Instruction-tuned LLMs with World Knowledge are More Aligned to the Human Brain

Khai Loong Aw, Syrielle Montariol*, Badr AlKhamissi*, Martin Schrimpf†, Antoine Bosselut†
EPFL
{khai.aw,syrielle.montariol,badr.alkhamissi,martin.schrimpf,antoine.bosselut}@epfl.ch

Abstract

Instruction-tuning is a widely adopted method of finetuning that enables large language models (LLMs) to generate output that more closely resembles human responses to natural language queries, in many cases leading to human-level performance on diverse testbeds. However, it remains unclear whether instruction-tuning truly makes LLMs more similar to how humans process language. We investigate the effect of instruction-tuning on *brain alignment*, the similarity of LLM internal representations to neural activity in the human language system. We assess 25 vanilla and instruction-tuned LLMs across three datasets involving humans reading naturalistic stories and sentences, and discover that instruction-tuning generally enhances brain alignment by an average of 6%. To identify the factors underlying LLM-brain alignment, we compute the correlation between the brain alignment of LLMs and various model properties, such as model size, performance ability on problem-solving benchmarks, and ability on benchmarks requiring world knowledge spanning various domains. Notably, we find a strong positive correlation between brain alignment and model size (r = 0.95), as well as performance on tasks requiring world knowledge (r = 0.81). Our results demonstrate that instruction-tuning LLMs improves both world knowledge representations and human brain alignment, suggesting that mechanisms that encode world knowledge in LLMs also improve representational alignment to the human brain.

1 Introduction

Instruction-tuning is a widely adopted method for finetuning large language models (LLMs) on datasets containing task-specific instructions. This approach enhances their ability to generalize effectively to previously unseen tasks by learning to follow provided instructions [25]. Instruction-tuning often costs only a small fraction of compute relative to pretraining [5], yet propels pretrained LLMs to incredible performance leaps on reasoning and problem-solving benchmarks. This transformation has enabled LLMs to approach human performance on many tasks, despite using only few (or zero) training examples, and tackle open-world reasoning tasks previously only achievable by humans [27].

In addition to teaching LLMs to understand and follow human instructions, instruction-tuning also improves the ability of LLMs to mimic the ground-truth outputs (often human-written) of the training data. This property allows them to produce more controllable and predictable output that is deemed (1) more desirable by human evaluators on various metrics [27, 5, 24], (2) more aligned to human values [4], and (3) more stylistically similar to human outputs [6, 15].

Consequently, instruction-tuning yields LLMs that are more similar to humans in both capability and output similarity. From a neuroscience perspective, these observations beg the question: **Does**

*Equal contribution
†Equal supervision / senior authors

Preprint.

instruction-tuning make LLMs more similar to the human language system? Previous work has shown that models with high performance on next-word prediction tasks are well-aligned to the human language system [16, 8, 3], and, on some datasets, even hit the estimated noise ceiling.[3] However, there has been no similar study on how instruction-tuning, the training method that enabled powerful LLMs such as ChatGPT, affects alignment to the human language system.

In this work, we explore the impact of instruction-tuning on *brain alignment*, how closely LLMs' internal representations match neural activity in the human language system. Both LLMs and human participants are presented with the same language stimuli comprised of naturalistic stories and sentences. For LLMs, we analyze their internal representations, while for humans, we use previously collected brain activity data from functional magnetic resonance imaging (fMRI) experiments.

To measure brain alignment, we use the Brain-Score [17] linear predictivity metric, assessing how well LLM representations predict human brain activity in response to the same language stimuli [10, 21, 16, 11], using three neural datasets: PEREIRA2018 [12], BLANK2014 [2], and WEHBE2014 [26]. As models vary in brain alignment across different architectures and training objectives [16], we estimate the effect of instruction-tuning across 17 instruction-tuned and 8 vanilla LLMs, and report a significant increase in brain alignment by instruction-tuned LLMs compared to vanilla ones.

To investigate *why* instruction-tuning increases alignment to human brain activity, we then estimate the contribution of various LLM properties towards brain alignment. Specifically, we compute Pearson correlations between an LLM's brain alignment and its properties, including next-word prediction (NWP) ability, model size, a range of problem-solving abilities, and world knowledge spanning different domains. The latter two properties are evaluated with the Big-Bench Hard benchmark (BBH) [19] and the Massive Multi-task Language Understanding benchmark (MMLU) [9], respectively.

We report two major findings:

1. Instruction-tuning generally improves the alignment of LLM representations to brain activity, increasing by 6.2% on average for the LLMs and neural datasets we tested (Figure 1).
2. Investigating the factors underlying LLM-brain alignment, we find that world knowledge and model size are strongly correlated with brain alignment (r = 0.81 and 0.95 for instruction-tuned models, respectively; Figure 2).

2 Language Models

We evaluate the brain alignment of 25 large language models (LLMs) from two model families: T5 [13] and LLaMa [22]. T5 models are encoder-decoder LLMs pre-trained on the Colossal Common Crawl Corpus (C4), a corpus of 356 billion tokens, using a masked infilling objective, and then further finetuned on multi-task mixture of unsupervised and supervised tasks converted into a text-to-text format. We use all five T5 models with sizes between 77M to 11B parameters. LLaMA models [22] are decoder-only LLMs trained on 1.6 trillion tokens from a mixture of corpora including C4, English CommonCrawl, Wikipedia, Github, and more. For LLaMA, we use the 7B, 13B, and 33B parameter versions in our study.

3 Brain Alignment

Brain alignment refers to the method of evaluating the representational similarity between LLMs and human brain activity (Figure 1). This assessment relies on fMRI recordings obtained from human subjects while they read specific language stimuli on potentially any topic [here: 12, 2, 26]. In brain alignment studies, these same language stimuli from prior brain recordings are provided as input to LLMs, whose intermediate layer activations are recorded to extract model representations of the language stimuli. To study the alignment of LLM and human data, we follow a general approach previously used in several works [17, 16, 10, 21, 11, 1]. Specifically, we use the linear predictivity metric implemented in Brain-Score [18, Figure 1], first training a linear function to predict fMRI voxels associated with the human language system using LLM representations as input features. We then apply this linear function to held-out brain activity data from the original corpus of recordings,

[3] In fMRI recordings, an upper limit of representation similarity can be computed by sampling from the same patient twice, deducing a threshold defined by the noise level of the data gathering process.

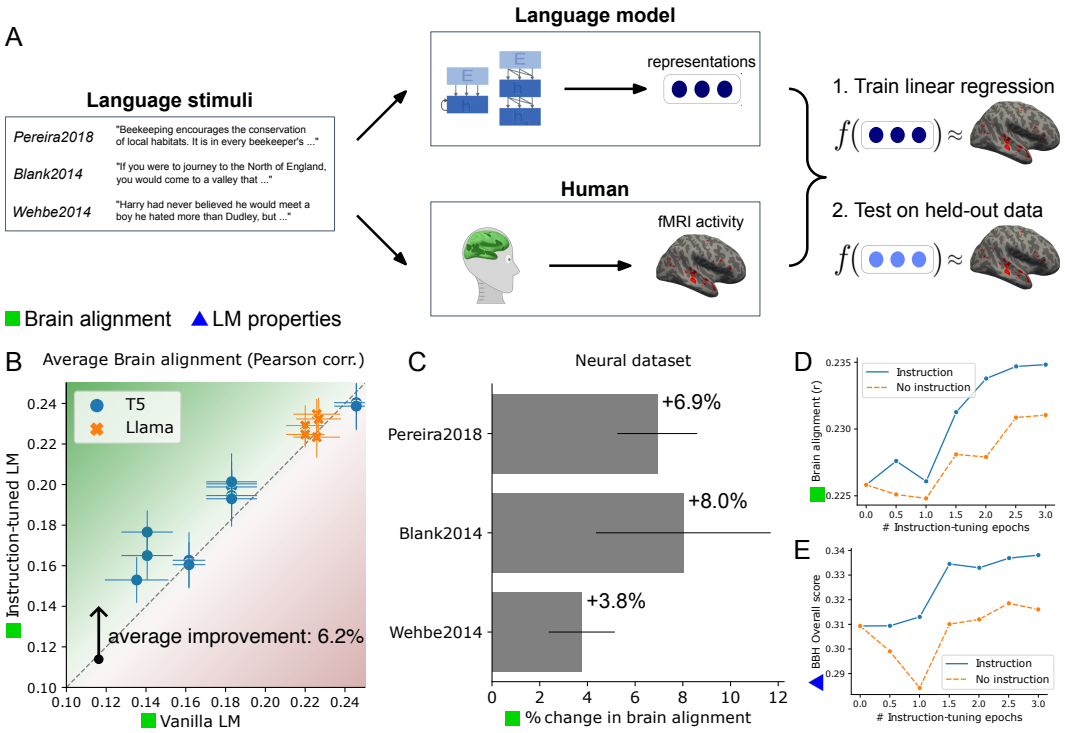

Figure 1: **Instruction-tuning aligns LLM representations to human brain activity. (A)** Method of brain alignment. **(B)** Instruction-tuning improves average brain alignment by 6.2%. We compute each LLM's average brain alignment using the mean of its brain alignment on the 3 neural datasets. Then, we compare the brain alignment of each instruction-tuned LLM against its vanilla counterpart. Each point above the identity line represents an instruction-tuned LLM that has greater brain alignment than its vanilla counterpart. Error bars (here and elsewhere) represent median absolute deviation over human participants. **(C)** Instruction-tuning generally improves brain alignment on all three neural datasets. **(D)** We instruction-tune LLaMA-7B using the Alpaca dataset. We also train an ablation model with the same process and training data, but remove the instruction portion from each training sample. This experiment demonstrates that improvements in brain alignment from instruction-tuning are due to both (1) training data (present in both models) and (2) the process of training LLMs to understand and follow instructions (present only in original model).

and evaluate the brain alignment of the LLM as the Pearson correlation between the predicted and actual brain activity data.

Datasets We use three fMRI datasets to measure the brain alignment of LLMs. Each neural dataset includes the brain activity of a different set of human participants, and uses a different set of language stimuli involving naturalistic stories and sentences. (1) PEREIRA2018 (experiments 2 and 3 from 12): In experiment 2, nine participants read 384 sentences organized into 96 text passages. In experiment 3, six participants read 243 sentences in 72 text passages. Each sentence was displayed for four seconds on a screen. (2) BLANK2014 [2]: The data consists of fMRI recordings of 5 human participants listening to naturalistic stories from the Natural Stories Corpus [7]. Participants listened to stories presented auditorily. (3) WEHBE2014 [26]: The data includes fMRI recordings of 8 human participants reading chapter 9 of the book *Harry Potter and the Sorceror's Stone* [14]. Participants read the chapter at a fixed interval of one word every 0.5 seconds.

Finding 1: Instruction-tuning aligns LLM representations to human brain activity First, we study the effect of instruction-tuning on brain alignment of LLMs. We compute each LLM's average brain alignment as the mean of its brain alignment scores on the 3 neural datasets and find that instruction-tuning improves alignment by an average of 6.2% across all tested LLMs (Figure 1B). We elaborate additional findings in Figure 1.

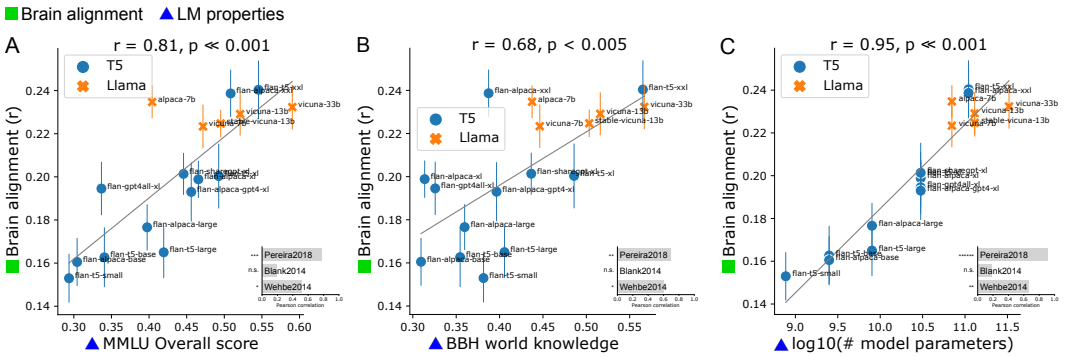

Figure 2: **World knowledge and model size are important factors underlying LLM-brain alignment.** Insets display results on individual datasets with stars reflecting statistical significance (n.s. = $p > 0.05$, * = $p < 0.05$, ** = $p < 0.005$, etc.) **(A)** Brain alignment is significantly and strongly correlated with world knowledge as evaluated by the MMLU Overall score (r = 0.81). This score reports the mean performance across all world knowledge subject domains on MMLU. **(B)** Brain alignment is significantly and strongly correlated with the world knowledge category on the BBH benchmark (r = 0.68). This score reports the mean performance on tasks included in the BBH world knowledge category. **(C)** Brain alignment is significantly and strongly correlated with model size (logarithm of the number of model parameters) (r = 0.95).

Finding 2: World Knowledge and Model Size are key factors underlying LLM-brain alignment To identify factors underlying the representational similarity between LLMs and human brains, we compute the Pearson correlation between LLM brain alignment and various properties of LLMs: performance on benchmarks involving different reasoning abilities (BBH benchmark; 19), performance on benchmarks requiring domain-specific world knowledge (MMLU; 9), language modeling ability, and model size. We elaborate additional findings in Figure 2 and Table 1. Finally, we provide our full results, as well as details about our models and code repositories, in the Appendix.

Table 1: **Brain alignment strongly correlates with world knowledge across all subject domains in MMLU, and the world knowledge problem-solving category in BBH.** At the same time, brain alignment is not significantly correlated with all other types of problem-solving abilities in BBH (e.g., algorithmic or multilingual reasoning). We obtain the p-value after false discovery correction.

Task category	**Brain Alignment Correlation** (r)	**corrected** p**-value**	**Number of tasks**	**Average Model Performance**
MMLU – Overall Score	**0.809**	**0.000329**	57	0.36
MMLU – STEM	**0.792**	**0.000343**	18	0.28
MMLU – Humanities	**0.791**	**0.000343**	13	0.34
MMLU – Social Sciences	**0.807**	**0.000329**	12	0.41
MMLU – Others	**0.809**	**0.000329**	14	0.40
BBH – Overall score	0.384	0.177	23	0.28
BBH – Algorithmic reasoning	0.194	0.558	8	0.22
BBH – Language understanding	0.163	0.585	3	0.43
BBH – World knowledge	**0.679**	**0.005**	5	0.36
BBH – Multilingual reasoning	-0.035	0.895	1	0.19
BBH – Others	0.478	0.083	6	0.27

4 Conclusions

We investigate whether instruction-tuning improves the alignment of LLMs to the human language system. We evaluate 25 LLMs with parameter sizes ranging from 77 million to 33 billion, across three neural datasets of humans reading naturalistic stories and sentences, and find that instruction-tuning generally improves the alignment of LLM representations to brain activity. Delving into the factors underlying LLM-brain alignment, we discover that world knowledge and model size are

key determinants of brain alignment. This correlation suggests that world knowledge helps shape representations in the human language system, and highlights the significance of integrating world knowledge in the development of future LLMs.

References

[1] K. L. Aw and M. Toneva. Training language models to summarize narratives improves brain alignment, Feb. 2023. arXiv:2212.10898 [cs, q-bio].

[2] I. Blank, N. Kanwisher, and E. Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, 112(5):1105–1118, Sept. 2014.

[3] C. Caucheteux and J.-R. King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, Feb. 2022.

[4] Y. K. Chia, P. Hong, L. Bing, and S. Poria. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models, June 2023. arXiv:2306.04757 [cs].

[5] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling Instruction-Finetuned Language Models, Dec. 2022. arXiv:2210.11416 [cs].

[6] I. Dasgupta, A. K. Lampinen, S. C. Y. Chan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill. Language models show human-like content effects on reasoning, July 2022. arXiv:2207.07051 [cs].

[7] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. The natural stories corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[8] A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. Brenner, Y. Matias, K. A. Norman, O. Devinsky, and U. Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, Mar. 2022.

[9] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring Massive Multitask Language Understanding, Jan. 2021. arXiv:2009.03300 [cs].

[10] S. Jain and A. G. Huth. Incorporating Context into Language Encoding Models for fMRI. preprint, Neuroscience, May 2018.

[11] S. R. Oota, M. Gupta, R. S. Bapi, G. Jobard, F. Alexandre, and X. Hinaut. Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey), July 2023. arXiv:2307.10246 [cs, q-bio].

[12] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, Mar. 2018.

[13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, July 2020. arXiv:1910.10683 [cs, stat].

[14] J. Rowling, M. GrandPre, M. GrandPré, T. Taylor, A. A. L. Books, and S. Inc. *Harry Potter and the Sorcerer's Stone*. Harry Potter. A.A. Levine Books, 1998.

[15] M. Safdari, G. Serapio-García, C. Crepy, S. Fitz, P. Romero, L. Sun, M. Abdulhai, A. Faust, and M. Matarić. Personality Traits in Large Language Models, June 2023. arXiv:2307.00184 [cs].

[16] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, Nov. 2021.

[17] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience, Sept. 2018.

[18] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. Ratan Murty, R. Ajemian, and J. J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 2020.

[19] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them, Oct. 2022. arXiv:2210.09261 [cs].

[20] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[21] M. Toneva and L. Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain), Nov. 2019. arXiv:1905.11833 [cs, q-bio].

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and Efficient Foundation Language Models, Feb. 2023. arXiv:2302.13971 [cs].

[23] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

[24] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, M. Patel, K. K. Pal, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. K. Sampat, S. Doshi, S. Mishra, S. Reddy, S. Patro, T. Dixit, X. Shen, C. Baral, Y. Choi, N. A. Smith, H. Hajishirzi, and D. Khashabi. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, Oct. 2022. arXiv:2204.07705 [cs].

[25] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. Reddy A, S. Patro, T. Dixit, and X. Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[26] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLoS ONE*, 9(11):e112575, Nov. 2014.

[27] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang. Instruction Tuning for Large Language Models: A Survey, Aug. 2023. arXiv:2308.10792 [cs].

A Language Models: Parameter count and Number of Layers

Table 2: **Parameter count and number of layers for all 25 vanilla and instruction-tuned LLMs.** The upper part contains encoder-decoder models of the T5 family, the lower parts decoder-only models of the LLaMA family. For the parameter count, “M" refers to million and “B" refers to billion. The number of layers for T5 models is a sum of the number of encoder and decoder layers.

Model	Parameter Count	Number of Layers
t5-small	77 M	16
flan-t5-small	77 M	16
t5-base	250 M	24
flan-t5-base	250 M	24
flan-alpaca-base	250 M	24
t5-large	800 M	48
flan-t5-large	800 M	48
flan-alpaca-large	800 M	48
t5-xl	3 B	48
flan-t5-xl	3 B	48
flan-alpaca-xl	3 B	48
flan-gpt4all-xl	3 B	48
flan-sharegpt-xl	3 B	48
flan-alpaca-gpt4-xl	3 B	48
t5-xxl	11 B	48
flan-t5-xxl	11 B	48
flan-alpaca-xxl	11 B	48
llama-7b	7 B	32
alpaca-7b	7 B	32
vicuna-7b	7 B	32
llama-13b	13 B	40
vicuna-13b	13 B	40
stable-vicuna-13b	13 B	40
llama-33b	33 B	60
vicuna-33b	33 B	60

B Language Models: Links to models weights

Table 3: **Link to model weights for all 25 vanilla and instruction-tuned LLMs.** The upper part contains encoder-decoder models of the T5 family, the lower parts decoder-only models of the LLaMA family. We provide these links for reproducibility purposes.

Model	Link to model weights
t5-small	www.huggingface.co/google/t5-v1_1-small
flan-t5-small	www.huggingface.co/google/flan-t5-small
t5-base	www.huggingface.co/google/t5-v1_1-base
flan-t5-base	www.huggingface.co/google/flan-t5-base
flan-alpaca-base	www.huggingface.co/declare-lab/flan-alpaca-base
t5-large	www.huggingface.co/google/t5-v1_1-large
flan-t5-large	www.huggingface.co/google/flan-t5-large
flan-alpaca-large	www.huggingface.co/declare-lab/flan-alpaca-large
t5-xl	www.huggingface.co/google/t5-v1_1-xl
flan-t5-xl	www.huggingface.co/google/flan-t5-xl
flan-alpaca-xl	www.huggingface.co/declare-lab/flan-alpaca-xl
flan-gpt4all-xl	www.huggingface.co/declare-lab/flan-gpt4all-xl
flan-sharegpt-xl	www.huggingface.co/declare-lab/flan-sharegpt-xl
flan-alpaca-gpt4-xl	www.huggingface.co/declare-lab/flan-alpaca-gpt4-xl
t5-xxl	www.huggingface.co/google/t5-v1_1-xxl
flan-t5-xxl	www.huggingface.co/google/flan-t5-xxl
flan-alpaca-xxl	www.huggingface.co/declare-lab/flan-alpaca-xxl
llama-7b	www.github.com/facebookresearch/llama
alpaca-7b	www.github.com/tatsu-lab/stanford_alpaca
vicuna-7b	www.huggingface.co/lmsys/vicuna-7b-v1.3
llama-13b	www.github.com/facebookresearch/llama
vicuna-13b	www.huggingface.co/lmsys/vicuna-13b-v1.3
stable-vicuna-13b	www.huggingface.co/CarperAI/stable-vicuna-13b-delta
llama-33b	www.github.com/facebookresearch/llama
vicuna-33b	www.huggingface.co/lmsys/vicuna-33b-v1.3

C Code Repositories

We use the Brain-Score repository to evaluate brain alignment for the PEREIRA2018 and BLANK2014 datasets. Link: www.github.com/brain-score/language.

We use an open-source repository to evaluate brain alignment for the WEHBE2014 dataset. Link: www.github.com/awwkl/brain_language_summarization, which builds on www.github.com/mtoneva/brain_language_nlp.

We use Instruct-Eval repository to evaluate MMLU and BBH scores. Link: www.github.com/declare-lab/instruct-eval.

We use Stanford Alpaca repository for instruction-tuning. Link: www.github.com/tatsu-lab/stanford_alpaca).

D Results for Brain alignment

Table 4: **Brain alignment results for all 25 vanilla and instruction-tuned LLMs.** We provide these results for reproducibility purposes.

	PEREIRA2018	BLANK2014	WEHBE2014	Average
t5-small	0.166	0.168	0.071	0.135
flan-t5-small	0.202	0.178	0.079	0.153
t5-base	0.222	0.188	0.074	0.162
flan-t5-base	0.234	0.178	0.076	0.163
flan-alpaca-base	0.227	0.179	0.076	0.161
t5-large	0.270	0.082	0.071	0.141
flan-t5-large	0.311	0.104	0.080	0.165
flan-alpaca-large	0.322	0.126	0.082	0.177
t5-xl	0.285	0.192	0.072	0.183
flan-t5-xl	0.314	0.215	0.072	0.200
flan-alpaca-xl	0.312	0.209	0.075	0.199
flan-gpt4all-xl	0.300	0.206	0.078	0.195
flan-sharegpt-xl	0.323	0.211	0.070	0.201
flan-alpaca-gpt4-xl	0.302	0.205	0.073	0.193
t5-xxl	0.343	0.297	0.096	0.246
flan-t5-xxl	0.350	0.268	0.103	0.240
flan-alpaca-xxl	0.346	0.268	0.102	0.239
llama-7b	0.405	0.154	0.118	0.226
alpaca-7b	0.420	0.167	0.118	0.235
vicuna-7b	0.399	0.152	0.119	0.223
llama-13b	0.412	0.133	0.115	0.220
vicuna-13b	0.423	0.148	0.116	0.229
stable-vicuna-13b	0.415	0.144	0.115	0.225
llama-33b	0.426	0.145	0.109	0.227
vicuna-33b	0.436	0.156	0.105	0.232

Table 5: **Noise ceiling estimates for all 3 neural datasets.** For PEREIRA2018 and BLANK2014, noise ceiling estimates are computed using the Brain-Score repository, with details provided in [16]. For WEHBE2014, noise ceiling estimates are also computed using a similar procedure.

	PEREIRA2018	BLANK2014	WEHBE2014	Average
Noise ceiling	0.32	0.20	0.10	0.21

E Results for Next-word prediction, MMLU, BBH

Table 6: **WikiText-2 NWP loss, MMLU Overall Score, and BBH Overall Score for all instruction-tuned LLMs.** Results for vanilla LLMs are not shown as they are not adapted for the question formats in the MMLU and BBH benchmarks. We provide these results for reproducibility purposes.

	WikiText-2 NWP Loss	MMLU Overall Score	BBH Overall Score
flan-t5-small	0.851	0.294	0.287
flan-t5-base	1.235	0.341	0.308
flan-alpaca-base	1.074	0.304	0.266
flan-t5-large	0.625	0.419	0.370
flan-alpaca-large	0.648	0.397	0.276
flan-t5-xl	0.650	0.493	0.402
flan-alpaca-xl	0.604	0.466	0.270
flan-gpt4all-xl	0.625	0.337	0.212
flan-sharegpt-xl	0.664	0.446	0.363
flan-alpaca-gpt4-xl	0.593	0.456	0.348
flan-t5-xxl	0.638	0.545	0.443
flan-alpaca-xxl	0.607	0.508	0.229
alpaca-7b	4.201	0.404	0.328
vicuna-7b	4.387	0.472	0.331
vicuna-13b	4.130	0.521	0.387
stable-vicuna-13b	4.623	0.495	0.380
vicuna-33b	3.940	0.590	0.426

F Results for Correlations of Brain Alignment with LM properties

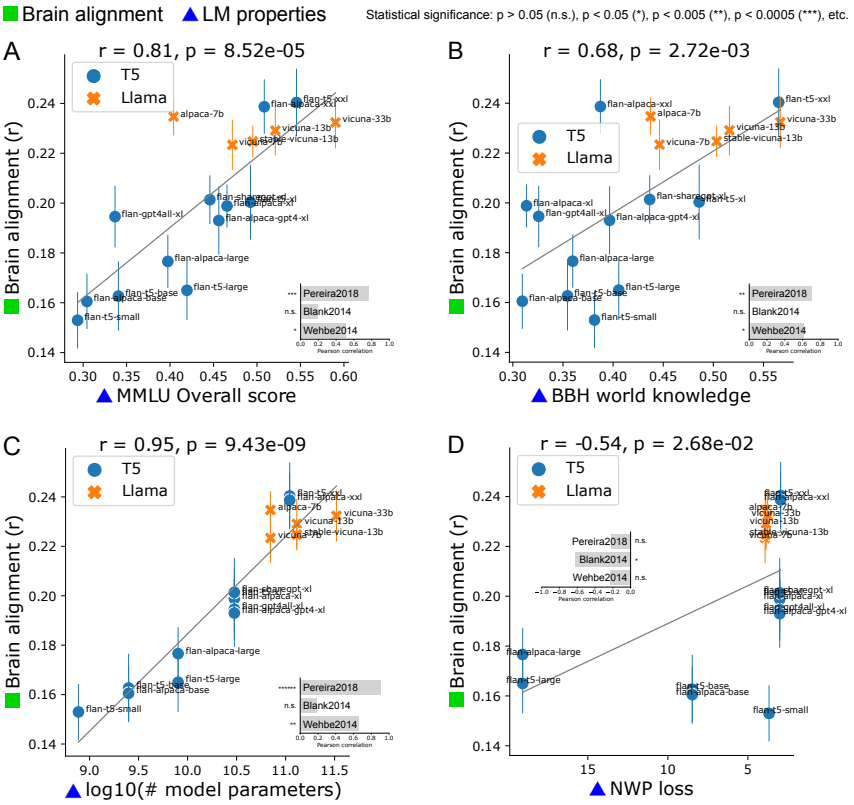

Figure 3: **Correlation between brain alignment and various LM properties:** (A) MMLU benchmark global score, (B) BBH benchmark score with only world knowledge tasks, (C) number of parameters of the model, and (D) Next word prediction (NWP) performance.

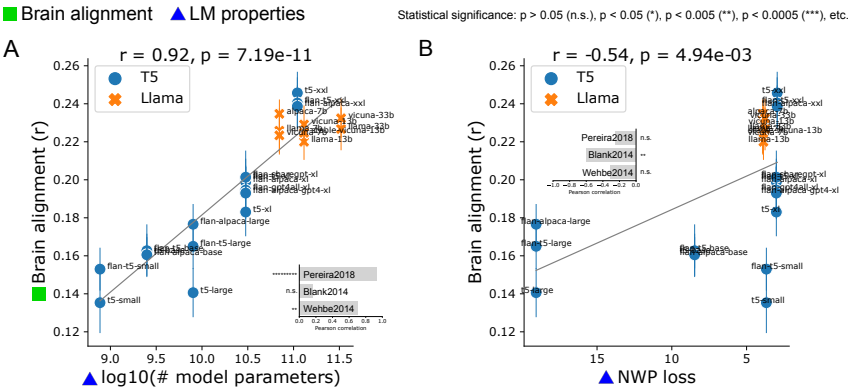

Figure 4: **Correlation between brain alignment and various LM properties for all 25 LLMs:** (A) number of parameters of the model, and (B) Next word prediction (NWP) performance.

G Results for Instruction-tuning LLaMA-7B on Alpaca dataset

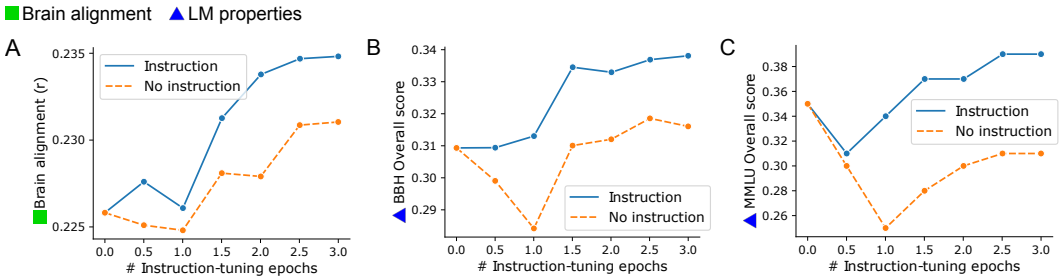

Figure 5: **Improvements in brain alignment from instruction-tuning are due to both additional training data, as well as training to understand and follow instructions.**

Instruction model We instruction-tune LLaMA-7B on the Stanford Alpaca dataset [20] using the default training process, following the code in `www.github.com/tatsu-lab/stanford_alpaca`. In particular, the model is instruction-tuned using 52K instruction-following examples generated through methods inspired by Self-Instruct [23]). This model is labeled “Instruction" in Figure 5.

No instruction model We also train an ablation model with the same process and training data as the default instruction-tuning, but remove the instruction portion from each training sample. This ablation model is labeled “No instruction" in Figure 5. This ablation experiment disentangles: (1) training data (present in both Instruction and No instruction), from (2) training LMs to understand and follow instructions (present only in Instruction).