to the anchor corresponds to the level of agreement between humans.

*obtained from `www.huggingface.co`

Model	Pretr.	I. T.	Behav.	Invalid
Gemma2-2B	0.77	**0.79**	0.70	0.01
Gemma2-9B	0.77	0.82	**0.83**	0.03
Llama-3.1-8B	0.81	**0.82**	**0.82**	0.00
Llama-3.2-1B	0.78	**0.80**	0.48	0.01
Llama-3.2-3B	0.80	**0.81**	0.69	0.00
Minitron-4B	0.59	**0.78**	0.67	0.07
Minitron-8B	**0.61**	-	-	-
OpenELM-270M	**0.79**	**0.79**	0.00	1.00
OpenELM-450M	**0.82**	0.81	0.00	1.00
OpenELM-1.1B	**0.81**	0.80	0.00	1.00
OpenELM-3B	**0.77**	**0.77**	0.00	1.00
Phi-1.5	**0.79**	-	-	-
Phi-2	**0.79**	-	-	-
Phi-3.5-mini	-	0.77	**0.78**	0.03
Qwen-2.5-0.5B	0.66	**0.80**	0.44	0.04
Qwen-2.5-1.5B	0.67	**0.78**	0.57	0.01
Qwen-2.5-3B	0.67	**0.79**	0.70	0.07
Qwen-2.5-7B	0.66	**0.79**	**0.79**	0.02
Models (mean)	0.79	**0.82**	**0.82**	-
LSN	-	-	**0.74**	-
Humans (mean)	-	-	**0.82**	-

Table 1: Maximum choice accuracy across all layers for (1) representations of pretrained, (2) representations of instruction-tuned (I.T.), and (3) behavior of instruction-tuned models. Bold numbers are the row-wise maximum, underlined is the overall maximum. The last column indicates the fraction of invalid model answers. The mean choice accuracy over human choices is "Humans (mean)" and over all valid model choices is "Models (mean)". A dash indicates inexistent models and evaluations.

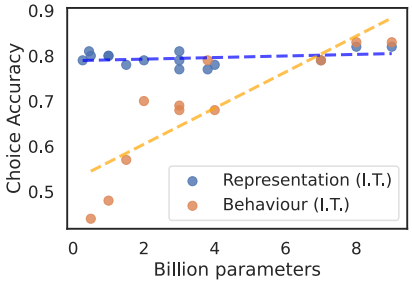

Figure 2: Representational and behavioral choice accuracy v.s. model size for instruction-tuned (I.T.) models. OpenELM models are excluded due to poor instruction following.

Do language models capture human term similarities well? In Tab. 1, we report the choice accuracy of the layer achieving the highest choice accuracy for each model. To contextualize the results, we provide the average human alignment with the majority vote (fraction of human choices agreeing with the human majority vote) as well as the neuro-cognitive inspired model Lancaster Sensorimotor Norms (Lynott et al., 2020) (LSN), both of which can be derived from the 3TT dataset. We find that (1) even the smallest models can be more accurate in predicting term similarities than LSN, which is a strong baseline (Borghesani et al., 2023), and (2) while there are variations across models, their choice accuracies are close to or even reach the average human choice accuracy. We provide examples of triplets for which the model choice consistently (dis)agrees with the human majority in Appx. D. For example, for the anchor *cat* and targets *rat* and *meow*, most humans choose *meow* but all pretrained models pick *rat* as the most similar term.

Does model size matter? Across all models, we find that a higher number of parameters neither consistently positively nor consistently negatively affects representational choice accuracy. Notably, the pretrained model with the highest choice accuracy is OpenELM with 450M parameters. We conclude that for models in the evaluated parameter range, a low number of parameters does not prevent learning representations aligned with human similarity judgments.

What is the impact of instruction tuning? For Gemma 2, Minitron, and Qwen-2.5 models we observe a substantial increase in choice accuracy, with the latter two families showing an increase of 0.1 to 0.2 in choice accuracy. In these cases, instruction tuning appears to have aligned the internal representation spaces with human similarity judgments. For no model does instruction tuning have a significant negative effect on choice accuracy.

Do relative similarities model human agreement? To further investigate the correspondence of models' and humans' similarity judgments, we evaluate whether the relative representational similarity of the two choices to the anchor corresponds to human disagreement. For this purpose, we define a quantity $\gamma := \rho(a, 1-c)$, where ρ is the Pearson correlation coefficient, $a \in [0,1]^N$ is a vector of human agreement scores per triplet (see Sec.3) and $c \in [0,1]^N$ is calculated as the *distance ratio* of the smaller and the larger cosine distance of the targets to the anchor. Intuitively, the distance ratio quantifies how clearly the model prefers one choice over the other.

We find that while there is a strong correlation of choice accuracy and γ ($r = .95$, $p < .001$ for instruction-tuned models and $r = .93$, $p < .001$ for pretrained models), the correlation of agreement and distance ratios is weak ($\gamma < 0.4$). This reveals that the distance ratio of models' representations poorly models human

disagreement and that this mismatch is unaffected by model size. We refer to Appx. C for more detailed results.

4.2 (Q2) How Does Alignment Vary Across Layers?

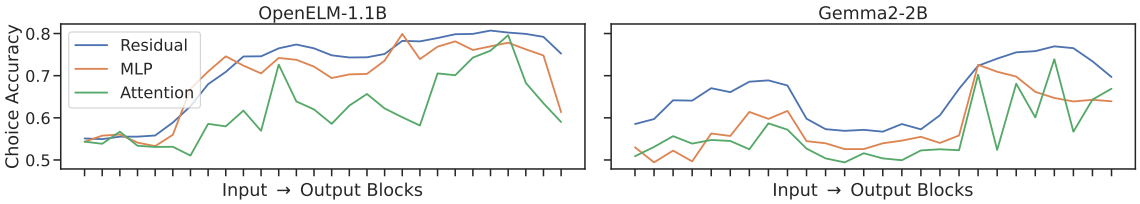

Figure 3: Choice accuracy across layers for two pretrained models. (**Left**) in OpenELM-1.1B, choice accuracy rises nearly monotonically, (**right**) in Gemma2-2B, a bimodal pattern can be observed.

We observe that even though all models follow the same architectural structure (blocks of attention and MLP layers, operating on the residual stream), the pattern of choice accuracy levels across layers differ by model family. Fig. 3 shows one such example: whereas in OpenELM-1.1B, choice accuracy increases rather monotonically until the last third of the model, in Gemma 2-2B, we can see a bimodal choice-accuracy profile. Across all models, choice accuracy appears to increase up until later layers, indicating that the relative arrangement of terms in representation space changes significantly over the layers, even if no context is provided. Instruction tuning appears to modulate the progression of choice accuracy across layers, e.g. in Gemma 2 models, the observed bimodality is considerably flattened. We refer to Appx. B for plots for the remaining models, as well as to Appx. E for additional results on reasoning models. Furthermore, we find that residual stream representations often achieve the models' highest choice accuracies and do not see as strong fluctuations in choice accuracy as attention or MLP layers (see Appx. B). Cases where residual stream layers are superseded in choice accuracy usually see single attention layers achieving the maximum.

4.3 (Q3) Does Representational Alignment Correlate with Behavioral Alignment?

One of the core questions of the representational alignment community is to what extent one can translate between observations of representational structure and behavioral outcomes. While the 3TT dataset only provides behavioral outcomes for human participants, we can observe in models to what extent representational choice accuracy is correlated with behavioral choice accuracy.

Unlike in the evaluation of representational alignment, it can be seen in Tab. 2 that behavioral alignment increases with model size. This pattern is evident in models such as Qwen-2.5-7B, where both metrics are closely aligned. In contrast, smaller models (e.g., Qwen-2.5-0.5B) show poor behavioral alignment, which cannot be attributed only to failures in adhering to the expected output format, indicated by higher rates of invalid answers. OpenELM models are the exception, almost always failing to match the answer format.

Overall, these results suggest that model scale plays a critical role in achieving behavioral alignment and opens the possibility that representational alignment forms an approximate upper bound on behavioral alignment. Furthermore, our results suggest that small language models may contain more knowledge than can be extracted from them in generative evaluations.

5 Discussion and Conclusion

In this work, we found that even small language models can show human-like agreement with the human majority choice on a term similarity task. This is remarkable since the task does not exactly specify what

type of similarity (e.g. lexical, concept feature, concept associative (Borghesani et al., 2023)) is to be used for making choices. We further find that alignment is often positively impacted by instruction tuning. Most studies so far have focused on behavioral alignment (Lampinen et al., 2024; Wang et al., 2023; Chia et al., 2024), with recent evidence suggesting a positive impact of instruction tuning also on representational alignment (Aw et al., 2024). Interestingly, a models' behavioral alignment, unlike representational alignment, is dependent on model size. While the choice accuracy of language models is high, we found that the anchor-target similarity in representation space does not capture human disagreement well.

We believe that extending these analyses to more complex textual data and task-based evaluations can pave the way for the automatic assessment of language models' representational structure, potentially uncovering spurious associations and discrepancies in concept alignment. Furthermore, triplet-term data may be used, similar to work on vision models, to encourage human-like representation structures, making the model more robust and trustworthy.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their constructive feedback. We acknowledge funding by the German Ministry for Education and Research (refs. 01IS18037A and 01IS18025A). This work was supported by the Novo Nordisk Foundation grant NNF22OC0076907 "Cognitive spaces - Next generation explainability".

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9. URL https://aclanthology.org/2021.conll-1.9/.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende (eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1003/.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret (eds.), *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 385–393, Montréal, Canada, 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1051/.

Maria Attarian, Brett D Roads, and Michael C. Mozer. Transforming neural network visual representations to predict human judgments of similarity. In *NeurIPS Workshop on Shared Visual Representations between Humans and Machines*, pp. 1–6, 2020. URL https://openreview.net/forum?id=8wNMPXWK5VX.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns LLMs to the human brain. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=nXNN0x4wbl.

Valentina Borghesani, Jonathan Armoza, Martin N. Hebart, Lune Bellec, and S. M. Brambati. The three terms task - an open benchmark to compare human and artificial semantic representations. *Scientific Data*, 10(1):117, 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02015-3. URL https://doi.org/10.1038/s41597-023-02015-3.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html. Last accessed: 2025-01-20.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 15–26, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2002. URL https://aclanthology.org/S17-2002/.

Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2), February 2021. ISSN 0360-0300. doi: 10.1145/3440755. URL https://doi.org/10.1145/3440755.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. InstructEval: Towards holistic evaluation of instruction-tuned large language models. In Antonio Valerio Miceli-Barone, Fazl Barez, Shay Cohen, Elena Voita, Ulrich Germann, and Michal Lukasik (eds.), *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pp. 35–64, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.scalellm-1.4/.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. doi: 10.1038/s41467-024-53147-y. URL https://doi.org/10.1038/s41467-024-53147-y.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1149–1161, 2020.

Kazuyoshi Fukuzawa, Motonobu Itoh, Sumiko Sasanuma, Tsutomu Suzuki, Yoko Fukusako, and Tohru Masui. Internal representations and the conceptual operation of color in pure alexia with color naming defects. *Brain and Language*, 34(1), 1988. ISSN 10902155. doi: 10.1016/0093-934X(88)90126-5.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A Nastase, Zaid Zada, Eric Ham, Amir Feder, et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature communications*, 15(1):2768, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-00951-3. URL https://doi.org/10.1038/s41562-020-00951-3.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12, 2023. doi: https://doi.org/10.7554/eLife.82580.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL https://aclanthology.org/J15-4004/.

Anastasiia Hrytsyna and Rodrigo Alves. From representation to response: Assessing the alignment of large language models with human judgment patterns. *ACM Trans. Intell. Syst. Technol.*, 2024. ISSN 2157-6904. doi: 10.1145/3709148. URL https://doi.org/10.1145/3709148.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3182–3196, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.181. URL https://aclanthology.org/2024.findings-emnlp.181/.

Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1926–1940, 2022.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. Resi: A comprehensive benchmark for representational similarity measures. *arXiv preprint arXiv:2408.00531*, 2024. URL https://arxiv.org/abs/2408.00531.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 07 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae233. URL https://doi.org/10.1093/pnasnexus/pgae233.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Linjie Li, Vicente Malave, Amanda Song, and Angela J. Yu. Extracting Human Face Similarity Judgments: Pairs or Triplets? In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016*, 2016. doi: 10.1167/16.12.719.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023. URL https://arxiv.org/abs/2309.05463.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2023. URL https://arxiv.org/abs/2211.09110.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. In P Gibbons, G Pekhimenko, and C De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper{_}files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper{_}files/paper/2023/file/43e9d647ccd3e4b7b5baab53f0368686-Paper-Conference.pdf.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52(3):1271–1291, 2020. ISSN 1554-3528. doi: 10.3758/s13428-019-01316-z. URL https://doi.org/10.3758/s13428-019-01316-z.

Jonathan Mamou, Hang Le, Miguel A. Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and Sueyeon Chung. Emergence of separable manifolds in deep language representations. In Hal Daume and Aarti Singh (eds.), *37th International Conference on Machine Learning, ICML 2020*, pp. 6669–6679. International Machine Learning Society (IMLS), 2020.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. OpenELM: An Efficient Language Model Family with Open Training and Inference Framework. *arXiv.org*, April 2024. URL https://arxiv.org/abs/2404.14619v1.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VtkGvGcGe3.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:252907685.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *arXiv preprint arXiv:2407.14679*, 2024. URL https://arxiv.org/abs/2407.14679.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. *International Conference on Learning Representations*, 11, 2023.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024a. URL https://arxiv.org/abs/2409.06509.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems*, 36, 2024b.

Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2409.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Rocco Robilotto and Qasim Zaidi. Limits of lightness identification for real objects under natural viewing conditions. *Journal of Vision*, 4(9), 2004. ISSN 15347362. doi: 10.1167/4.9.9.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Alexandros Vasileiou and Oliver Eberle. Explaining text similarity in transformer models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7859–7873, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.435. URL https://aclanthology.org/2024.naacl-long.435/.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*, 2024.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

A Details to Model Response Extraction

In this section, we provide additional details on how model responses were extracted to calculate choice accuracy from both representational and behavioral responses.

A.1 Representational responses

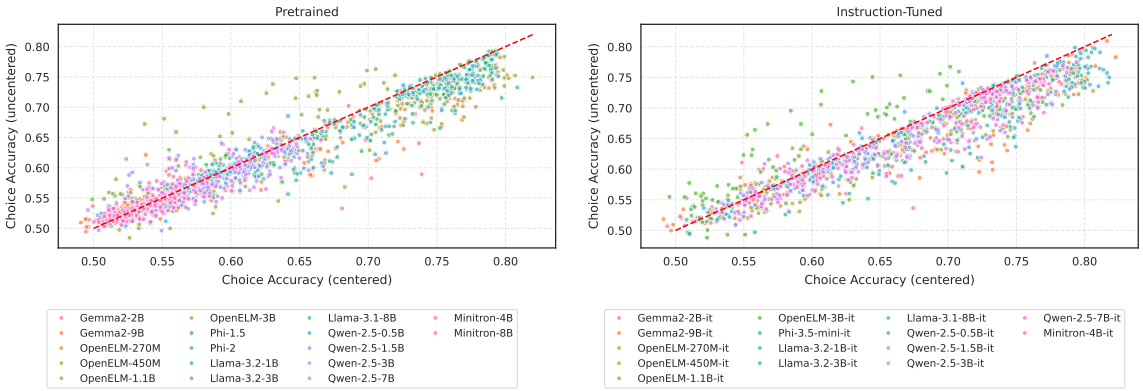

Figure 4: Choice-accuracy before and after representation centering for pretrained (**left**) and instruction tuned (**right**) models. In most cases, centering leads to higher choice accuracy (below the diagonal).

For extracting representations from pretrained (not instruction-tuned) models, we embed each term individually. Special `<bos>` tokens are prepended if required by the model. For each word, we record the representation after each sub-block (attention, MLP, residual stream). Instruction-tuned model variants are fed the individual terms in the `user` part of the corresponding chat template and empty `system` prompts. In both cases, the last token corresponding to the input term is recorded.

All representations used for calculating representational alignment are centered per layer (i.e. the mean over the layer-specific representations of all terms is subtracted). As we use cosine similarity as a basis of the three-terms-task evaluation, moving the origin of the representation space can significantly impact the results. It can be seen in Fig.4 that in most cases, centering leads to small improvements in choice accuracy.

A.2 Behavioral responses

To extract behavioral responses from language models for the three terms task, we prompt the instruction-tuned models with an adapted version of the instructions for human raters used for the creation of the 3TT dataset (Borghesani et al., 2023). The adapted prompt was designed to reduce the rate of invalid answers: *"Which of the words A or B is closer in meaning with the word C? Answer with exactly one word: either A or B. Do not answer with C. Do not answer in a full sentence."* We post-process the models' responses by removing special characters and transforming them to lowercase. To compute the choice accuracy, we determine whether the post-processed response equals *A* or *B* or neither, which we count as invalid choice. To preclude potential bias introduced by the ordering of the presented choices, we randomize their order and report the average choice accuracy over 3 seeds. In Appx. F, we provide additional analyses on the robustness of the prompt concerning term order and instruction complexity.

B Choice Accuracy Across Layers

In this section, we report the choice accuracy of the representations obtained from the attention layer, MLP layer, and residual stream of every block of every evaluated pretrained model. It can be seen in Fig.6 and Fig.7 that qualitatively, the choice-accuracy dynamics over layers varies across models.

To quantify the observation of relatively high smoothness of residual stream layers, we calculate the total variation across layers for every layer type separately:

$$\mathrm{tv} = \sum_{l=2}^{L} |\mathrm{ca}_l - \mathrm{ca}_{l-1}| \tag{1}$$

Here ca_l is the choices accuracy at layer l of L layers of the same layer type. The aggregated results over all models in Fig. 5 confirm that residual stream layers are the most consistent in their similarity structure, whereas attention layers are the most volatile.

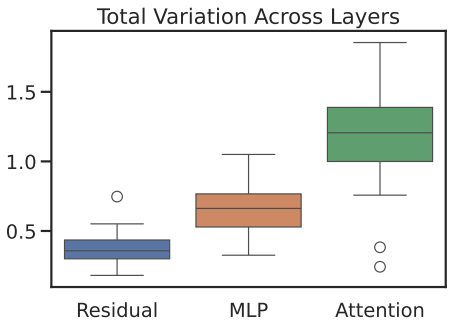

Figure 5: Total variation of choice accuracy over all layers of a particular type. Each box aggregates all pretrained models.

C Additional Analysis on γ

In this section, we report results for the experiment on γ, correlating human agreement and $1-$ distance ratio, from Sec. 4.1 of the main text. We use Spearman and Pearson correlation coefficients, as well as pretrained and instruction-tuned models.

As can be seen in Fig. 8, the trend that models showing a higher choice accuracy also achieve higher γ is preserved when using Spearman correlation. For pretrained models, results are mostly comparable to the ones obtained on instruction-tuned models, with the only notable difference being a set of layers underperforming in γ, relative to their choice accuracy. This set is mainly comprised of the residual stream layer of Qwen and Minitron models – the two model families seeing the largest improvement by instruction tuning.

D Examples of Triplets

Tab. 2 shows examples of triplets where the maximum-choice-accuracy layers of all pretrained models consistently agree or disagree with the human majority. All of these triplets have been selected for high inter-human agreement and are sorted by this agreement.

While we cannot draw strong conclusions from this limited set of examples alone, it seems possible that human similarity judgments are at times based on association (*cat* and *meow*, *dryer* and *lint*, *surfboard* and *pier*) whereas language models' similarity judgments are more often based on type (*cat* and *rat*, *dryer* and *dishwasher*, *surfboard* and *toothbrush*). Future work should include larger, quantitative evaluations to assess the difference in how humans and language models construct similarity judgments.

Tab. 3 and Tab. 4 show examples for representational and behavioral choices of instruction-tuned models. We note some overlap between the pretrained and instruction-tuned representational examples for which models disagree with the human majority.

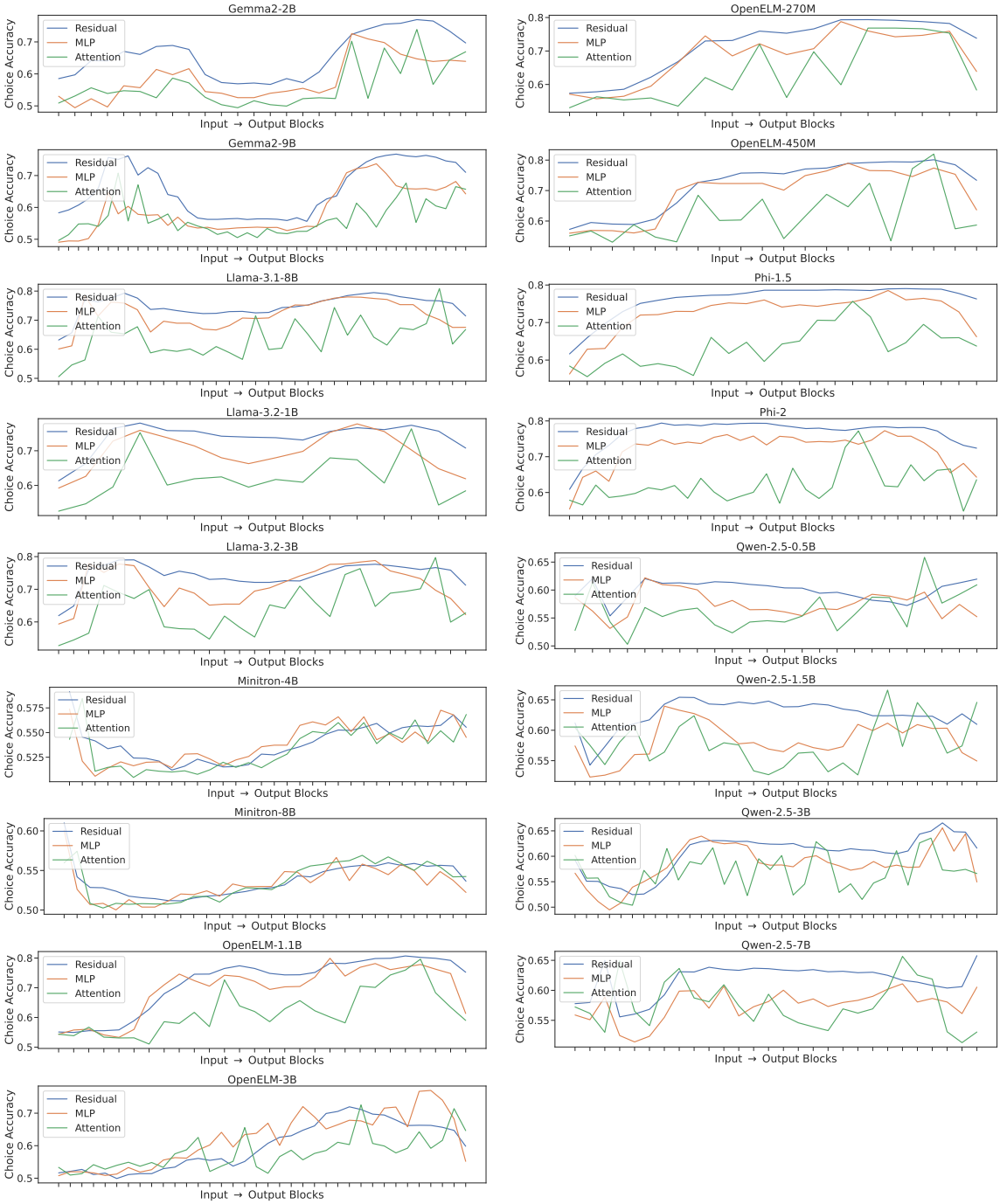

Figure 6: Development of choice accuracy over layers for all pretrained models.

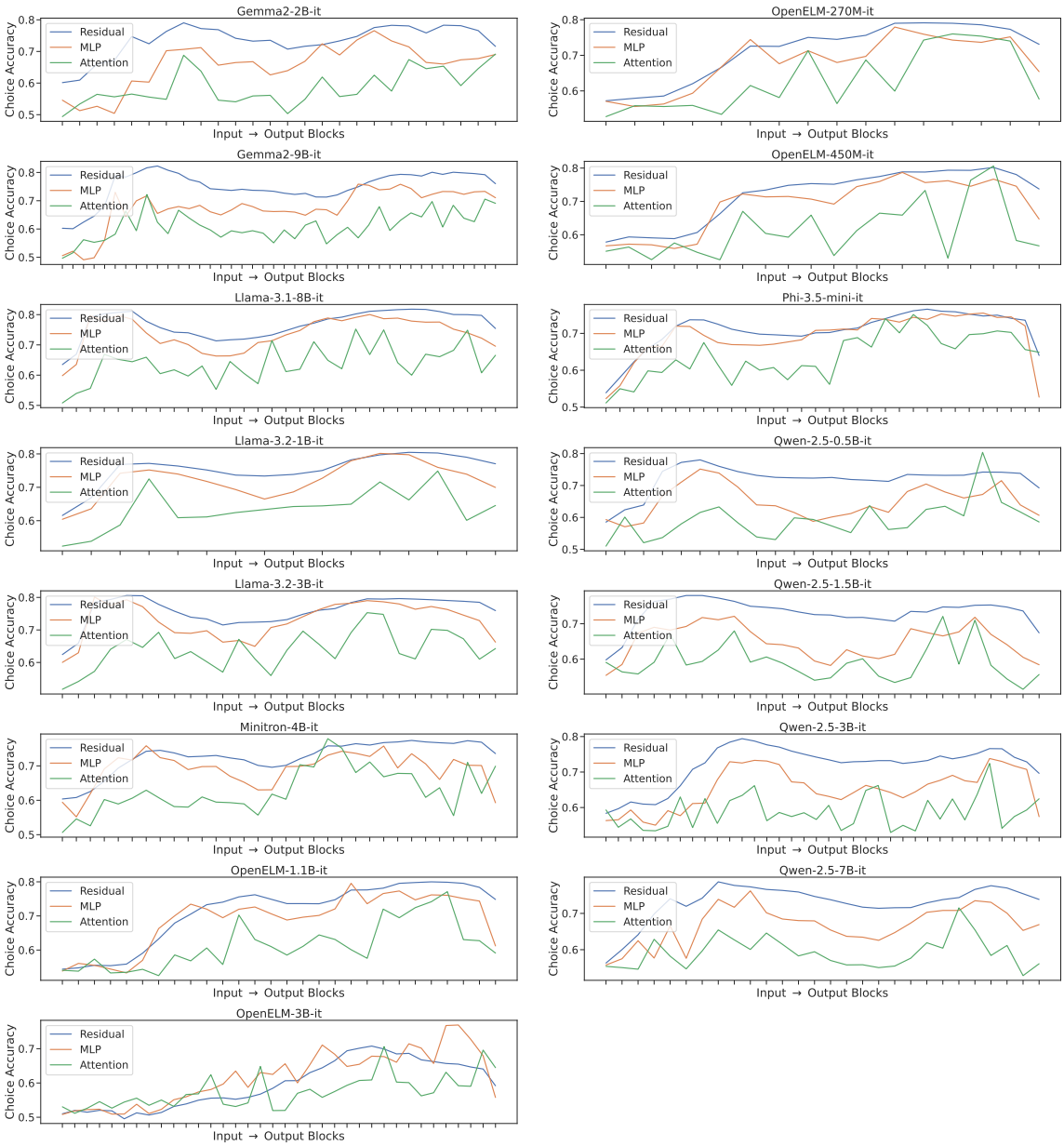

Figure 7: Development of choice accuracy over layers for all instruction-tuned models.

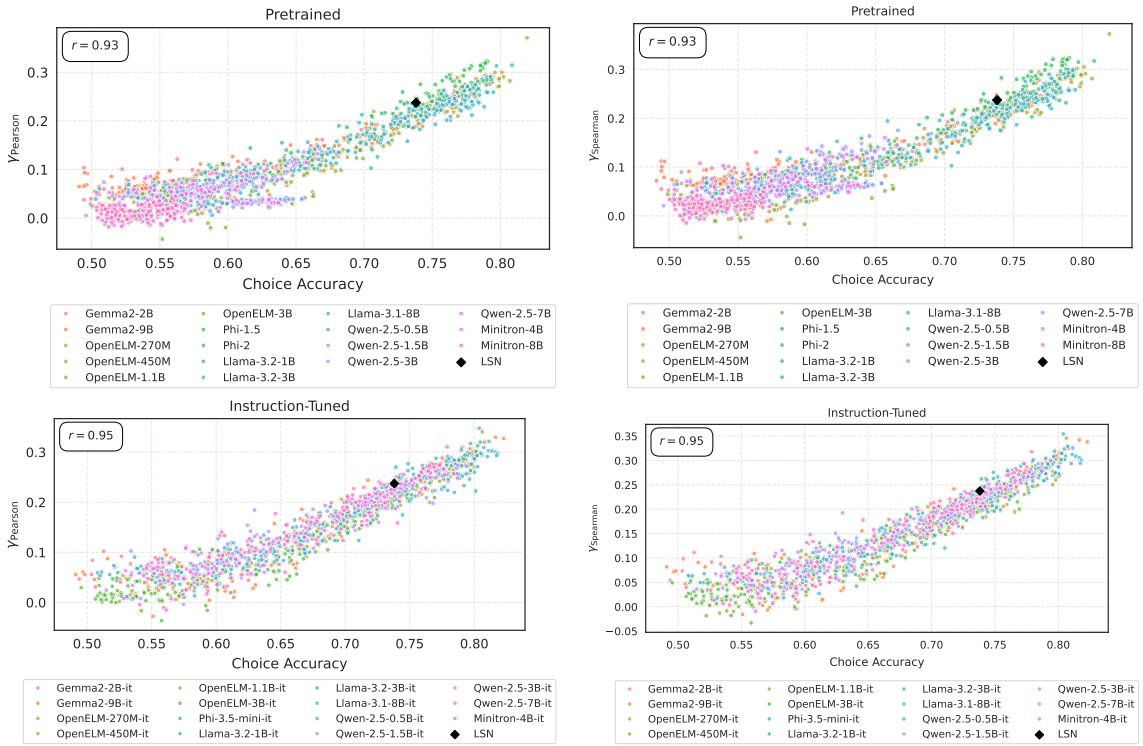

Figure 8: γ v.s. choice-accuracy for pretrained (**top**) and instruction-tuned (**bottom**) models, as well as for computations of γ based on Pearson (**left**) and Spearman (**right**) correlation coefficients. The r-value at the top left of each plot provides the Pearson correlation coefficient of the values on the x-axis and the y-axis.

Anchor	Target 1	Target 2
waterwheel	**watermill**	fingerlike
touchpad	**mousepad**	midfield
fruit	**pear**	merchandise
truck	**tanker**	upholstery
clay	**sandstone**	javelin
chocolate	**butterscotch**	tourist
stroller	**pushchair**	harbor
telescope	**binoculars**	university
lollipop	**lollypop**	officer
hand	**finger**	face

Anchor	Target 1	Target 2
baseball	**outfielder**	cricket
jam	**blackberry**	jump
root	**woody**	proxy
surfboard	**pier**	toothbrush
curb	spur	**floorboard**
screw	**prosthetics**	buck
cat	rat	**meow**
dryer	**lint**	dishwasher
guacamole	margarita	**chip**
comb	**flyer**	wort

Table 2: Examples of triplets with low and high human-model agreement for *pretrained* models. (**Left**) all models chose the same target as the human majority, (**right**) all models chose the other target. Inter-human agreement is high in all examples. Terms in **bold** were chosen by the human majority vote.

Anchor	Target 1	Target 2
waterwheel	**watermill**	fingerlike
fruit	**apple**	spice
door	**gate**	desk
money	**cash**	planetarium
pepperoni	**pizza**	scrubber
pocket	**bag**	panhandle
sailboat	**powerboat**	grandmother
prism	**refraction**	mentality
projector	**auditorium**	collie
scarf	**hairnet**	land

Anchor	Target 1	Target 2
baseball	**outfielder**	cricket
jam	**blackberry**	jump
box	picture	**tupperware**
breakfast	**shaker**	gratitude
filter	speed	**rainwater**
train	training	**railroad**
goldfish	**pond**	hamster
surfboard	**pier**	toothbrush
screw	**prosthetics**	buck
hanger	hangar	**garter**

Table 3: Examples of triplets with low and high human-model agreement for *instruction-tuned* models. (**Left**) all models chose the same target as the human majority, (**right**) all models chose the other target. Inter-human agreement is high in all examples. Terms in **bold** were chosen by the human majority vote.

Anchor	Target 1	Target 2
robe	**turban**	coliseum
guardrail	**roadway**	jabbed
prism	**refraction**	mentality
surfboard	**paddle**	hairstylist
mustache	**beard**	catsup
mustache	**beard**	carburetor
jellyfish	passage	**plankton**
headband	**headpiece**	ritalin
crystal	gut	**sapphire**
dress	**apparel**	pee

Anchor	Target 1	Target 2
thumbtack	**corkboard**	prong
avocado	**macadamia**	mayo
noodle	**popsicle**	teahouse
pigeon	**kingfisher**	courier
meatloaf	pantry	**pumpernickel**
kite	**paraglider**	string
swing	**paddle**	jazz
mixer	encoder	**sealer**
retainer	solicitor	**spacer**
bike	**backpack**	jockey

Table 4: Examples of triplets with low and high human-model agreement for *behavioral responses of instruction-tuned* models. (**Left**) all models chose the same target as the human majority, (**right**) all but one models chose the other target. Inter-human agreement is high in all examples. Terms in **bold** were chosen by the human majority vote.

E Results for Distilled DeepSeek-R1

In this section, we extend our basic choice accuracy evaluation to two distilled versions of DeepSeek-R1 (Guo et al., 2025), that use Qwen-Math-1.5B[†] and Qwen-Math-7B[‡] as base models.

For behavioral response extraction, we allowed the models to generate up to 2000 tokens, and then post-processed the output following the reasoning as detailed in Appx. A. We observe that for some triplets, the models appear to get stuck in a reasoning loop without making a choice within the allotted token budget. We count these cases as invalid answers, resulting in an invalid answer fraction of 0.22 for the 1.5B model and 0.07 for the 7B model. We note that this fraction may be reduced by tuning the temperature parameter.

In Fig. 9, it can be seen that the representational choice accuracy is comparably low for both models (0.64 and 0.71). The 8B variant achieves a significantly higher behavioral choice accuracy of 0.80. We speculate that this discrepancy of representational and behavioral performance may stem from a reduced need for easily decodable representations, as this decoding can be done by the model over lengthy reasoning chains. We note that further analyses are warranted, as in these two models, the reasoning component is confounded with the math-focused training of the base model.

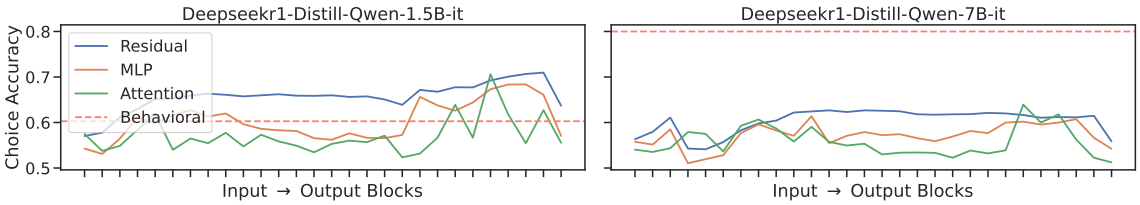

Figure 9: Choice accuracy across layers for two distilled DeepSeek-R1 variants.

F Prompt Robustness Analysis

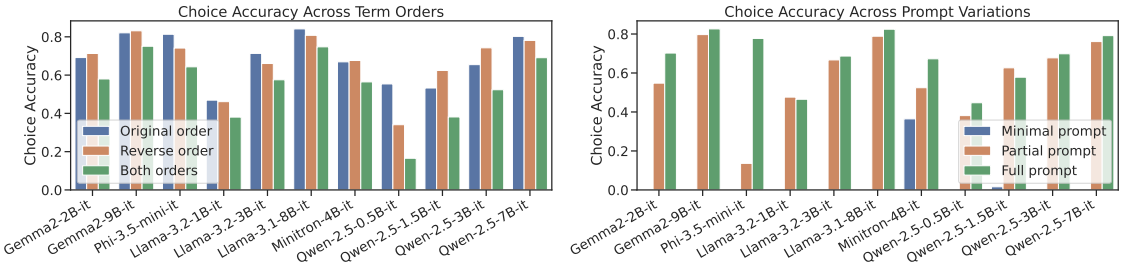

Figure 10: (**Left**) Behavioral choice accuracy for the original target term order in the 3TT datasets, reversed term order, and when only counting triplets as correct if they have been answered correctly in both orders. (**Right**) Behavioral choice accuracy for three variations of the prompt. Only instruction-tuned models are shown. OpenELM models are excluded due to poor instruction following.

[†]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
[‡]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

To evaluate the robustness of our results w.r.t prompt design, we provide two additional experiments investigating (1) the effect of the order in which the two target terms are presented within the prompt, and (2) the effect of varying the prompt formulation.

In Fig. 10 (left), it can be seen that the order in which the target terms are presented can impact choice accuracy. Yet, this impact is only large for the smaller Qwen models. It should be noted that due to the dataset construction, the first term, according to the original order, has a higher chance of being more similar to the anchor. This results in the human choice being the first term in 70% of the evaluated triplets. This explains why term order can influence choice accuracy and supports our strategy of evaluating both orders.

In Fig. 10 (right) we compare three variations of the prompt: The **minimal** prompt reads *"Which of the words A or B is closer in meaning with the word C?"*, the **partial** prompt additionally ends with *"Answer with exactly one word: either A or B."*, and the **full** prompt adds to the partial prompt *"Do not answer with C. Do not answer in a full sentence."* The minimal prompt was used for gathering the human responses in the 3TT dataset, whereas the full prompt was used in the main part of this paper. It can be seen that the relative ordering of the models' choice accuracies does not greatly change across the latter two variations, indicating some robustness to the prompt formulation. Here, the exception is Phi-3.5-mini, which only performs well on the full prompt. Overall, the shorter variations of the prompt show reduced choice accuracy (around 0 for most models in the minimal formulation). We attribute this to answers being incompatible with our answer evaluation schema (see Appx. A), and then being counted as invalid answers.