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 LLM-human similarity in two ways: (1) *brain alignment*, the similarity of LLM internal representations to neural activity in the human language system, and (2) *behavioral alignment*, the similarity of LLM and human behavior on a reading task. 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%, but does not have a similar effect on behavioral alignment. 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 (Wang et al., 2022c). Instruction-tuning often costs only a small fraction of compute relative to pretraining (Chung et al., 2022), 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, as well as to tackle open-world reasoning tasks previously only achievable by humans (Zhang et al., 2023).

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 (Zhang et al., 2023; Chung et al., 2022;

*Equal contribution
†Equal supervision / senior authors

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

Wang et al., 2022b), (2) more aligned to human values (Chia et al., 2023), and (3) more stylistically similar to human outputs (Dasgupta et al., 2022; Safdari et al., 2023).

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 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 (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux and King, 2022), 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 the alignment between LLMs and the human language system, considering two aspects: (1) *brain alignment*, which assesses how closely LLMs' internal representations match neural activity patterns in the human language system, and (2) *behavioral alignment*, which evaluates the similarity between LLM behavior and human behavior. To conduct this study, 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 and per-word perplexity, while for humans, we use previously collected brain activity data from functional magnetic resonance imaging (fMRI) experiments and per-word reading times.

To measure brain alignment, we use the Brain-Score (Schrimpf et al., 2018) linear predictivity metric, assessing how well LLM representations predict human brain activity in response to the same language stimuli (Jain and Huth, 2018; Toneva and Wehbe, 2019; Schrimpf et al., 2021; Oota et al., 2023), using data from three neural datasets: Pereira et al. (2018), Blank et al. (2014), and Wehbe et al. (2014). To evaluate behavioral alignment, we use a benchmark in Brain-Score which calculates the Pearson correlation between LLM per-word perplexity and human per-word reading times from the Futrell et al. (2018) dataset. Perplexity for LLMs and reading times for humans offer insights into comprehension difficulty (Ehrlich and Rayner, 1981; Hale, 2001; Smith and Levy, 2013), allowing us to examine whether LLMs and humans share similarities in terms of which words and sentences they find challenging or surprising. Because models vary in their brain and behavioral alignment across different architectures and training objectives (Schrimpf et al., 2021), we estimate the effect of instruction-tuning across 17 instruction-tuned LLMs and 8 vanilla LLMs, and report a significant increase in brain alignment by instruction-tuned models 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 evaluation of the latter two properties is based on the Big-Bench Hard benchmark (BBH) (Suzgun et al., 2022) and the Massive Multi-task Language Understanding benchmark (MMLU) (Hendrycks et al., 2021), respectively.

We report three major findings:

1. Instruction-tuning generally improves the alignment of LLM representations to brain activity, increasing brain alignment 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).
3. Surprisingly, our results generally indicate that instruction-tuning LLMs does not enhance behavioral alignment with human reading times. Furthermore, behavioral alignment on this dataset demonstrates poor correlations with all other measures we investigate, including task performance and model size (Figure 3).

[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.

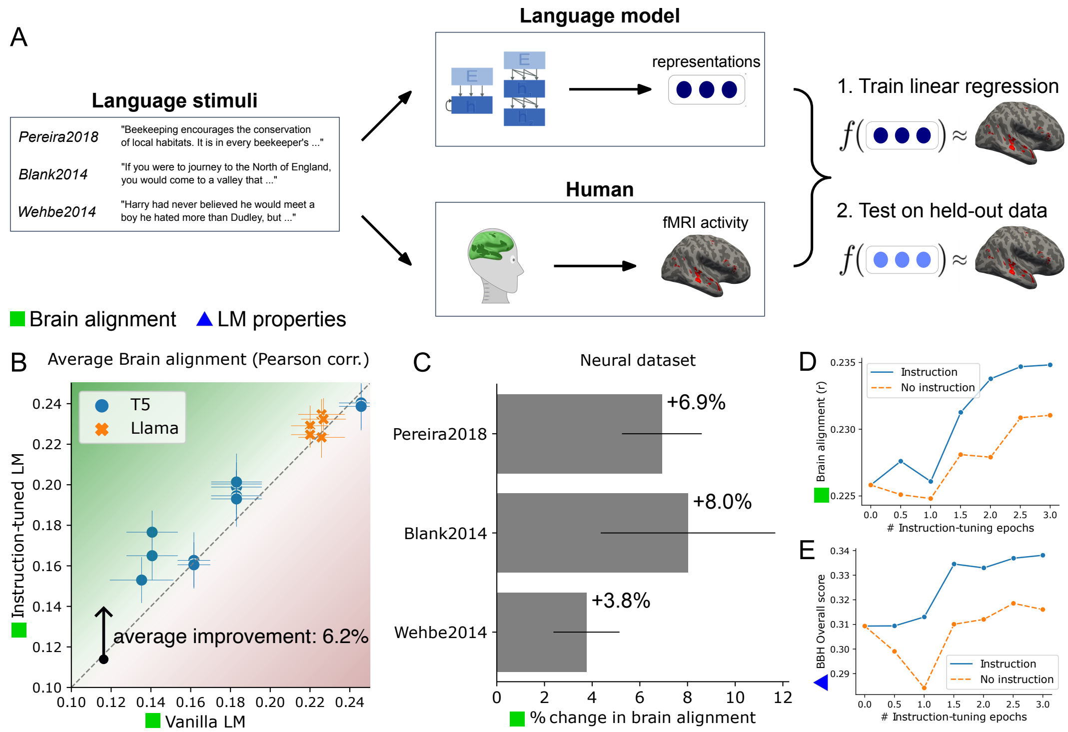

Figure 1: **Instruction-tuning aligns LLM representations to human brain activity. (A)** Method of brain alignment: We measure *brain alignment* as the similarity of an LLM's internal representations to human brain activity, using a linear predictivity metric implemented in Brain-Score. We evaluate the brain alignment of 25 vanilla and instruction-tuned LLMs with sizes between 80M and 33B parameters. We use 3 neural datasets of humans reading naturalistic stories and sentences: PEREIRA2018, BLANK2014, and WEHBE2014. **(B)** Instruction-tuning improves average brain alignment by 6.2% on average. 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 as the default instruction-tuning, 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).

2 Background & Related Work

Effect of Instruction-tuning on LLMs Instruction-tuning is an effective method to enhance the capabilities and controllability of LLMs. It entails additional training of LLMs using pairs of human instructions and desired model outputs. Zhang et al. (2023) categorized the benefits of instruction-tuning into three key aspects: (1) it helps bridge the disparity between the pretraining objective of LLMs on next-word prediction, and the goal of accurately following human instructions, (2) it provides a means for achieving more control and predictability over the behavior of the model compared to standard LLMs, allowing researchers to make them more similar to humans in both capability and output similarity (Chia et al., 2023; Dasgupta et al., 2022; Safdari et al., 2023), and (3) it often costs only a small fraction of compute relative to pretraining, enabling LLMs to swiftly adapt to specific domains (Chung et al., 2022). We contribute to this research area from a neuroscience perspective, by studying whether instruction-tuning makes LLMs more aligned to the human language system in terms of brain and behavioral alignment.

Effect of Finetuning on Brain alignment Our work builds on prior works that study how finetuning affects LMs' alignment to human brain activity. These include studying how brain alignment is modified as a result of finetuning on a wide range of downstream NLP tasks (Oota et al., 2022), finetuning to summarize narratives (Aw and Toneva, 2023), and finetuning LMs to predict brain activity recordings (Schwartz et al., 2019). One goal of these works is to use brain alignment to study how finetuning affects LMs and their representations. Our work demonstrates that instruction-tuning aligns LLM representations to human brain activity. In addition, we investigate why instruction-tuned LLMs align to brain activity by testing the correlation of brain alignment with various domains of world knowledge and problem-solving abilities.

LM properties linked to Brain alignment There is a growing body of work on disentangling the contribution of various model properties towards brain alignment. These include studying how brain alignment is driven by next-word prediction ability (Schrimpf et al., 2021; Caucheteux and King, 2022), multi-word semantics (Merlin and Toneva, 2022), performance on various NLP tasks (Oota et al., 2022), and model size (Antonello et al., 2023). To disentangle the contribution of various LM properties toward brain alignment, instruction-tuned LLMs are especially useful. They have been trained to respond to the question-answer format, allowing us to evaluate various LM properties on a wide range of tasks and in a more fine-grained manner. We expand this area of research by studying instruction-tuned LLMs to identify two properties underlying LLM-brain alignment: world knowledge and model size.

3 Language Models

We evaluate the brain alignment of 25 large language models (LLMs) from two model families: T5 (Raffel et al., 2020) and LLaMa (Touvron et al., 2023). 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. In our study, we use all five T5 models with sizes between 77M to 11B parameters. LLaMA models (Touvron et al., 2023) 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.

For the instruction-tuned variants of T5 models, we utilize a variety of models finetuned on the FLAN suite (15M examples for 1,836 different tasks accompanied by instructions, Chung et al., 2022), Alpaca (52K instruction-following examples generated through methods inspired by Self-Instruct, Wang et al. (2022a), Taori et al., 2023), and GPT4ALL (437K instruction-following examples generated with GPT-3.5-turbo, Anand et al., 2023) datasets. As for the LLaMa model family, we employ Vicuna's 7B, 13B, and 33B models (Chiang et al., 2023), which are finetuned on user-shared conversations. Additionally, we incorporate the StableVicuna-13B model, which further refines the Vicuna-13B model using reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) on a range of conversational and instructional datasets. We also use the 7B version of Alpaca (Taori et al., 2023). Additional details about these LLMs can be found in Appendix A.

4 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: Pereira et al., 2018; Blank et al., 2014; Wehbe et al., 2014). 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 (Schrimpf et al., 2018, 2021; Jain and Huth, 2018; Toneva and Wehbe, 2019; Oota et al., 2023; Aw and Toneva, 2023). Specifically, we use the linear predictivity metric implemented in Brain-Score (Schrimpf et al., 2020, 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, 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.

PEREIRA2018 (experiments 2 and 3 from Pereira et al., 2018): 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.

BLANK2014 (Blank et al., 2014): The data consists of fMRI recordings of 5 human participants listening to naturalistic stories from the Natural Stories Corpus (Futrell et al., 2018). Participants listened to stories presented auditorily.

WEHBE2014 (Wehbe et al., 2014): The data includes fMRI recordings of 8 human participants reading chapter 9 of the book *Harry Potter and the Sorceror's Stone* (Rowling et al., 1998). Participants read the chapter at a fixed interval of one word every 0.5 seconds.

4.1 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). This result holds across all three neural datasets, with average improvements (across models) of +6.9% on PEREIRA2018, +8.0% improvement on BLANK2014, +3.8% on WEHBE2014 (Figure 1C). This provides a strong signal that instruction-tuning aligns LLM representations to human brain activity. Moreover, a smaller instruction-tuned model can have a higher brain alignment than a larger model from the same family that hasn't been instruction-tuned (e.g., Alpaca-7B v.s. LLaMa-13B, see detailed results in Appendix D).

To longitudinally study how instruction-tuning aligns LLM representations to brain activity, we separately instruction-tune a LLaMA-7B model on the Stanford Alpaca instruction dataset (Taori et al., 2023) for 3 epochs. By measuring checkpoints regularly during training, we find that instruction-tuning progressively improves brain alignment (Figure 1D). While it does not monotonically increase alignment across data windows, we do observe a significant increase as the model is tuned. We perform an ablation study to disambiguate the effect on brain alignment of the instruction-following ability provided by the instruction-tuning step, and the effect of the added training data. We fine-tune LLaMA-7B with the same process as we did on Stanford Alpaca, but remove the instruction portion from each training sample. We observe that brain alignment of this ablated model increases during fine-tuning but stays lower than its instruction-following counterpart (Figure 1D).

To identify possible factors for why instruction-tuning improves brain alignment, we investigate other effects of instruction-tuning on LLaMA-7B. We evaluate each training checkpoint's performance on various types of problem-solving and reasoning abilities (BBH benchmark), as well as world knowledge across various domains (MMLU benchmark). Over the process of 3 training epochs, instruction-tuning increases both problem-solving abilities (BBH: Figure 1E) and world knowledge (MMLU: Appendix H). Although these results hint at problem-solving ability and world knowledge as possible factors underlying LLM-brain alignment, they were only conducted on a single LLM. Hence, we perform a rigorous analysis in Section 4.2, where we test the full set of instruction-tuned LLMs, in order to identify important factors underlying LLM-brain alignment.

4.2 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; Suzgun et al., 2022), performance on benchmarks requiring domain-specific world knowledge (MMLU; Hendrycks et al., 2021), language modeling ability, and model size.

World Knowledge and Reasoning abilities MMLU contains 57 tasks, categorized by the subject domain of world knowledge tested: STEM, Humanities, Social Sciences, and Others (a very broad category ranging from finance and marketing to professional medicine). BBH contains 23 tasks, categorized by the type of problem-solving ability tested: Algorithmic reasoning, World knowledge, etc.

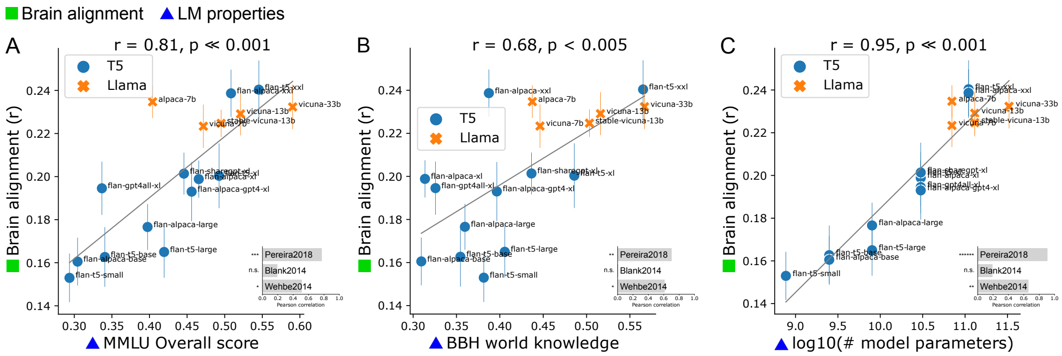

Figure 2: **World knowledge and model size are important factors underlying LLM-brain alignment.** To identify factors underlying the representational similarity between LLMs and human brains, we test the Pearson correlation between LLM brain alignment and various LLM properties, such as model size, various types of problem-solving abilities (BBH benchmark) and world knowledge in various domains (MMLU benchmark). We evaluate all instruction-tuned LLMs on the MMLU and BBH benchmarks. 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).

We measure the performance of LLMs on BBH and MMLU using the `instruct-eval` repository[4] with default settings (3-shots, 5-shots respectively) and preset prompts. We measure the Pearson correlation between LLM brain alignment and performance on each category of the MMLU and BBH benchmarks. We compute the p-value after false discovery rate correction.

We find that brain alignment is significantly and strongly correlated with world knowledge. On the MMLU benchmark, we observe a high correlation between brain alignment scores and the MMLU Overall score (r = 0.81), which reports the mean performance across all world knowledge subject domains on MMLU (Figure 2A). Similarly, brain alignment is also significantly and strongly correlated with the mean performance on tasks included in the world knowledge category of the BBH benchmark (r = 0.68; Figure 2B). Interestingly, we do not find strong correlations with other dimensions of the BBH benchmark (e.g., Algorithmic reasoning and Multilingual reasoning, see Table 1), though this could also be due to limitations of the tested models (most are primarily pretrained on English language data), as indicated by their low raw performance scores on some tasks. Overall, our results provide a strong signal that more accessible representations of world knowledge are a key factor in aligning LLM representations to human brain activity.

Language Modeling Ability Prior works have demonstrated the correlation between brain alignment and next-word prediction (Caucheteux and King, 2022; Schrimpf et al., 2021). We follow this experimental setting for instruction-tuned LLMs and validate a similar finding for the correlation between brain alignment and next-word prediction (r = -0.54, Appendix F). Interestingly, however, we observe that, for instruction-tuned models, the strength of correlation between brain alignment and world knowledge performance (r = 0.81) is greater than the strength of the correlation between brain alignment and next-word prediction (r = -0.54). This result suggests that brain alignment is a better predictor of world knowledge understanding than language modeling, even as the brain alignment scores are derived from a language modeling task.

Model Size Finally, we find that brain alignment is significantly and strongly correlated with model size (as measured by the logarithm of the number of model parameters) with a correlation of 0.95 (Figure 2C). Schrimpf et al. (2021) observe such a pattern for language models, and we find the pattern holds for instruction-tuned models, and models trained at a larger scale than their study (7B+ parameters).

[4] `https://github.com/declare-lab/instruct-eval`

Surprisingly, the LLaMA models attain lower brain alignment scores on BLANK2014 than the T5-XL and T5-XXL models, despite having larger sizes and performance on MMLU and BBH in general. Consequently, we do not observe significant correlations between brain alignment and world knowledge or model size for BLANK2014.

Table 1: **Brain alignment strongly correlates with world knowledge across all subject domains (e.g., STEM, Humanities) in MMLU, as well as 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 evaluate all instruction-tuned LLMs on the MMLU and BBH benchmarks. We test the Pearson correlation between LLM brain alignment and performance on each category of MMLU and BBH. 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

5 Behavioral Alignment

In the previous section, we show that instruction-tuning aligns the internal representations of LLMs to human *brain recordings* (Section 4.1). In this section, we explore whether instruction-tuning also aligns LLM behavior to human *behavior*.

Following the approach previously proposed by Schrimpf et al. (2021) and implemented in the Brain-Score package (Schrimpf et al., 2020), we measure behavioral alignment by evaluating the similarity between LLM per-word perplexity and human per-word reading times, given the same language stimuli (Figure 3A). We use the self-paced reading times dataset from Futrell et al. (2018), consisting of the reading times of 179 human participants recorded while they were visually presented with 10 naturalistic stories. We provide language stimuli from this data as input to LLMs and measure their per-word perplexity. Finally, we evaluate behavioral alignment by computing the Pearson correlation between per-word LLM perplexity and per-word human reading times.

Using the same set of models as in the previous section, we compare the behavioral alignment of each instruction-tuned LLM against its vanilla counterpart. Our results generally indicate that instruction-tuning LLMs does not improve behavioral alignment to human reading times (Figure 3B). For half of the LLMs tested, it results in no change or reduced behavioral alignment. Then, we test the correlation between LLM behavioral alignment and model size, next-word prediction ability, various reasoning abilities (as measured by performance on the BBH benchmark), and world knowledge across various domains (as measured by performance on the MMLU benchmark). Contrary to our findings on the correlation between brain alignment and model size and world knowledge (Section 4.2), we do not find that these factors are correlated with the behavioral alignment of models: world knowledge in Figure 3C ($r = 0.08$, $p = 0.76$), model size in Figure 3D ($r = 0.26$, $p = 0.31$), next-word prediction loss for T5 models in Figure 3E ($r = -0.2$, $p = 0.54$), and next-word prediction loss for LLaMA models in Figure 3F ($r = 0.68$, $p = 0.21$). We discuss our interpretations of these results and possible explanations in Section 6.2.

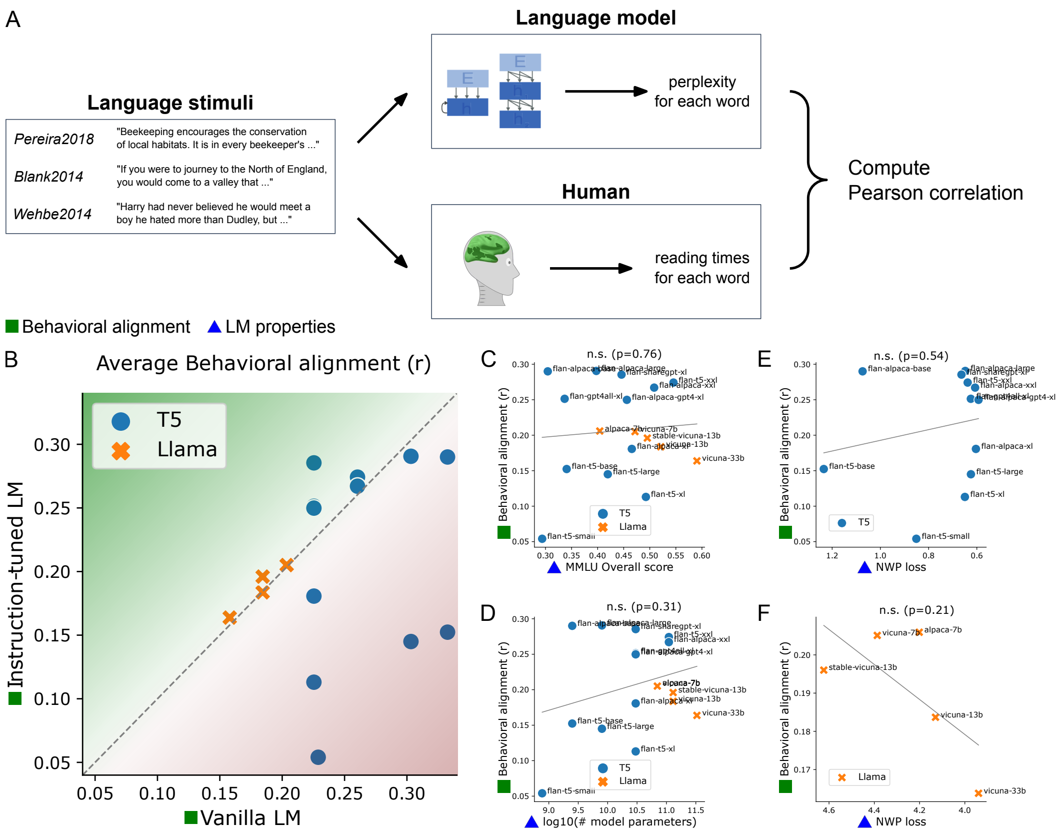

Figure 3: **Instruction-tuning LLMs generally does not improve behavioral alignment to human reading times. Furthermore, behavioral alignment correlates poorly with all other tested measures:** world knowledge, model size, and next-word prediction ability. **(A)** We estimate behavioral alignment on the *Futrell2018* benchmark in Brain-Score which uses a dataset of humans reading naturalistic stories and presents the same language stimuli to the models. The score (r) reflects the similarity between LLM perplexity for each word and human reading times for each word. **(B)** We compare the behavioral alignment of each instruction-tuned LLM against its vanilla counterpart. Instruction-tuning does not generally improve behavioral alignment to humans on a reading task. Furthermore, behavioral alignment is poorly and not significantly correlated with all other measures: **(C)** world knowledge as evaluated using the MMLU benchmark (r = 0.08, p = 0.76), **(D)** model size (r = 0.26, p = 0.31), **(E)** next-word prediction loss value for T5 models (r = -0.2, p = 0.54), and **(F)** next-word prediction loss value for LLaMA models (r = 0.68, p = 0.21).

6 Discussion

6.1 Implications for NLP: Building LLMs

In our work, we investigate the brain alignment of models to evaluate the potential of guiding future LLMs with the human language system. Representations in the human language system support a wide range of downstream tasks that current models still struggle on and, as shown in this work, improved brain alignment correlates with improved performance on downstream tasks such as those requiring world knowledge.

More brain-like internal representations could be a proxy metric for the quality of instruction-tuned models. Prior works have used brain activity results to interpret neural networks (Dong and Toneva, 2023), and use findings around brain alignment to release more performant models (Dapello et al., 2020, 2022; Safarani et al., 2021). Instruction-tuning has emerged as a breakthrough technique to improve LLM benchmark performance, improve the quality of LLM outputs, and allow LLMs to adapt to new tasks using minimal task-specific training examples. However, the manner in which instruction-tuning alters the internal representations of LLMs to achieve these improvements remains unclear. Instruction-tuning might result only in superficial and external changes to LLM behavior, or it might significantly affect how LLMs represent and process language internally. Brain alignment provides a proxy metric for how these representations are structured, and perhaps a roadmap for how LLMs should be instruction-tuned, as we find significant correlation between brain alignment and performance on downstream tasks requiring world knowledge.

6.2 Implications for Neuroscience: Studying LLM-Human Alignment

Instruction-tuned LLMs are useful for studying LLM properties underlying brain and behavioral alignment. To identify why LLM and human brains exhibit representational similarities, prior work has mostly focused on high-level properties such as model size (Antonello et al., 2023), and external behaviors such as predicting missing words Schrimpf et al. (2021); Caucheteux and King (2022). However, a key to understanding these similarities is to identify lower-level or internal properties of LLMs that underlie brain alignment. This includes the amount of knowledge LLMs contain, e.g., factual (AlKhamissi et al., 2022) and commonsense (Sap et al., 2020; Bosselut et al., 2019). Our work shows that we can harness instruction-tuned LLMs for this purpose as they have been trained to respond to a general question format, allowing us to evaluate LLMs in a more fine-grained manner. This allows the study of both internal (e.g., knowledge) and external (e.g., behavior) properties of LLMs, and how they correlate with brain and behavioral alignment.

Examining more dimensions of behavior. To evaluate behavioral alignment, our work and many prior works compare LM and human next-word surprisal on reading tasks (Wilcox et al., 2020; Schrimpf et al., 2021; Eghbal A. Hosseini et al., 2023). This evaluates only a single dimension of LM and human behavior (per-word perplexity and reading times). On the models we test here, behavioral alignment is not significantly correlated with model size, world knowledge, or next-word prediction ability. While next-word prediction performance correlates with alignment to human reading times across a broad band of models (Schrimpf et al., 2021), this trend does not hold up in recent Transformer-based LMs (Oh and Schuler, 2023), having a surprising negative correlation with parameter count (Oh et al., 2022). Our results highlight the need to create more benchmarks to expand the dimensions of behavior examined for both LLMs and humans, in order to holistically evaluate LLM behavior, as well as LLM-human behavioral alignment.

Brain alignment datasets with humans performing diverse tasks. Our work studies brain alignment using neural datasets that are limited to humans reading naturalistic stories and sentences in English. Unfortunately, there does not exist brain activity data for human participants attempting the BBH and MMLU benchmarks. This may explain why brain alignment is not significantly correlated with many categories of problem-solving on BBH, e.g., language understanding. In the future, we hope to study brain alignment with human participants performing more diverse sets of tasks, e.g., reading computer program code (Ivanova et al., 2020). This can identify more factors underlying LLM-brain alignment, and provide insights into how brain activity and the human language system may be shaped by various forms of problem-solving. Furthermore, for the neural datasets in our work, many of the larger models exceed the noise ceiling estimates (Appendix D), highlighting the need for more neural datasets (with better ways of computing noise ceiling estimates).

World knowledge shapes brain activity. Our results show that world knowledge is a key factor in aligning LLM representations to human brain activity. LLMs with greater world knowledge across all tested subject domains produce representations that align more closely to human brain activity. Analogously, this suggests that world knowledge influences human brain activity, shaping the language comprehension systems in the brain.

7 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

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A Review on Language Models as Knowledge Bases, April 2022. URL http://arxiv.org/abs/2204.06031. arXiv:2204.06031 [cs].

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all, 2023.

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fMRI, May 2023. URL http://arxiv.org/abs/2305.11863. arXiv:2305.11863 [cs].

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

Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, 112(5):1105–1118, September 2014. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00884.2013. URL https://www.physiology.org/doi/10.1152/jn.00884.2013.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:189762527.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, February 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models, June 2023. URL http://arxiv.org/abs/2306.04757. arXiv:2306.04757 [cs].

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, December 2022. URL http://arxiv.org/abs/2210.11416. arXiv:2210.11416 [cs].

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. In *Neural Information Processing Systems (NeurIPS)*, June 2020. doi: 10.1101/2020.06.16.154542.

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

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning, July 2022. URL http://arxiv.org/abs/2207.07051. arXiv:2207.07051 [cs].

Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. URL `https://openreview.net/forum?id=p-vL3rmYoqh`.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *bioRxiv*, page 2022.10.04.510681, January 2023. doi: 10.1101/2022.10.04.510681. URL `http://biorxiv.org/content/early/2023/09/19/2022.10.04.510681.abstract`.

Susan F. Ehrlich and Keith Rayner. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655, December 1981. ISSN 00225371. doi: 10.1016/S0022-5371(81)90220-6. URL `https://linkinghub.elsevier.com/retrieve/pii/S0022537181902206`.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina 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). URL `https://aclanthology.org/L18-1012`.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-022-01026-4. URL `https://www.nature.com/articles/s41593-022-01026-4`.

John Hale. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*, pages 1–8, Pittsburgh, Pennsylvania, 2001. Association for Computational Linguistics. doi: 10.3115/1073336.1073357. URL `http://portal.acm.org/citation.cfm?doid=1073336.1073357`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. URL `http://arxiv.org/abs/2009.03300`. arXiv:2009.03300 [cs].

Anna A Ivanova, Shashank Srikant, Yotaro Sueoka, Hope H Kean, Riva Dhamala, Una-May O'Reilly, Marina U Bers, and Evelina Fedorenko. Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife*, 9:e58906, December 2020. ISSN 2050-084X. doi: 10.7554/eLife.58906. URL `https://elifesciences.org/articles/58906`.

Shailee Jain and Alexander G Huth. Incorporating Context into Language Encoding Models for fMRI. preprint, Neuroscience, May 2018. URL `http://biorxiv.org/lookup/doi/10.1101/327601`.

Gabriele Merlin and Mariya Toneva. Language models and brain alignment: beyond word-level semantics and prediction, December 2022. URL `http://arxiv.org/abs/2212.00596`. arXiv:2212.00596 [cs, q-bio].

Byung-Doh Oh and William Schuler. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350, March 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00548. URL `https://doi.org/10.1162/tacl_a_00548`.

Byung-Doh Oh, Christian Clark, and William Schuler. Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators. *Frontiers in Artificial Intelligence*, 5:777963, March 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.777963. URL `https://www.frontiersin.org/articles/10.3389/frai.2022.777963/full`.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.235. URL https://aclanthology.org/2022.naacl-main.235.

Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey), July 2023. URL http://arxiv.org/abs/2307.10246. arXiv:2307.10246 [cs, q-bio].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://api.semanticscholar.org/CorpusID:246426909.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, March 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03068-4. URL https://www.nature.com/articles/s41467-018-03068-4.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, July 2020. URL http://arxiv.org/abs/1910.10683. arXiv:1910.10683 [cs, stat].

J.K. Rowling, M. GrandPre, M. GrandPré, T. Taylor, Arthur A. Levine Books, and Scholastic Inc. *Harry Potter and the Sorcerer's Stone*. Harry Potter. A.A. Levine Books, 1998. ISBN 9780590353403. URL https://books.google.de/books?id=zXgTdQagLGkC.

Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, 34:739–751, 2021.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality Traits in Large Language Models, June 2023. URL http://arxiv.org/abs/2307.00184. arXiv:2307.00184 [cs].

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense Reasoning for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.7. URL https://www.aclweb.org/anthology/2020.acl-tutorials.7.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/407007.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.07.040.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2105646118. URL https://pnas.org/doi/full/10.1073/pnas.2105646118.

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

Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2013.02.013. URL https://www.sciencedirect.com/science/article/pii/S0010027713000413.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them, October 2022. URL http://arxiv.org/abs/2210.09261. arXiv:2210.09261 [cs].

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain), November 2019. URL http://arxiv.org/abs/1905.11833. arXiv:1905.11833 [cs, q-bio].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022a.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, October 2022b. URL http://arxiv.org/abs/2204.07705. arXiv:2204.07705 [cs].

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong 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, December 2022c. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLoS ONE*, 9(11):e112575, November 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL https://dx.plos.org/10.1371/journal.pone.0112575.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior, June 2020. URL http://arxiv.org/abs/2006.01912. arXiv:2006.01912 [cs].

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction Tuning for Large Language Models: A Survey, August 2023. URL http://arxiv.org/abs/2308.10792. 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, as well as behavioral alignment for the FUTRELL2018 dataset. 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 Schrimpf et al. (2021). 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

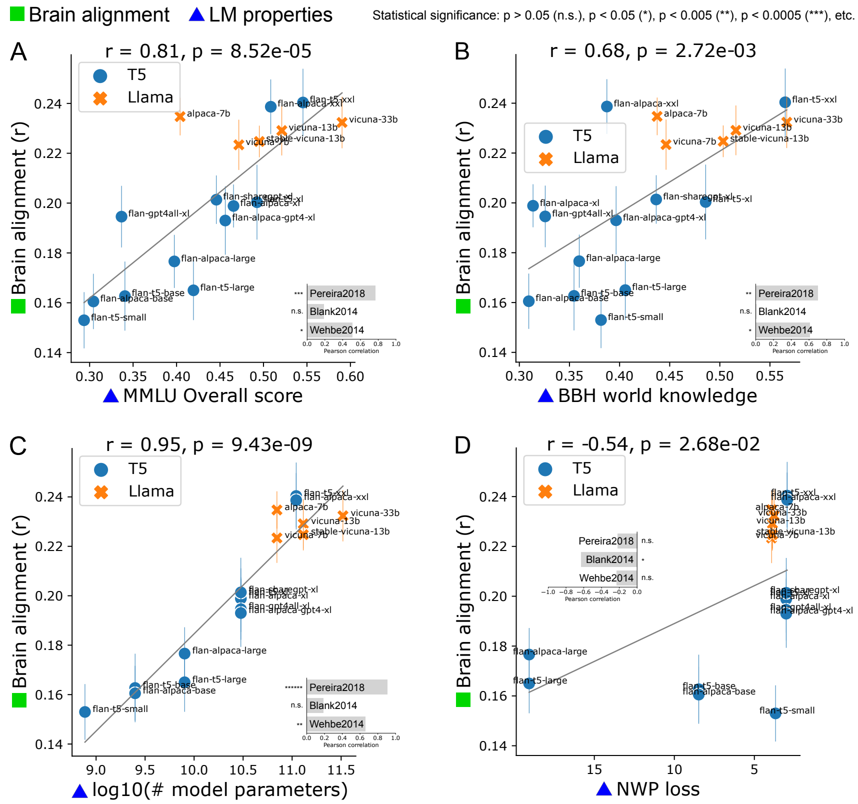

Figure 4: **Correlation between brain alignment and various LM properties:** (A) MMLU benchmark global score, (B) BBH benchmark score with only world knowledge tasks, (C) number of parameters of the model, and (D) Next word prediction (NWP) performance.

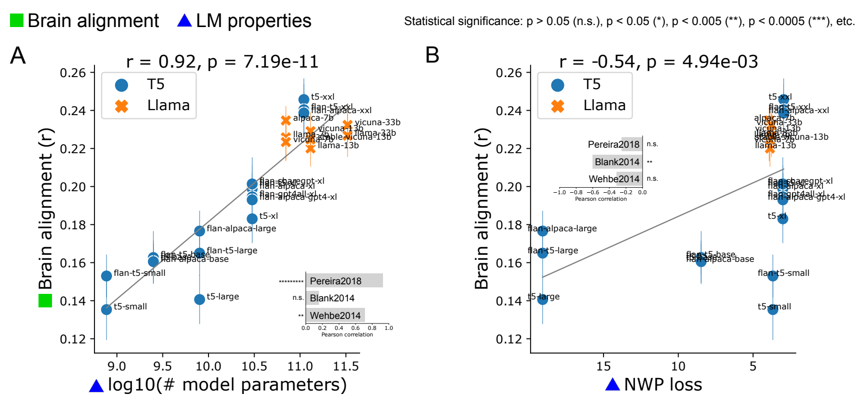

Figure 5: **Correlation between brain alignment and various LM properties for all 25 LLMs:** (A) number of parameters of the model, and (B) Next word prediction (NWP) performance.

G Results for Behavioral alignment

Table 7: **Noise ceiling estimates for the Futrell2018 reading-times dataset.** Noise ceiling estimates are computed using the Brain-Score repository, with details provided in Schrimpf et al. (2021).

	Futrell2018
Noise ceiling	0.76

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

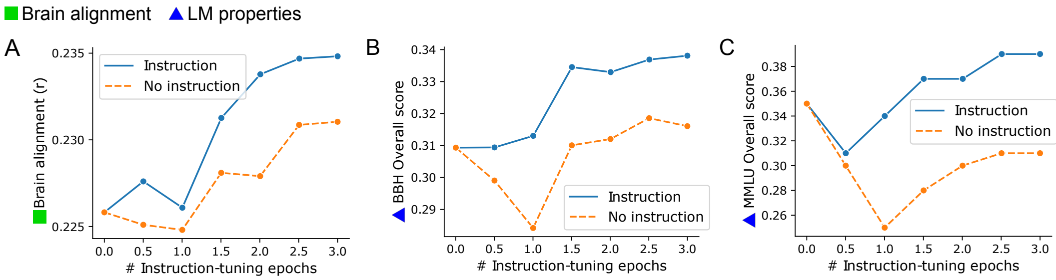

Figure 6: **Improvements in brain alignment from instruction-tuning are due to both additional training data, as well as training to understand and follow instructions.**

Instruction model We instruction-tune LLaMA-7B on the Stanford Alpaca dataset (Taori et al., 2023) 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 (Wang et al., 2022a)). This model is labeled "Instruction" in Figure 6.

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 6. 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).