nature portfolio

Peer Review File

Increasing large language models' alignment with language processing in the human brain

Corresponding Author: Dr Jixing Li

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Professor Li,

Your manuscript "Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain" has now been seen by 4 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- The alignment during naturalistic reading does not capture the full spectrum of human language processing. Please address this.
- Broader generalizations about instruction tuning should test additional instruction-tuned models.
- Datasets with diverse linguistic and cultural contexts should be added to provide a more comprehensive understanding of LLM alignment with human language processing.
- Please address instruction tuning might improve alignment in task-specific or instruction-heavy scenarios.
- Please add a little more narrative "setup" before reporting the results.
- Please perform the model-brain alignment analysis for all voxels across the entire brain.
- Please make use of training/test sets or cross-validation in the description of the model-brain alignment analyses.
- Please align the fMRI analysis more closely with prior work so that readers have more schema in place to understand what you're doing.
- It's unclear whether the reported results on eye movement alignment control for the different number of attention heads.
- Please include modeling results from other GPT-2 variants.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

Link Redacted

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best regards,

Ananya Rastogi, PhD
Senior Editor

Nature Computational Science

Reviewers comments:

Reviewer #1 (Remarks to the Author):

Summary of the Paper:
The paper investigates the alignment of large language models (LLMs) with human cognitive processes during naturalistic reading by comparing their self-attention mechanisms with human eye movement and functional MRI (fMRI) data. The study evaluates two dimensions of LLM development: scaling (increasing the model size) and instruction tuning (fine-tuning on instruction-specific datasets). Using a range of models, including LLaMA and its fine-tuned versions, Alpaca and Vicuna, the authors assess alignment based on regression analyses of attention matrices against human eye-tracking and neural activity patterns. The results demonstrate that: 1. Model scaling consistently enhances alignment with human data, following a predictable scaling law. 2. Instruction tuning does not improve alignment, suggesting that making LLMs instruction-sensitive introduces deviations from human-like language processing. Overall, this work contributes to the broader discourse on whether LLMs can serve as cognitive models by elucidating the effects of scaling and fine-tuning on their brain-encoding performance.

Technical Contributions: The paper makes the following notable contributions-

a. Systematic Comparison of Scaling and Instruction Tuning
The study uniquely isolates the effects of scaling and instruction tuning by employing a consistent experimental design across multiple LLMs, ranging from 774M to 65B parameters. By including instruction-tuned variants (Alpaca and Vicuna), the authors explore whether alignment improvements from scaling also extend to fine-tuned models.

b. Multimodal Cognitive Evaluation
Using the Reading Brain dataset, which combines eye-tracking and fMRI data, the paper employs two complementary modalities to measure alignment:

1. Eye-tracking captures immediate behavioral responses during reading.
2. fMRI provides insights into neural processing patterns in language-relevant brain regions.

c. Quantitative Evidence for the Scaling Law
The study reinforces the scaling law for brain-encoding performance by demonstrating consistent improvements in alignment across models as their size increases. Regression analyses reveal that larger models, such as LLaMA-65B, achieve better alignment with both eye movement and neural activity data.

d. Critical Examination of Instruction Tuning
By analyzing the sensitivity of instruction-tuned models to linguistic instructions and trivial patterns, the paper shows that instruction tuning introduces artifacts in attention distributions. This finding suggests that instruction tuning aligns LLMs with human intentions rather than naturalistic language processing.

e. Open-Source Contribution
The authors provide access to their attention matrices, code, and the experimental dataset, enabling reproducibility and further research in the domain.

Limitations of the Study:

a. Narrow Focus on Naturalistic Reading
The study primarily examines alignment during naturalistic reading, which may not capture the full spectrum of human language processing. For instance, alignment during tasks involving discourse comprehension, inference, or multi-modal integration remains unexplored.

b. Limited Scope of Instruction Tuning
While the study evaluates instruction-tuned models, it does so using only two variants (Alpaca and Vicuna) derived from LLaMA. Broader generalizations about instruction tuning require testing additional instruction-tuned models, such as Gemma, Phi, Mistral, etc., which have demonstrated superior performance in other contexts.

c. Trivial Patterns and Attention Analysis
The study highlights the increased sensitivity of larger models to trivial patterns (e.g., focusing on the first word of a sentence) but does not deeply explore how these patterns affect the cognitive plausibility of LLMs. Are these patterns an artifact of scaling, or do they also manifest in human cognition?

d. Dataset Limitations
The reliance on a single dataset (Reading Brain) may limit the generalizability of the findings. Other datasets with diverse linguistic and cultural contexts could provide a more comprehensive understanding of LLM alignment with human language processing.

e. Instruction Tuning in Task-Specific Settings
The study does not address whether instruction tuning might improve alignment in task-specific or instruction-heavy scenarios. For example, tasks involving multi-step reasoning or complex commands could reveal different patterns of alignment.

Final Verdict:
The paper provides valuable insights into the relationship between LLM development techniques and their alignment with human cognitive processes. By rigorously isolating the effects of scaling and instruction tuning, the authors advance our understanding of the scaling law's role in brain-encoding performance. However, the study's conclusions about instruction tuning should be interpreted cautiously, given the limited scope of models and tasks examined. Future work could address these limitations by:

1. Expanding the analysis to include diverse datasets and tasks.
2. Investigating the cognitive implications of instruction tuning in more complex settings.
3. Exploring the role of fine-tuning techniques in improving task-specific alignment.
4. Exploring and experimentations on more models.

Reviewer #2 (Remarks to the Author):

Gao and colleagues pursue what I think is a very interesting and timely question: does instruction-tuning in LLMs bring them closer to human behavior and neural activity. My intuition would have been "yes", but interestingly the authors show that this does not appear to be the case. In the process, they do replicate scaling effects previously reported in the literature where increasing model size does yield better alignment to humans. While the majority of recent studies on this topic have used passive story-listening fMRI datasets (no overt behavior), the authors use a very neat naturalistic reading fMRI dataset with concurrent eye-tracking. I really like their holistic approach of first evaluating model performance in its own right (with and without instruction tuning over multiple model sizes), then evaluating whether model attention aligns with human gaze behavior (e.g. fixation duration, regression to previous words), and then evaluate the models against human brain activity. In general, I think there's a lot to like about this paper. That said, I had a hard time understanding how and why certain analyses were performed, and I feel very uncertain about the fMRI results as they're currently presented. Below, I list some clarifying questions and suggestions.

Major comments:

My first comment is really just a request to hold the reader's hand a bit more when working through the results. First, readers might benefit from a little more narrative "setup" before you report the results; some kind of one-sentence articulation of a question or hypothesis or expectation or motivation prior to each major result—for those of us with limited working memory :) Second, you introduce a couple of the analyses with a very brief "we regressed X onto Y", and I had a difficult time understanding the structure of what was being mapped onto what. I think a little more detail (e.g., just one sentence) about the shape of the matrices or where they're coming from would be a big help. As the results are currently written, I had to flip back and forth to methods a lot: "is this a wide predictor matrix?" "are they predicting individual voxel time series?"—a little more clarifying detail here would reduce the amount of work the reader needs to do to understand what's going on. Third, I would try to unpack your terminology a bit more for the sake of clarity; for example, on page 6, you first mention "eye regressions" in the context of a "regression" model. It's worth unpacking what an "eye regression" is, or this will be pretty confusing for folks who aren't familiar with the eye-tracking/reading literature.

My main concern with the manuscript has to do with the fMRI analysis and results. First of all, the large cluster reported in Figure 4A looks very weird to me (to the point that, at first, I wondered if it might be a plotting error). I don't really understand regressing eye movements onto fMRI data would yield an ROI effectively centered on the left insula. I understand that there's IFG and a little bit of STG in there, like we would expect from the language network, but still the brain map does not look convincing. For example: this is a reading task, and you're using eye movement as a regressor—wouldn't you expect to see some visual areas (e.g., Hsu et al., 2019)? The significant clusters in panel D also look a bit suspicious to me: again, they seem to be centered on the insula (not a region I would expect) and the t-values within each cluster seem to have a random spatial distribution (like TV static). I would triple-check to make sure there's nothing going wrong with projecting the data onto the cortical surface, nothing wrong with the cluster correction, make sure you're accounting for potential confounds in the time series, etc.

Following on the previous comment, I have questions about the rationale for two analytic choices in the fMRI analysis. First, the authors use the eye-tracking data to identify a cluster of voxels, then examine model-brain alignment for voxels in this cluster? Is this correct? I understand this to be the case because it looks like the maps you obtain in Fig. 4D are a subset of the voxels in Fig. 4A. If this is the case, I think we need a much stronger motivation for constraining the analysis in this way. If you did not perform the analysis within the ROI, then (in line with the previous comment), I am even more suspicious about the resulting brain maps. In either case, I think the authors should either try performing the model-brain alignment analysis for all voxels across the entire brain, or at least try using a more standard set of language areas (e.g., Lipkin et al., 2022). Second, I'm not really sure that I understand why the authors formulate their model-based fMRI encoding analysis the way they do. Most work evaluating LLM representations with fMRI data use ridge regression to map the model embeddings (at a given layer) onto voxelwise activity time series (e.g., Schrimpf et al., 2021; cf. Kumar, Sumers et al., 2024). If I understand correctly, rather than modeling voxel time series (or a sequence of words), the authors construct a word-by-word matrix of the sum(?) of activity corresponding to both words before and after a saccade for each voxel, then use ridge regression to predict the vectorized version of this matrix. Is this correct? Is the reason for doing this to manipulate the fMRI data into a format that resembles the fixation transitions and/or attention matrices? Or does self-paced reading task demand this kind of design somehow? Is there any precedent for this approach in the field? What's the neurobiological motivation for summing the activity for a pair of words? How are you accounting for the hemodynamic lag? My point here is that if the authors are effectively inventing a novel way to construct voxelwise encoding models, we'll need a bit more motivation and detailed explanation.

I don't see any mention of training/test sets or cross-validation in the description of the model-brain alignment analyses. All prior work that I'm aware of on voxelwise encoding models uses cross-validation of some kind (e.g., splitting the stimuli into train and test sets; Huth et al., 2016; Schrimpf et al., 2021; Kumar, Sumers et al., 2024). If I understand correctly, the attention matrices used to predict the neural activity will be quite wide relative to the number of "samples" (vectorized off-diagonal word pairs)—this is presumably why the authors are using ridge regression rather than ordinary least-squares. But simply using ridge regression won't protect against overfitting all on its own—you absolutely need to use some kind of training/test split. (Relatedly, how do the authors select the penalty hyperparameter for ridge without some kind of grid search/nested cross-validation?)

One more point related to the previous two comments: Have the authors tried simply comparing how well the embeddings from models with and without instruction tuning predict voxelwise time series (or voxelwise sequence of words). This kind of analysis would make closer contact with prior work and was the analysis I was expecting when I read the title of the manuscript. In general, I would try to align the fMRI analysis more closely with prior work so that readers have more schema in place to understand what you're doing.

Minor comments:

"For instance, Ouyang et al. (2022) fine-tuned GPT-3 using reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020), and showed that the finetuned models were more aligned with human preferences than GPT-3, despite GPT-3's much larger model size." I don't understand this sentence; sounds like you're saying "GPT-3 is larger than GPT-3".

Figure 2B: In the rightmost plot, I'm not sure I understand why you have four lines total in the legend (or maybe I just can't tell the colors apart for Alpaca versus Vicuna).

Figure 3A: Can you add a legend to the figure indicating what the two groups of lines correspond to?

As far as I understand, MNE is typically used for EEG/MEG analyses. Does MNE's cluster correction software work properly on fMRI data? Does it properly take into the account the 3D (or 2D on the surface) spatial structure of voxels?

"given that GPT2 already predicts neural responses more accurately than an average human"—I'm not sure this is really true; GPT-2 gets close to the noise ceiling for the simple sentence stimuli in the Pereira2018 and Fedorenko2016 datasets, but it doesn't get even close to the noise ceiling for naturalistic stories in the Blank2014 dataset. In Kumar, Sumers et al. (2024, Fig. 2, S1, S4), models like GPT-2 only reach ~30% of the noise ceiling at best.

"our results suggest that human language processing in a naturalistic setting is not sensitive to instructions"—the logic here seems a bit odd, since the humans aren't really performing an "instruction-following" task; consider re-wording this.

We have a recent paper on scaling that extends the fMRI results from Antonello et al., 2023, to ECoG data—we see similar scaling effects: Hong, Wang et al., 2024.

References:

Antonello, R., Vaidya, A., & Huth, A. (2024). Scaling laws for language encoding models in fMRI. Advances in Neural Information Processing Systems, 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/4533e4a352440a32558c1c227602c323-Abstract-Conference.html

Hong, Z.*, Wang, H.*, Zada, Z., Gazula, H., Turner, D., Aubrey, B., ... & Goldstein, A. (2024). Scale matters: large language models with billions (rather than millions) of parameters better match neural representations of natural language. eLife. https://doi.org/10.7554/eLife.101204.1

Hsu, C. T., Clariana, R., Schloss, B., & Li, P. (2019). Neurocognitive signatures of naturalistic reading of scientific texts: a fixation-related fMRI study. Scientific Reports, 9, 10678. https://doi.org/10.1038/s41598-019-47176-7

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458. https://doi.org/10.1038/nature17637

Kumar, S.*, Sumers, T. R.*, Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., ... & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. Nature Communications, 15, 5523. https://doi.org/10.1038/s41467-024-49173-5

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., ... & Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from> 800 individuals. Scientific Data, 9, 529. https://doi.org/10.1038/s41597-022-01645-3

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Signed: Samuel A. Nastase

Reviewer #3 (Remarks to the Author):

Review of "Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain"

According to Nature Computational Science's guidelines for writing review reports (https://www.nature.com/natcomputsci/for-reviewers/writing-your-report), I comment on following aspects of the manuscript below.

Key results.
This article examines the influence of Transformer-based large language models' (LLMs') size and instruction tuning on the alignment between their attention weights and eye movements and blood-oxygen-level-dependent (BOLD) signals. The authors derive attention-based predictors from publicly available LLMs and train ridge regression models to evaluate their fit to eye movements and BOLD signals. The consistent finding across the two experiments is that the increase in model size improves the fit of attention-based predictors to eye movements and BOLD signals, while there is no clear effect of instruction tuning.

Validity.
The interpretation of the experimental results generally seems valid.

Significance.
On first read, the reported results seem like a replication of previous studies on the alignment between LM representations and brain data (Schrimpf et al. 2021 and Antonello et al. 2023 cited in the manuscript, and also Hong et al. 2024 suggested below). Since the data is from a pre-existing study, I guess the new contribution here is the evaluation of instruction tuning and showing that "the scaling law of model-brain alignment holds even with shorter text stimuli and smaller fMRI data," which isn't very exciting to me as it currently stands. I think the authors should aim to bring out their contributions through a deeper discussion in light of previous 'brain scaling law' results.

Data and methodology.
As the data is from a pre-existing publication, I assume the collection and post-processing procedures of the data analyzed in the manuscript have already been validated. It might be more helpful to readers if the authors could include the exact source of the text stimuli (currently it only says "five English STEM articles").

Analytical approach.
I found many aspects of the modeling decisions not clearly justified, and therefore the modeling protocols don't seem as solid as they should be. Below are my concerns:

1) What justifies the use of attention weights as a predictor of BOLD signal and/or measure of processing difficulty? To my knowledge, most computational neuroscience work have used word-by-word surprisal or the raw vectors as predictors. The use of (unidirectional) attention weights seems to have resulted in other modeling decisions that seem pretty unusual to me.

2) The first of these has to do with how the dependent variables are filtered. Both the eye movement and fMRI data seem to be based on those for right-to-left eye regression only. I guess this results in a match in the shape of the dependent variables and the attention weights (a lower triangular matrix), but does this mean the data associated with forward saccades are thrown out completely? To what extent is this decision justified?

3) The authors also mention that attention weights can show "trivial patterns." Putting attention on the first ~~token is perhaps uninterpretable, but why is attending to the immediately previous word or the current word a trivial pattern? Maybe this further suggests that attention weights are not the most appropriate predictors to model fMRI data to begin with.~~

~~4) I like the authors' efforts to ensure a fair comparison between LLMs by controlling for the number of attention heads. However, it's currently unclear whether the reported results on eye movement alignment control for the different number of attention heads (cf. Section on 'Alignment between LLMs and fMRI data' says "We then subtracted the regression scores LLMs with random weights from the resulting regression scores of corresponding LLMs at all the voxels for each subject, ...").~~

~~5) While the authors aimed to control for the number of attention heads, they don't seem to have controlled for the number of layers. Admittedly, I'm not sure how this might be done cleanly, but it seems important to do so as the core 'scaling results' are based on the best performing layer for both the eye movement and fMRI data.~~

~~6) On that note, Figures 3 and 4 seem a bit strange. First, the stars in Figures 3B and 4C seem to be miscolored. Additionally, none of the layers seem to exceed a normalized r-squared of 0.6 in Figure 3A, but the orange star in Figure 3B seems to exceed it.~~

~~7) The next-token prediction loss is not comparable across LLMs if they have different tokenizers. Instead, I would recommend reporting the average word-level surprisal or perplexity (i.e. exponentiated average word-level surprisal), which is not subject to this confound.~~

~~8) I think including modeling results from other GPT-2 variants (i.e. small, medium, XL) would greatly strengthen the evidence for the scaling behavior reported in this study (and be relatively lightweight to conduct).~~

~~Suggested improvements.~~
~~Addressing the points raised in the previous section, either by including a discussion of the rationales underlying modeling decisions or re-conducting some of the experiments, will greatly improve the quality of the manuscript.~~

~~Clarity and context.~~
~~The prose is generally clear, but I have a high-level suggestion to further improve clarity. The Introduction section situates the current work by outlining studies that examine the influence of LLM size on various tasks. Currently, the authors appear to be using the term "performance" to mean many different things, which may be confusing to readers who aren't familiar with all these studies. I would recommend specifying performance on natural language processing tasks, natural language generation tasks, modeling brain imagery, modeling human reading times, etc. where applicable.~~

~~References.~~
~~The following three groups of studies seem relevant to the current manuscript.~~

~~1) These studies have shown that surprisal from larger LLMs yield poorer fits to word-by-word human reading times: Oh and Schuler (2023, "Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?"); Shain, Meister, Pimentel, and Levy (2024, "Large-scale evidence for logarithmic effects of word predictability on reading time").~~

~~2) This study shows instruction tuning LLMs does not improve surprisal's fit to reading times: Kuribayashi, Oseki, and Baldwin (2024, "Psychometric predictive power of large language models").~~

~~3) This concurrent preprint also demonstrates the 'brain scaling law:' Hong et al. (2024, "Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language").~~

~~Your (the reviewer's) expertise.~~
~~I do not have expertise in the collection, validation, and pre-processing of fMRI data, but since this manuscript utilizes publicly available data from a previously published study, I think these aspects are less relevant for evaluating this manuscript.~~

~~Reviewer #3 (Remarks on code availability):~~

~~I only checked whether the repository was empty or not. But based on the methodology described in the manuscript, the modeling experiments seem very straightforward to replicate in any case.~~

~~Reviewer #4 (Remarks to the Author):~~

~~This paper demonstrates that the scaling law holds for alignment between LLMs and human behavioral and neural data, while instruction-tuning is inert for that purpose. I'm generally positive for endorsing this paper for publication, but the following major and minor comments should be addressed before publication.~~

~~Major comments:~~
~~- This paper states "our work is the first to systematically investigate the impact of both scaling and finetuning on the alignment...", but this is an overstatement because the previous literature have already investigated the impact of scaling/instruction tuning on the alignment with human~~

language processing. For example, Kuribayashi et al. (2024) "Psychometric Predictive Power of Large Language Models" has shown that both scaling (as measured in perplexity) and instruction tuning have negative impacts on the alignment with human behavior. Please relax this statement (i.e. the impact of scaling/instruction tuning on the alignment with human neural activity may be new) and, more importantly, discuss how to reconcile the results of this paper and Kuribayashi et al. (2024).
- This paper sometimes performs statistical analyses on "every pair of models", but I don't think that any comparisons except the ones between models of the same sizes are meaningful. Please justify the reason why the comparisons of "every pair of models" is necessary.
- The result that the D_JS of model attentions was significantly larger only between LLaMA-13B and Vicuna-13B is interesting, which may be attributed to the difference between Vicuna and Alpaca in how those models are instruction-tuned. Please explain how differently Vicuna and Alpaca are fine-tuned, and discuss the results from that perspective.
- The results of this paper are based on the models after "we subtracted these patterns from all the attention matrices of the LLMs for the following analyses", which seems to be arbitrary and not usual in the literature. Please justify this design choice.
- This paper performs regression analyses on human behavior and neural activity with "the attention matrices of the LLMs", but how matrices (i.e. model attentions) can predict vectors (i.e. human behavior and neural activity) is not clear. While the details seem to be described in the Methods section, please explain the details briefly in the Results section as well.
- This paper defined fROI based on "the eye movement regression number patterns", which is not a standard practice in the literature to my knowledge. Please justify this design choice.

Minor comments:
- The notations are inconsistent for the GPT family (e.g. GPT-2 vs. GPT4).
- What does the sentence "GPT2 already predicts neural responses more accurately than an average human" mean?

Version 1:

Decision Letter:

** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Professor Li,

Your manuscript "Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain" has now been seen by 4 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Starting at line 447, when describing the regression models, please explicitly state the dimensionality of both the model (X) and the outcome variable (y).
- Are these models fit to one long vector of 7,388 samples, or are these fit for separate models for each sentence? How is overfitting avoided, particularly for the shortest sentences, making ridge regression more necessary.
- Please use out-of-sample prediction (cross-validation) for the core, main-text analyses.
- In the response letter, it is mentioned that "two-level GLM approach" is used in typical fMRI analyses, but this approach is typically used to compare regression coefficients assigned to individual regressors in a regression model. This approach is not valid for comparing R-squared scores for models of varying dimensionality.
- The study normalizes R-squared values and use a one-sample one-tailed t-test with a cluster-based permutation test. However, the one-sample test is problematic as the R-squared values are positive.
- The method of using nilearn's vol_to_surf to resample volumetric data onto the cortical surface may have a low-quality surface projection due to the different cortical anatomy of the fsaverage(5) surface template, which is aligned to MNI space. Best practice is to resample individual-subject, native-space data onto subject-specific FreeSurface reconstruction.
- The negative impact of instruction tuning on alignment with human language processing seems to be consistent with the previous literature, but the impact of scaling is unexplored.
- Please discuss how it is concluded that "the significant voxels identified in the whole-brain analysis largely overlap with our fROI".

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

Link Redacted

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best regards,

Ananya Rastogi, PhD
Senior Editor
Nature Computational Science

Reviewers comments:

Reviewer #1 (Remarks to the Author):

Thank you for your thoughtful and thorough revisions to the manuscript. I have carefully reviewed the updated version and am pleased to see that all the suggested changes and comments have been appropriately addressed. The revisions have significantly improved the clarity and overall quality of the paper. I have no further comments at this stage except the one that the codebase for the newly conducted experiments is still not added to the shared code repository.

Reviewer #1 (Remarks on code availability):

The codebase for the newly conducted experiments is not added to the shared code repository.

Reviewer #2 (Remarks to the Author):

The authors have put considerable effort into improving the clarity of the manuscript and bolstering their findings with additional language models. I still have some concerns about the methodology that I try to unpack below.

Starting at line 447, when describing the regression models, can the authors please explicitly state the dimensionality of both the model (X) and the outcome variable (y)? The new paragraph added at line 140 in the Results is extremely helpful and at least that level of detail should also be present in the Methods (some redundancy is warranted in this case). I had initially thought the models were quite wide, but I think now I understand that the width of X corresponds to the number of heads at a given layer—ranging from 12 to 64? The authors need to be very clear about this. For example, in the response letter, you say you "used ridge regression because our regressor consists of 7388*N columns"—this could be interpreted as a wide model with thousands of columns (typical in fMRI encoding analyses), or a relatively narrow model with only n_heads columns. My understanding is that the latter interpretation is correct: the models have a number of columns equal to the number of heads in a given layer—please confirm.

This raises a question: (1) do you fit these models to one long vector of 7,388 samples, or (2) do you fit separate models for each sentence? If the answer is 1, then I don't think you need to worry about ridge regression at all, OLS would probably be fine. If the answer is 2, then you might encounter overfitting, particularly for the shortest sentences, making ridge regression more necessary. It's important for readers to understand how the regression models are fit at each stage of analysis.

Regarding ridge, there is no reason to think that including a ridge penalty (especially with an arbitrary value of 1) will adequately mitigate overfitting in a training set (i.e. when the R-squared is computed "in-sample"). The appropriate ridge parameter is non-trivially related to the structure of the data/model (e.g., variance, dimensionality) and is typically determined empirically; e.g. using grid search with nested cross-validation. The "correct" ridge penalty is the one that yields the best out-of-sample prediction performance. Note that, when fitting wide models to fMRI data, ridge penalties vary wildly (e.g., Huth et al., 2016, use a grid of ridge coefficients ranging from 10 to 1,000 and settle on coefficients on the scale of ~200).

Whether the authors need to use ridge or not, I would strongly encourage the authors to use out-of-sample prediction (cross-validation) for their core, main-text analyses. The authors try to maneuver around this in several ways, but it ends up complicating the analysis pipeline quite a bit.

In the response letter, the authors mention the "two-level GLM approach" used in typical fMRI analyses, but this approach is typically used to compare regression coefficients (not R-squared values!) assigned to individual regressors in a regression model. Respectfully, I do not think that this approach is valid for comparing R-squared scores for models of varying dimensionality. As the authors acknowledge (line 466), the problem is that bigger models (i.e., with a larger number of heads) will obtain trivially higher fits in a training set based on their dimensionality. They will overfit noise in the training set by definition, and this is equally true of the behavioral and fMRI analyses (see Yarkoni & Westfall, 2017, for a great discussion of this issue). I made a little demo showing that R-squared increases monotonically with model dimensionality, even with completely random model features and target variables, and that using a ridge penalty in-sample does not really help at all: https://github.com/snastase/tutorials/blob/main/overfitting.ipynb

The authors try to get around this problem by using a control (line 488): "we included a control model of matching size for each LLM, where the model weights were randomized to address this potential confound" and "we then subtracted the regression scores of randomly weighted LLMs from the corresponding LLM regression scores for each subject." To me, this raises more questions than it resolves. When you say the model weights were randomized, do you mean you used a randomly-initialized untrained language model? Or that you randomized the encoding weights and re-generated predictions? Note that the variance of possible R-squared values across random models also increases with increasing dimensionality, meaning that this subtraction method could yield larger differences for larger models based on the randomization alone (see the demo).

After this normalization step, the authors then feed these R-squared values into a "one-sample one-tailed t-test with a cluster-based permutation test". It's not clear to me what is being permuted in this one-sample test (maybe sign-flipping?). There's an issue with feeding in-sample R-squared values into a one-tailed test, which is that the R-squared values will be positive by definition: a parametric test against mean zero will be trivially "significant". The authors try to get around this by z-scoring the R-squared values, which will forcibly mean-center them… but z-score them across what, sentences? If you forcibly mean-center the values (with z-score), I don't understand how you can then expect to find a significant effect of the mean being greater than zero (rejecting null hypothesis mean = 0) by means of any statistical test, parametric or otherwise.

Line 536: “Since it is unlikely that ridge regression would cause overfitting at the same voxels in all participants”—this should be removed; when evaluated in-sample, ridge regression will very likely yield overfitting in *all* voxels and participants.

In the revision, they include a supplementary figure (Fig. S1b) where they use a train-test split; what kind of train-test split? Split-half? K-fold? I would use e.g. 5- or 10-fold splitting across sentences. This analysis yields a quite different-looking brain map than the in-sample model fits—why? To me, this brain map actually looks to better match the language network than the others (although it looks strangely spotty/patchy, see the comment on surface sampling below).

I'm truly sorry to be annoying about this, but I am still very hung up on these brain visualizations. I understand that I might have different priors than the authors, but to my mind, there appears to be something wrong. The authors point to Li et al., 2024, for similar looking brain maps—but those maps are derived from MEG, not fMRI. (Really no right hemisphere results for any of these fMRI analyses?) I'm willing to concede that source-reconstructed MEG data may yield maps like this, but they do not seem typical for fMRI. The authors also cite Maris & Oostenveld, 2007, and Brodbeck et al., 2023, for their clustering analysis—but both of these toolboxes are designed for EEG/MEG data, not fMRI data. I trust the authors to coerce their fMRI data into a format that EEG/MEG software can accommodate (line 506), but this is one place I would have a close look when trying to figure out why the maps look so strange. I had a quick look through the code but couldn't find where the fMRI resampling, statistical clustering, etc is happening. For example, does MNE properly deal with the neighborhood structure of the surface vertices for computing clusters? Does it properly deal with the two separate hemispheres? Maybe using a simpler method of correction for multiple tests would be helpful here? Regardless of the cluster coordinate table the authors report in the response letter, the brain maps seem uncomfortably centered around the insula (especially Fig. S2b) and do not very closely reflect the functional anatomy we might expect for language processing (including Lipkin et al., 2022, and Malik-Moraleda et al., 2022, that the authors cite).

Another way to build confidence here would be to run a simple “sanity check” analysis to ensure the brain maps look sensible. Here's an example that might work: For each subject, average the fMRI volumes across TRs within each sentence, resulting in an n_sentences-long vector per voxel. Now, for each voxel, compute the intersubject correlation of these sentence-level time series. If there's reliable variance across sentences—which could be due to interesting things like meaning, or uninteresting things like word rate—I would expect to see ISC in the typical language network.

Taking a closer look at the methods, another possible issue here is in using nilearn's vol_to_surf to resample volumetric data onto the cortical surface. The volumetric data are aligned to MNI space using SPM—but we need to know which MNI template; e.g., “MNI 152 linear”, “MNI 152 nonlinear 6th generation”, “MNI 152 nonlinear 2009”? (See the Lead-DBS website for details: https://www.lead-dbs.org/about-the-mni-spaces/). The potential issue here is that the fsaverage(5) surface template will have different cortical anatomy and is aligned to MNI 305 space (see, e.g., https://surfer.nmr.mgh.harvard.edu/fswiki/CoordinateSystems). This will result in a fairly low-quality surface projection. This is why best practice is to resample individual-subject, native-space data onto the subject-specific FreeSurface reconstruction (as in, e.g., fMRIPrep). You might be able to get a better volume-to-surface projection from neuromaps: https://netneurolab.github.io/neuromaps/generated/neuromaps.transforms.mni152_to_fsaverage.html#neuromaps.transforms.mni152_to_fsaverage

I apologize for putting the authors through the ringer here, and I hope that this manuscript can be published soon—but I really think it will improve the impact of manuscript to make sure the modeling procedure and brain visualizations are as convincing as possible.

Reviewer #3 (Remarks to the Author):

I thank for authors for addressing all of my methodological concerns. Expanding the experiments to include additional models/datasets and re-writing parts of the prose further improved the quality of the manuscript. I agree with the other reviewers that this work makes a nice contribution in investigating the impact of scaling and instruction-tuning on brain-LLM alignment.

I have another set of minor suggestions:

1) I don't think it's ever stated explicitly that each sentence is provided separately as input to the LLMs (cf. providing the entire passage in one context window). I think this is an important modeling decision that should be mentioned, maybe in the "LLMs" paragraph under the "Methods" section.

2) Lines 5, 38: Very minor, but "hundreds" could be replaced with "thousands" for more rhetorical effect.

3) Line 42: In "language comprehension processing," "comprehension" and "processing" seem redundant.

4) Lines 459-465: It seems like this was written to describe model-brain alignment, when the paragraph is about "Alignment between LLMs and eye movement."

Reviewer #3 (Remarks on code availability):

I only checked whether the repository was empty or not. But based on the methodology described in the manuscript, the modeling experiments seem very straightforward to replicate in any case. The code was helpful for me to understand that the stimuli sentences were presented separately to the LLMs.

Reviewer #4 (Remarks to the Author):

All the comments have been sufficiently addressed, so that I would be happy to recommend this paper for publication provided that the following major and minor comments are resolved.

Major comments:
- The negative impact of instruction tuning on alignment with human language processing seems to be consistent with the previous literature, but how about scaling? The previous literature have reported the positive impact of scaling on alignment with human language processing (Kuribayashi et al. 2021, Oh & Schuler 2023, etc.).
- Now I understand that fROI is defined from human eye movement to identify "regions that have already demonstrated significant model-behavior alignment", but how did you conclude "the significant voxels identified in the whole-brain analysis largely overlap with our fROI"? Any

quantification, or just visual inspection?
- The "trivial patterns" seem to be subtracted because the authors "believe they do not reflect underlying human cognitive processes", but in order to empirically test whether this belief is true, please perform the regression analyses without any subtraction of "trivial patterns" and show the results as supplementary information.

Minor comments:
- l.57: alignment between -> alignment with
- l.96: rephrase/delete "as the size increases", because D_JS does not increase with the model size
- l.202: relationship -> relationships
- l.240: applies -> applies to
- l.254: has -> have
- l.282: instruction tuning and prompting -> instruction-tuned and prompted LLMs
- l.283: delete "direct probability measurements from", because API vs. prompting is not relevant here
- l.287: IT -> instruction-tuned
- l.443: the new sentence starting with "We subtracted..." seems to be ungrammatical

Reviewer #4 (Remarks on code availability):

The code provides a README file with enough instructions for installing and running the application.

Version 2:

Decision Letter:

Our ref: NATCOMPUTSCI-24-2037B

21st May 2025

Dear Dr. Li,

Thank you for submitting your revised manuscript "Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain" (NATCOMPUTSCI-24-2037B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please remember to choose, using the manuscript system, whether or not you want to participate in transparent peer review.**
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science. Please do not hesitate to contact me if you have any questions.

Sincerely,

Ananya Rastogi, PhD
Senior Editor
Nature Computational Science

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #2 (Remarks to the Author):

I very very much appreciate all the time and effort the authors have put into addressing my comments! I personally find the results much more convincing now, and I hope other readers will too. I'm satisfied with the current version of the paper and happy endorse the manuscript for publication. —Sam

Reviewer #2 (Remarks on code availability):

I made a quick pass through the GitHub repo and the code seems complete enough.

Reviewer #4 (Remarks to the Author):

Regarding RE16, yes, thank you for reading my intention correctly; I was referring to the "negative" impact of scaling, so the "positive" impact of

scaling was a typo! Now I'd be happy to recommend this paper for publication.

Version 3:

Decision Letter:

Dear Dr Li,

We are pleased to inform you that your Article "Increasing large language models' alignment with language processing in the human brain" has now been accepted for publication in Nature Computational Science.

Once your manuscript is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles or the NIH public access policy) then you should select the gold OA route, and we will direct you to the compliant route where possible. Because authors warrant under our subscription licensing terms that they haven't committed to licensing any version of their article under a licence inconsistent with the terms of our agreement – including the applicable embargo period – publication under the subscription model isn't suitable for authors whose funders require no embargo.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see https://www.nature.com/natcomputsci/for-authors). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Computational Science as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). We also welcome suggestions for the Hero Image, which appears at the top of our home page; these should be 72 dpi at 1400 x 400 pixels in JPEG format. Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Computational Science logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

We look forward to publishing your paper.

Best regards,

Kaitlin McCardle, PhD
Senior Editor
Nature Computational Science

P.S. Click on the following link if you would like to recommend Nature Computational Science to your librarian: https://www.springernature.com/gp/librarians/recommend-to-your-library

Dear Editors and Reviewers,

Thank you for your letter of Jan. 15th, 2025 regarding our submission to Nature Computational Science (MS#: NATCOMPUTSCI-24-2037). We would like to express our sincere gratitude for your insightful comments and valuable suggestions. All four reviewers provided constructive comments that in several cases, point to the same issues or concerns with the previous version of our submission. We believe that these inputs have significantly strengthened and enhanced the quality of our work. In this revision, we have carefully considered and addressed all the points raised, and we hope you and the reviewers will find this revision acceptable for publication in your journal.

In response to the editors and the reviewers' comments, we have made several key revisions that have improved the clarity, depth, and rigor of our manuscript, along the following lines (text highlighted represents our revisions):

1. The alignment during naturalistic reading does not capture the full spectrum of human language processing; Datasets with diverse linguistic and cultural contexts should be added to provide a more comprehensive understanding of LLM alignment with human language processing: We have expanded our analysis to include an fMRI dataset in which participants listened to narratives in Chinese and answered multiple-choice comprehension questions. This dataset introduces diversity in modality (listening), linguistic and cultural context (Chinese), and task engagement. We believe these results support the scaling law in model-brain alignment across a broader spectrum of human language processing (see RE1,4,5).

2. Broader generalizations about instruction tuning should test additional instruction-tuned models: We have now included results from two other fine-tuned models: Gemma-Instruct 7B and Mistral-Instruct 7B. (see RE2).

3. Please address instruction tuning might improve alignment in task-specific or instruction-heavy scenarios: We have added results from regressing fine-tuned LLaMA3 7B and LLaMA3 70B models against the fMRI data collected while participants answered multiple-choice comprehension questions about the preceding listening session (see RE1,RE5).

4. Please add a little more narrative "setup" before reporting the results: We have added some narratives at the beginning of major results sections (see RE7).

5. Please perform the model-brain alignment analysis for all voxels across the entire brain: We have performed model-brain analysis for all voxels using LLaMA 7B and 65B as a comparison with our fROI results (see RE11).

6. Please make use of training/test sets or cross-validation in the description of the model-brain alignment analyses: We have employed the train-test split approach using LLaMA 7B and compared the results with our two-level GLM approach without train-test split (see RE13).

7. Please align the fMRI analysis more closely with prior work so that readers have more schema in place to understand what you're doing: We have added justification for our approach in the context of prior work and showed similar results using the more widely applied RSA methods for the additional naturalistic fMRI dataset (see RE12).

8. It's unclear whether the reported results on eye movement alignment control for the different number of attention heads: We have clarified the control of different number of attention heads in both model-eye and model-brain alignments (see RE27,28).

9. Please include modeling results from other GPT-2 variants: We have included results from all GPT-2 variants (see RE31).

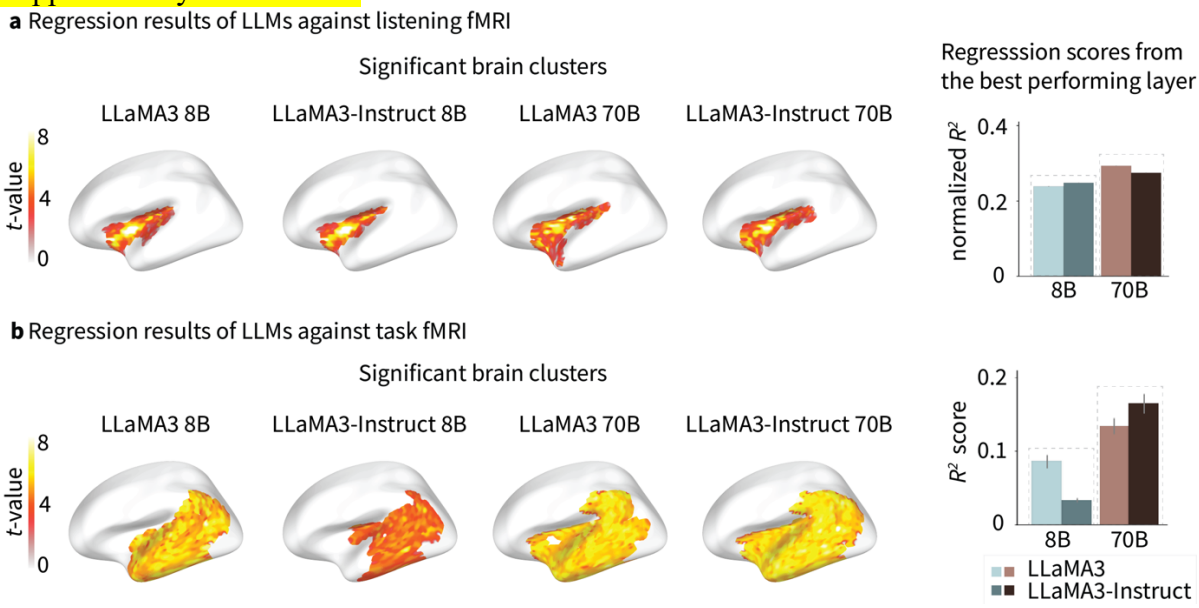
Additionally, we made two minor revisions to the plots: (1) We removed the regression plots across layers in Figures 3 and 4 as they became overly cluttered with the inclusion of five additional models, as requested by Reviewers R1 and R3. Moreover, the regression scores from individual layers did not offer additional insights beyond those from the best-performing layer, which remains the primary finding of this study. (2) We excluded the results for regressive eye duration in the model-eye movement analyses, as they closely mirrored the findings for regressive eye number but with a smaller effect size. Given that this metric does not contribute meaningfully to the primary results, we opted to remove it for clarity.

Below, we address each issue raised by the reviewers in a point-by-point format, providing the corresponding revised text or a description of the revisions made. All our responses are highlighted in yellow, as well as the corresponding revisions in the revised manuscript.

Reviewers comments:
Reviewer #1 (Remarks to the Author):
Limitations of the Study:
a. Narrow Focus on Naturalistic Reading
The study primarily examines alignment during naturalistic reading, which may not capture the full spectrum of human language processing. For instance, alignment during tasks involving discourse comprehension, inference, or multi-modal integration remains unexplored.
RE1: We have now incorporated additional results from participants who listened to a 20-minute audiobook in Chinese while undergoing fMRI scanning. We regressed the attention weights of the base and fine-tuned LLaMA3 7B and LLaMA3 70B models against the fMRI data matrices at the paragraph level (See "Additional results from fMRI data of naturalistic listening" in the Supplementary information for detailed description of the dataset and methods). We used the LLaMA3 models for their better performance in Chinese. This analysis extends beyond sentence-level comprehension to discourse-level processing and introduces both a different modality (listening vs. reading) and a different language (Chinese vs. English). Our findings remained consistent: Model scaling had a significant effect on model-brain alignment, while fine-tuned and base models of the same size showed no difference in brain encoding performance (see Supplementary Fig. 2a and Supplementary Table 12-13).

Additionally, we regressed the predictions from the fine-tuned LLaMA3 7B and LLaMA3 70B models against the fMRI data collected while participants answered multiple-choice comprehension questions about the preceding listening session (See "Additional results from fMRI data of comprehension tasks" in the Supplementary information for detailed description of the methods). Our results showed that in smaller sizes, LLaMA3 exhibited a significantly higher regression score (mean=0.219±0.004) compared to LLaMA3-Instruct (mean=0.211±0.004, $t=58.073$, $p<0.001$). In the larger size, LLaMA3-Instruct showed a higher mean regression score (mean=0.259±0.005) across participants compared to the base model (mean=0.243±0.005, $t=80.528$, $p<0.001$), but no significant brain cluster has been found between the contrast of the two models' R^2 maps (see Supplementary Fig. 2b and Supplementary Table 14-15). We believe these results support the scaling law in model-brain alignment across a broader spectrum of human language processing. We have added results of

the additional analyses on pp.6-7 of the revised manuscript, Supplementary Fig. 2 and Supplementary Table 12-15.



Supplementary Fig. 2 Regression results from additional fMRI datasets. **a,** Significant brain clusters showing alignment between LLMs and fMRI data during naturalistic listening. The right panel displays the averaged regression scores from the best-performing layer across participants. **b,** Significant brain clusters showing alignment between LLMs and fMRI data during the question-answering task. The right panel presents the averaged regression scores from the best-performing layer across participants.

b. Limited Scope of Instruction Tuning

While the study evaluates instruction-tuned models, it does so using only two variants (Alpaca and Vicuna) derived from LLaMA. Broader generalizations about instruction tuning require testing additional instruction-tuned models, such as Gemma, Phi, Mistral, etc., which have demonstrated superior performance in other contexts.

RE2: We selected only two fine-tuned models based on LLaMA because we want to investigate the effect of fine-tuning while controlling for potential confounding factors such as variations in model architecture and training data. Obviously, it is not possible to test all fine-tuned models given that newer models are emerging so rapidly. Nevertheless, we have now included additional results from Gemma-Instruct 7B (Gemma Team, 2024) and Mistral-Instruct 7B (Jiang et al., 2023). We excluded the Phi models, as Phi1 and Phi2 are relatively small in size (1.3B and 2.7B, respectively), and Phi3 employs a tiktoken tokenizer that is not comparable with those used by other LLMs (Abdin et al., 2024). Our results showed that Gemma-Instruct 7B and Mistral-Instruct 7B (1) did not improve next-word prediction for the text stimuli (see Fig. 2a and Supplementary Table 1-2); (2) did not significantly differ from LLaMA 7B in their reliance on trivial patterns (see Fig. 2d and Supplementary Table 5-6); (3) did not improve their alignment with the eye movement patterns compared to LLaMA 7B (see Fig. 3a and Supplementary Table 7-8); (4) did not improve in brain encoding performance compared to LLaMA 7B (see Fig. 4b and Supplementary Table 10-11). We have added the results and figures to the corresponding sections in the revised manuscript. Details of the models are provided in Table 1 and in the "LLMs" section in Methods.

c. Trivial Patterns and Attention Analysis

The study highlights the increased sensitivity of larger models to trivial patterns (e.g., focusing on the first word of a sentence) but does not deeply explore how these patterns affect the cognitive plausibility of LLMs. Are these patterns an artifact of scaling, or do they also manifest in human cognition?

RE3: Our apologies for plotting the figure for the trivial pattern results with wrong labels, which caused confusion. Our analyses actually suggest decreased sensitivity of larger models to trivial patterns, not the other way round. We showed the correct results in our earlier work using only the eye-tracking data (see Figure 8 and Figure 9 in Gao et al., 2023). We have now revised the plot in Fig. 2d, Supplementary Table 5-6. Given that similar patterns were not observed in human eye movement data, we believe they do not reflect underlying human cognitive processes. Since the attention weights of larger models display fewer trivial patterns compared to smaller models, this reduced sensitivity may contribute to their greater cognitive plausibility. We have added this discussion on p.5 of the revised manuscript.

d. Dataset Limitations

The reliance on a single dataset (Reading Brain) may limit the generalizability of the findings. Other datasets with diverse linguistic and cultural contexts could provide a more comprehensive understanding of LLM alignment with human language processing.

RE4: As mentioned in our response for the first comment, we have now included additional results from a naturalistic listening dataset where Chinese speakers listen to 2 sections of The Little Prince (20 minutes long in total) in the fMRI scanner. This dataset sufficiently differ from the English reading task with a different linguistic and cultural context and a different comprehension modality. Our results still supported a scaling effect over finetuning (see Supplementary Fig.2a). Details of the analyses are added to the "Additional results from fMRI data of naturalistic listening" section in the Supplementary information.

e. Instruction Tuning in Task-Specific Settings

The study does not address whether instruction tuning might improve alignment in task-specific or instruction-heavy scenarios. For example, tasks involving multi-step reasoning or complex commands could reveal different patterns of alignment.

RE5: As mentioned in our response for the first comment, we have now included results from task fMRI data where participants completed multi-choice questions based on previous listening sessions. We believe this task involves relatively complex reasoning as well as memorizing previous contents (see an example question below). We prompted the fine-tuned LLMs with the text from the listening session along with the multiple-choice questions. We then extracted each model's probability for the correct answer and computed its surprisal (negative log probability) to regress against the fMRI scans time-locked to the button press for each question (see the "Additional results from fMRI data of comprehension tasks" in the Supplementary information for detailed description of the methods). We still did not observe significantly higher regression scores for the fine-tuned vs. base LLaMA3 for the task fMRI data (see Supplementary Fig. 2b).

Why is it difficult to talk to the Little Prince?
a. He doesn't speak.
b. He doesn't ask questions.
c. He speaks an alien language.
d. He doesn't answer questions directly.

Final Verdict:

The paper provides valuable insights into the relationship between LLM development techniques and their alignment with human cognitive processes. By rigorously isolating the effects of scaling and instruction tuning, the authors advance our understanding of the scaling law's role in brain-encoding performance. However, the study's conclusions about instruction tuning should be interpreted cautiously, given the limited scope of models and tasks examined. Future work could address these limitations by:

1. Expanding the analysis to include diverse datasets and tasks.
2. Investigating the cognitive implications of instruction tuning in more complex settings.
3. Exploring the role of fine-tuning techniques in improving task-specific alignment.
4. Exploring and experimentations on more models.

RE6: (1) We have expanded our analysis to include an fMRI dataset in which participants listened to narratives in Chinese and answered multiple-choice comprehension questions. This dataset introduces diversity in modality (listening), linguistic and cultural context (Chinese), and task engagement. (2) We regressed model predictions for the comprehension questions against the fMRI data and found no significant advantage for the fine-tuned models. (3 & 4) We also included results from additional fine-tuned models, namely Gemma-Instruct 7B and Mistral-Instruct 7B. Despite differences in their finetuning techniques, these models did not demonstrate significantly improved brain encoding performance compared to LLaMA. However, we agree that caution is warranted when interpreting these results, as we have not specifically evaluated the fine-tuned LLMs on other task-based fMRI datasets where participants performed tasks aligned with the instruction-following nature of the fine-tuning process. This limitation is due to the lack of such openly available datasets, and we have now acknowledged this in the Discussion section on p.8 of the revised manuscript.

Reviewer #2 (Remarks to the Author):

Gao and colleagues pursue what I think is a very interesting and timely question: does instruction-tuning in LLMs bring them closer to human behavior and neural activity. My intuition would have been “yes”, but interestingly the authors show that this does not appear to be the case. In the process, they do replicate scaling effects previously reported in the literature where increasing model size does yield better alignment to humans. While the majority of recent studies on this topic have used passive story-listening fMRI datasets (no overt behavior), the authors use a very neat naturalistic reading fMRI dataset with concurrent eye-tracking. I really like their holistic approach of first evaluating model performance in its own right (with and without instruction tuning over multiple model sizes), then evaluating whether model attention aligns with human gaze behavior (e.g. fixation duration, regression to previous words), and then evaluate the models against human brain activity. In general, I think there’s a lot to like about this paper. That said, I had a hard time understanding how and why certain analyses were performed, and I feel very uncertain about the fMRI results as they’re currently presented. Below, I list some clarifying questions and suggestions.

Major comments:

My first comment is really just a request to hold the reader’s hand a bit more when working through the results. First, readers might benefit from a little more narrative “setup” before you report the results; some kind of one-sentence articulation of a question or hypothesis or expectation or motivation prior to each major result—for those of us with limited working memory :)

RE7: We thank the Reviewer for the enthusiastic endorsement of our approach and methodology, along with the suggestions. We have now added some narratives at the beginning of major results sections (highlighted in green) and condensed the descriptions of statistical results to comply with the journal’s 3,500-word limit.

Second, you introduce a couple of the analyses with a very brief "we regressed X onto Y", and I had a difficult time understanding the structure of what was being mapped onto what. I think a little more detail (e.g., just one sentence) about the shape of the matrices or where they're coming from would be a big help. As the results are currently written, I had to flip back and forth to methods a lot: "is this a wide predictor matrix?" "are they predicting individual voxel time series?"—a little more clarifying detail here would reduce the amount of work the reader needs to do to understand what's going on.

RE8: Thank you for pointing this out. We kept the results briefly mainly due to the word-limit restriction. We have now added more details on the regression analyses on p.5 in the revised manuscripts: We extracted the lower-triangle portions (excluding the diagonal line) of the attention matrix $n_{word} \times n_{word} \times n_{head}$ from all attention heads for every sentence. The attention matrices for all sentences were concatenated to create a regressor with dimensions $7388 \times n_{head}$ for each layer, where 7388 represents the total number of elements obtained after concatenating the lower triangles of the attention matrices across all sentences in our stimuli. For the human eye saccade data, we constructed matrices for saccade number $E_{num} \in \mathbb{R}^{n_{word} \times n_{word}}$ for each sentence. Each cell at row l and column m in E_{num} represents the number of eye fixation moving from the word in row l to the word in column m, respectively. We then extracted the lower-triangle parts of the matrices which marks right-to-left eye movement. Similar to the models' attention matrices, we flattened the regressive eye saccade number matrices for all sentences and concatenated them to create 7388-length vectors for each subject. We then performed ridge regression for each model layer, using the $7388 \times n_{head}$ regressor to predict each subject's regressive eye saccade number vectors. The final R^2_{model} was normalized by the $R^2_{ceiling}$, where the ceiling model represents the mean of all subjects' regressive eye saccade number vectors.

Third, I would try to unpack your terminology a bit more for the sake of clarity; for example, on page 6, you first mention "eye regressions" in the context of a "regression" model. It's worth unpacking what an "eye regression" is, or this will be pretty confusing for folks who aren't familiar with the eye-tracking/reading literature.

RE9: This "eye regression" terminology is indeed confusing and we have revised them to "regressive eye saccades". We also added an explanation of the term on p.5 of the revised manuscript: Regressive saccades occur when readers revisit earlier text, highlighting the importance of previous words in understanding the current word (Liversedge & Findlay, 2000)—similar to how attention weights function in LLMs.

My main concern with the manuscript has to do with the fMRI analysis and results. First of all, the large cluster reported in Figure 4A looks very weird to me (to the point that, at first, I wondered if it might be a plotting error). I don't really understand regressing eye movements onto fMRI data would yield an ROI effectively centered on the left insula. I understand that there's IFG and a little bit of STG in there, like we would expect from the language network, but still the brain map does not look convincing. For example: this is a reading task, and you're using eye movement as a regressor—wouldn't you expect to see some visual areas (e.g., Hsu et al., 2019)? The significant clusters in panel D also look a bit suspicious to me: again, they seem to be centered on the insula (not a region I would expect) and the t-values within each cluster seem to have a random spatial distribution (like TV static). I would triple-check to make sure there's nothing going wrong with projecting the data onto the cortical surface, nothing wrong with the cluster correction, make sure you're accounting for potential confounds in the time series, etc.

RE10: We understand why you have concerns regarding the lack of correlation between eye movements and visual regions in the brain. However, our regressor is not time-locked to

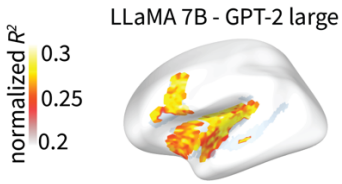
individual eye movements. Instead, it uses the number of regressive saccades to construct a relational matrix for each word in every sentence. These matrices reflect the relative importance of words within a sentence, much like attention weights in LLMs. As a result, the significant brain regions identified for this regressor are within the language network rather than the visual regions. Our ROI shown in Fig. 4a covered a broad region in the LSTG as well as the LIFG, with the largest cluster centered around the LSTG, not the left insula. See the attached screenshot from our spatial clustering analyses (Maris & Oostenveld, 2007) using eelbrain (Brodbeck et al., 2023).

```
In [144]: res = pickle.load(open('Results/fmri/res_eyefmri.p','rb'))
     ...: print(res.clusters)
     ...:
```

id	n_sources	location	hemi	tstart	tstop	duration	v	p	sig
1	823	superiortemporal-lh	lh	0	0.001	0.001	7565.6	0	***
8	193	parsopercularis-lh	lh	0	0.001	0.001	2173.1	0.0001	***
9	65	parsorbitalis-lh	lh	0	0.001	0.001	299.98	0.0966	

NDVars: cluster

For the figures in panel d, we understand that the t-values may appear distributed. However, this is a common pattern observed in results from spatiotemporal clustering analyses using z-scored R^2 maps (see Li et al., 2024). We have carefully reviewed our analyses and confirmed that there are no errors. We plotted the z-transformed R^2 values within the significant cluster for the contrast between LLaMA 7B and GPT-2 large, the results appear less distributed.



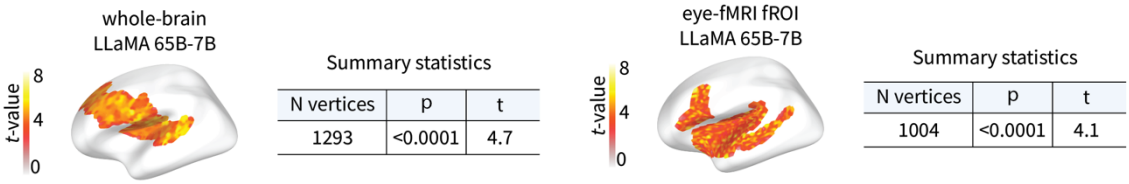
Following on the previous comment, I have questions about the rationale for two analytic choices in the fMRI analysis. First, the authors use the eye-tracking data to identify a cluster of voxels, then examine model-brain alignment for voxels in this cluster? Is this correct? I understand this to be the case because it looks like the maps you obtain in Fig. 4D are a subset of the voxels in Fig. 4A. If this is the case, I think we need a much stronger motivation for constraining the analysis in this way. If you did not perform the analysis within the ROI, then (in line with the previous comment), I am even more suspicious about the resulting brain maps. In either case, I think the authors should either try performing the model-brain alignment analysis for all voxels across the entire brain, or at least try using a more standard set of language areas (e.g., Lipkin et al., 2022).

RE11: Yes, we used the ROI from the eye-brain regression analysis, as explicitly stated in the "Alignment between LLMs and fMRI data" section of the Methods (p. 12): "After aligning the eye-tracking and fMRI data, we extracted the significant clusters as our fROI." Additionally, we indicated the fROI on all the brain surfaces in Fig. 4c (light blue). We have now added this in the caption of Fig.4. Our rationale for using the fROI is that we constructed the fMRI data matrices using the same method as the regressive eye saccades matrices, both of which capture word-to-word relationships rather than linear sequences, similar to attention weights in LLMs. We believe it is appropriate to examine model-brain alignment within regions that have already demonstrated significant model-behavior alignment. Additionally, we find that this fROI

largely overlaps with the language network reported in the literature (e.g., Lipkin et al., 2022; Malik-Moraleda et al., 2022).

We also conducted the same analysis across all brain voxels, as suggested, for LLaMA 7B and LLaMA 65B. The significant clusters from the contrast R^2 maps are displayed below. Notably, the significant voxels identified in the whole-brain analysis largely overlap with our fROI. Given this alignment, we opted to restrict our analyses to the fROI, as we believe that eye saccade patterns and brain signals for word-to-word relationship within a sentence should be correlated. We have added the motivation for our analyses in the "Alignment between LLMs and fMRI data" section in Methods on p.13 of the revised manuscript. The results from our additional whole-brain analysis are added to Supplementary Fig. 1a.

a Significant clusters from the contrasts of regression results of LLMs against whole-brain and fROI data



Summary statistics

N vertices	p	t
1293	<0.0001	4.7

Summary statistics

N vertices	p	t
1004	<0.0001	4.1

Second, I'm not really sure that I understand why the authors formulate their model-based fMRI encoding analysis the way they do. Most work evaluating LLM representations with fMRI data use ridge regression to map the model embeddings (at a given layer) onto voxelwise activity time series (e.g., Schrimpf et al., 2021; cf. Kumar, Sumers et al., 2024). If I understand correctly, rather than modeling voxel time series (or a sequence of words), the authors construct a word-by-word matrix of the sum(?) of activity corresponding to both words before and after a saccade for each voxel, then use ridge regression to predict the vectorized version of this matrix. Is this correct? Is the reason for doing this to manipulate the fMRI data into a format that resembles the fixation transitions and/or attention matrices? Or does self-paced reading task demand this kind of design somehow? Is there any precedent for this approach in the field? What's the neurobiological motivation for summing the activity for a pair of words? How are you accounting for the hemodynamic lag? My point here is that if the authors are effectively inventing a novel way to construct voxelwise encoding models, we'll need a bit more motivation and detailed explanation.

RE12: Yes most prior model-brain alignment studies regressed embeddings at each model layer onto voxel-wise activity time series (e.g., Gao et al., 2024; Huth et al., 2016; Kumar et al., 2024; Schrimpf et al., 2021). However, this method is not feasible for the current study due to the non-linear nature of reading (Note that our task is not self-paced reading at word level where each word appears on the screen sequentially, instead, we presented the whole sentence on the screen and relied on eye-tracking to identify the timepoints for each word). We cannot directly regress the embeddings for each sentence with the fMRI data because reading order is not strictly sequential. For example, our first participant read the sentence "Could humans live on Mars some day" in the following order based on their eye fixations: "humans humans Could humans on Mars on some some day". We could not simply input this disordered sequence into the LLM, as it would generate meaningless representations. Similarly, we cannot directly regress embeddings from the correctly ordered sentence onto the fMRI data, as the recorded neural responses correspond to the actual reading sequence, which does not follow the original sentence structure.

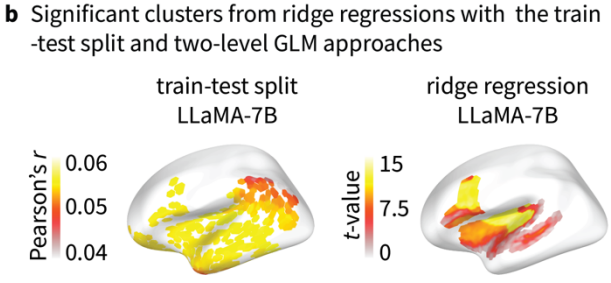
We transformed the fMRI data into a format that mirrors regressive eye saccade numbers. We summed fMRI activity for word pairs where there are more regressive saccades from the latter word to the former word, ensuring that our fMRI data matrices capture word-

to-word relationships within a sentence. To account for the hemodynamic lag, we added five scans to each word's offset. Note that our TR of 400 ms based on multiband scanning protocols (Hsu et al., 2019) is relatively short compared to typical fMRI studies, resulting in fewer words overlapping within a single scan. To the best of our knowledge, no prior studies have employed this approach, likely because most previous research relied on naturalistic listening or self-paced reading paradigms at the word level, which inherently enforce sequential word processing. We hope our rationale is now clearer and that future studies incorporating concurrent eye-tracking and fMRI will consider applying our methods. We have added motivation for our approach in the section "Alignment between LLMs and fMRI data" in Methods on p.13 of the revised manuscript.

I don't see any mention of training/test sets or cross-validation in the description of the model-brain alignment analyses. All prior work that I'm aware of on voxelwise encoding models uses cross-validation of some kind (e.g., splitting the stimuli into train and test sets; Huth et al., 2016; Schrimpf et al., 2021; Kumar, Sumers et al., 2024). If I understand correctly, the attention matrices used to predict the neural activity will be quite wide relative to the number of "samples" (vectorized off-diagonal word pairs)—this is presumably why the authors are using ridge regression rather than ordinary least-squares. But simply using ridge regression won't protect against overfitting all on its own—you absolutely need to use some kind of training/test split. (Relatedly, how do the authors select the penalty hyperparameter for ridge without some kind of grid search/nested cross-validation?)

RE13: Yes, we used ridge regression because our regressor consists of 7388*N columns, where N is the number of the LLM's attention heads. Our method follows a two-level general linear model (GLM) approach, which has been widely applied in fMRI studies for decades. The primary modification we made was replacing linear regression with ridge regression and conducting second-level statistical tests on the z-transformed R^2 maps instead of beta maps from the first-level GLM. The second-level statistical tests identify voxels that are better explained by the regressors across participants. Since it is unlikely that ridge regression would cause overfitting at the same voxels in all participants, we believe that this two-level GLM approach is as valid as the train-test split method utilized in prior studies (Huth et al., 2016; Kumar et al., 2024; Schrimpf et al., 2021).

To verify this, we conducted the same regression analyses for LLaMA 7B using the train-test split method, where we used 90% of the fMRI data to fit the ridge regression and then computed the correlation between the predicted and observed time courses for the remaining 10% of the data. The p-value for each voxel's correlation coefficient (r) was obtained by permuting the predicted time course 10,000 times and comparing the observed r against the distribution of r-values from the permutations. Our results showed that the train-test split approach identified a broader frontal-temporal region, whereas our two-level GLM approach also revealed significant model-brain alignment within this network. Given that our method produced more conservative results, we opted to retain the findings from our original two-level GLM approach. We have added this information on p.13 of the revised manuscript. The additional results from the train-test split approach are shown in Supplementary Fig. 1b.



Regarding the penalty hyperparameter, we retained the default value of 1, as we believe that keeping the encoding model simple facilitates a more interpretable model-brain alignment. While adjusting the penalty parameter for each voxel or employing more complex, non-linear models could potentially capture more nuanced model-brain relationships, it may also reduce generalizability across participants and datasets. Additionally, it is unclear what the motivation would be for assuming different regularization strengths at different voxels for each participant. We have added the explanation on in the "Alignment between LLMs and eye movement" section in Methods on p.12 of the revised manuscript.

One more point related to the previous two comments: Have the authors tried simply comparing how well the embeddings from models with and without instruction tuning predict voxelwise time series (or voxelwise sequence of words). This kind of analysis would make closer contact with prior work and was the analysis I was expecting when I read the title of the manuscript. In general, I would try to align the fMRI analysis more closely with prior work so that readers have more schema in place to understand what you're doing.

RE14: As previously mentioned, we could not directly regress the embeddings for each sentence with the fMRI data due to the non-sequential nature of reading order. However, in response to R1's suggestion, we have now incorporated results from an fMRI dataset of naturalistic listening (Wang et al., 2025). To ensure a fair comparison, we continued to use the attention matrices from the LLMs and constructed a representational dissimilarity matrix (RDM) for each sentence of the speech stimuli using searchlight representation similarity analysis (RSA; Kriegeskorte et al., 2008). Details on the dataset and methods can be found in "Additional results from fMRI data of naturalistic listening" in the Supplementary information. The RSA method has been widely applied in the literature for model-brain alignment in both language (Yu et al., 2024) and vision research (Xie et al., 2022; Yamins et al., 2014). Our findings remained consistent, showing a significant effect of scaling on model-brain alignment, while fine-tuned and base models of the same size exhibited no significant difference in brain encoding performance (see Supplementary Fig. 2a and Supplementary Tables 12–13).

Minor comments:

"For instance, Ouyang et al. (2022) fine-tuned GPT-3 using reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020), and showed that the finetuned models were more aligned with human preferences than GPT-3, despite GPT-3's much larger model size." I don't understand this sentence; sounds like you're saying "GPT-3 is larger than GPT-3".

RE15: Thank you for pointing this out. We have now revised the sentence to be "For instance, (Ouyang et al., 2022) fine-tuned GPT-3 of varying sizes using reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020), and showed that the finetuned models with only 1.3B parameters were more aligned with human preferences than the 175B base GPT-3." on p.3 of the manuscript.

Figure 2B: In the rightmost plot, I'm not sure I understand why you have four lines total in the legend (or maybe I just can't tell the colors apart for Alpaca versus Vicuna).

RE16: The 4 lines represent (1) the divergence between Alpaca 7B and LLaMA 7B; (2) the divergence between Alpaca 13B and LLaMA 13B; (3) the divergence between Vicuna 7B and LLaMA 7B; (4) the divergence between Vicuna 13B and LLaMA 13B. Lighter shades of red and orange correspond to the 7B models, while darker shades represent the 13B models. We have revised the legend and adjusted the color coding to enhance the distinction between Alpaca and Vicuna across all plots.

Figure 3A: Can you add a legend to the figure indicating what the two groups of lines correspond to?

RE17: The two groups of lines represent the regression scores from the real models (solid lines) and their controls (models of the same size with random weights, dashed lines). Due to the addition of 5 more models as requested by R1 and R3, the regression plots across layers became overly cluttered, so we removed these plots. Moreover, the regression scores from individual layers do not provide additional insights beyond those from the best-performing layer, which is the primary finding of this study.

As far as I understand, MNE is typically used for EEG/MEG analyses. Does MNE's cluster correction software work properly on fMRI data? Does it properly take into the account the 3D (or 2D on the surface) spatial structure of voxels?

RE18: Yes MNE is typically used for EEG/MEG analyses. Here is how we applied MNE to the fMRI results: We first transformed the volume-based fMRI data into surface-based data (see p.8: "We further projected the volume-based data for each run onto a brain surface using nilearn's vol_to_surf function with default parameters (Abraham et al., 2014) and a 'fsaverage5' template (Fischl, 2012). Next, we performed cluster analyses on the R^2 map of the surface fMRI data, treating the R^2 values from all model layers as "timepoints." (i.e., results from 32 layers is like MEG data with 32 timepoints). Since it remains unclear whether adjacent model layers perform similar functions, we set the temporal clustering threshold to one, ensuring clustering occurs only in space, not across time. This transformation resulted in a data format identical to MEG data, allowing us to analyze it using MNE. We have now added this explanation in the section "Alignment between LLMs and fMRI data" in Methods on p.13 of the manuscript.

"given that GPT-2 already predicts neural responses more accurately than an average human"—I'm not sure this is really true; GPT-2 gets close to the noise ceiling for the simple sentence stimuli in the Pereira2018 and Fedorenko2016 datasets, but it doesn't get even close to the noise ceiling for naturalistic stories in the Blank2014 dataset. In Kumar, Sumers et al. (2024, Fig. 2, S1, S4), models like GPT-2 only reach ~30% of the noise ceiling at best.

RE19: Thank you for pointing this out. We have now revised the sentence on p.7 of the revised manuscript: "As it has been shown that GPT-2 predicts more variance relative to the ceiling in some neuroimaging datasets (Fedorenko et al., 2016; Pereira et al., 2018)—defined as the mean of the neural responses from all participants in these datasets."

"our results suggest that human language processing in a naturalistic setting is not sensitive to instructions"—the logic here seems a bit odd, since the humans aren't really performing an "instruction-following" task; consider re-wording this.

RE20: Yes we agree and we have deleted this sentence.

We have a recent paper on scaling that extends the fMRI results from Antonello et al., 2023, to ECoG data—we see similar scaling effects: Hong, Wang et al., 2024.

RE21: We thank you for referring us to this important paper, which we have now added in the Introduction (p.2) and Discussion (p.7) of the revised manuscript.

References:

Antonello, R., Vaidya, A., & Huth, A. (2024). Scaling laws for language encoding models in fMRI. Advances in Neural Information Processing Systems, 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/4533e4a352440a32558c1c227602c323-Abstract-Conference.html

Hong, Z.*, Wang, H.*, Zada, Z., Gazula, H., Turner, D., Aubrey, B., ... & Goldstein, A. (2024). Scale matters: large language models with billions (rather than millions) of parameters better match neural representations of natural language. eLife. https://doi.org/10.7554/eLife.101204.1

Hsu, C. T., Clariana, R., Schloss, B., & Li, P. (2019). Neurocognitive signatures of naturalistic reading of scientific texts: a fixation-related fMRI study. Scientific Reports, 9, 10678. https://doi.org/10.1038/s41598-019-47176-7

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458. https://doi.org/10.1038/nature17637

Kumar, S.*, Sumers, T. R.*, Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., ... & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. Nature Communications, 15, 5523. https://doi.org/10.1038/s41467-024-49173-5

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., ... & Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from> 800 individuals. Scientific Data, 9, 529. https://doi.org/10.1038/s41597-022-01645-3

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Signed: Samuel A. Nastase

Reviewer #3 (Remarks to the Author):

Significance.

On first read, the reported results seem like a replication of previous studies on the alignment between LM representations and brain data (Schrimpf et al. 2021 and Antonello et al. 2023 cited in the manuscript, and also Hong et al. 2024 suggested below). Since the data is from a pre-existing study, I guess the new contribution here is the evaluation of instruction tuning and showing that "the scaling law of model-brain alignment holds even with shorter text stimuli and smaller fMRI data," which isn't very exciting to me as it currently stands. I think the authors should aim to bring out their contributions through a deeper discussion in light of previous 'brain scaling law' results.

RE22: We thank the reviewer for pointing out that the additional contribution of this work lies in (a) contrasting scaling with instruction tuning of LLMs, and (b) even with shorter text and smaller fMRI data the scaling low of model-brain alignment holds. We have revised our Introduction (p.3) and Discussion (pp. 7-8) along the following lines: Our results highlight the greater impact of scaling over instruction tuning in model-brain alignment, contributing to the existing literature on the scaling law in brain encoding performance (Antonello et al., 2023; Hong et al., 2024; Schrimpf et al., 2021). In addition, the use of a smaller brain data in our study suggests a robust effect of scaling vs. fine-tuning for model-brain alignment. These findings are particularly relevant in the current landscape, where reasoning LLMs such as DeepSeek-R1 (DeepSeek-AI et al., 2025) achieve state-of-the-art performance by using similar or fewer activated parameters than existing open-source LLMs. Such models' efficiency may

rest on their ability to integrate chain-of-thought reasoning with reinforcement learning during fine-tuning. However, our results do not support their roles in predicting brain responses during language comprehension. We acknowledge that caution is needed when interpreting these results. Since instruction tuning effectively realigned the model weights in response to instructions, these realigned model weights may better fit brain activity patterns where participants performed tasks aligned with the instruction-following nature of the fine-tuning process. Yet, due to the lack of such openly available neuroimaging datasets, we cannot evaluated the fine-tuned LLMs on these task-specific brain data.

Data and methodology.

As the data is from a pre-existing publication, I assume the collection and post-processing procedures of the data analyzed in the manuscript have already been validated. It might be more helpful to readers if the authors could include the exact source of the text stimuli (currently it only says "five English STEM articles").

RE23: The text stimuli were constructed using materials from established sources, including the NASA science website, the GPS.gov website (http://www.gps.gov), and Wikipedia. These texts underwent an extensive revision process to ensure content accuracy and stylistic consistency (see Follmer et al., 2018). We have added the information on p.8 of the manuscript.

Analytical approach.

I found many aspects of the modeling decisions not clearly justified, and therefore the modeling protocols don't seem as solid as they should be. Below are my concerns:

1) What justifies the use of attention weights as a predictor of BOLD signal and/or measure of processing difficulty? To my knowledge, most computational neuroscience work have used word-by-word surprisal or the raw vectors as predictors. The use of (unidirectional) attention weights seems to have resulted in other modeling decisions that seem pretty unusual to me.

RE24: Yes R2 has brought up the same concern, and here is our rationale: While most prior model-brain alignment studies regressed embeddings at each model layer onto voxel-wise activity time series (e.g., Gao et al., 2024; Huth et al., 2016; Kumar et al., 2024; Schrimpf et al., 2021), this method is not feasible for the current study due to the non-linear nature of reading (Note that our task is not self-paced reading at word level where each word appears on the screen sequentially, instead, we presented the whole sentence on the screen and relied on eye-tracking to identify the timepoints for each word). We cannot directly regress the embeddings for each sentence with the fMRI data because reading order is not strictly sequential. For example, our first participant read the sentence "Could humans live on Mars some day" in the following order based on their eye fixations: "humans humans Could humans on Mars on some some day". We could not simply input this disordered sequence into the LLM, as it would generate meaningless representations. Similarly, we cannot directly regress embeddings from the correctly ordered sentence onto the fMRI data, as the recorded neural responses correspond to the actual reading sequence, which does not follow the original sentence structure.

We transformed the fMRI data into a format that mirrors regressive eye saccade numbers. We summed fMRI activity for word pairs where there are more regressive saccades from the latter word to the former word, ensuring that our fMRI data matrices capture word-to-word relationships within a sentence. To account for the hemodynamic lag, we added five scans to each word's offset. Note that our TR of 400 ms due to the multiband scanning protocols (Hsu et al., 2019) is relatively short compared to typical fMRI studies, resulting in fewer words overlapping within a single scan. To the best of our knowledge, no prior studies have employed this approach, likely because most previous research relied on naturalistic listening or self-paced reading paradigms at the word level, which inherently enforce sequential word

processing. We hope our rationale is now clearer and that future studies incorporating concurrent eye-tracking and fMRI will consider applying our methods. We have added motivation for our approach in the section "Alignment between LLMs and fMRI data" in Methods on p.13 of the revised manuscript.

2) The first of these has to do with how the dependent variables are filtered. Both the eye movement and fMRI data seem to be based on those for right-to-left eye regression only. I guess this results in a match in the shape of the dependent variables and the attention weights (a lower triangular matrix), but does this mean the data associated with forward saccades are thrown out completely? To what extent is this decision justified?

RE25: Yes, we did not include forward saccades, not only due to the unidirectional nature of LLMs but also because regressive saccades may carry more informative value in reading. Regressive saccades occur when readers revisit earlier text, emphasizing the importance of previous words in comprehending the current word (Liversedge & Findlay, 2000). We believe this process is more analogous to attention weights in LLMs, which also prioritize contextual dependencies between words. We have now added this information on p.5 of the revised manuscript.

3) The authors also mention that attention weights can show "trivial patterns." Putting attention on the first token is perhaps uninterpretable, but why is attending to the immediately previous word or the current word a trivial pattern? Maybe this further suggests that attention weights are not the most appropriate predictors to model fMRI data to begin with.

RE26: We consider the model's tendencies to attend to the immediately preceding or current word "trivial patterns" because these behaviours are exhibited by all LLMs. As a result, it is not relevant to the effects of scaling or fine-tuning on LLMs' brain encoding performance, which is the primary focus of this study. We have added this explanation on p.4 of the manuscript.

4) I like the authors' efforts to ensure a fair comparison between LLMs by controlling for the number of attention heads. However, it's currently unclear whether the reported results on eye movement alignment control for the different number of attention heads (cf. Section on 'Alignment between LLMs and fMRI data' says "We then subtracted the regression scores LLMs with random weights from the resulting regression scores of corresponding LLMs at all the voxels for each subject, ...").

RE27: Yes we controlled for the number of attention heads both in the regression for eye movement patterns and the regression for the fMRI data. We have added this information in the Sections "Alignment between LLMs and eye movement" and "Alignment between LLMs and fMRI data" in Methods on pp.11-12 of the manuscript.

5) While the authors aimed to control for the number of attention heads, they don't seem to have controlled for the number of layers. Admittedly, I'm not sure how this might be done cleanly, but it seems important to do so as the core 'scaling results' are based on the best performing layer for both the eye movement and fMRI data.

RE28: We included all attention heads in our regression model for each layer, and larger models had more regressors. To control for the number of layers, we included a model of matching size for each LLM with randomized weights to control for both the numbers of attention heads and layers. The regression scores of the control models were subtracted from the corresponding LLM regression scores at all voxels for each subject. We have added explanation of this in the Sections "Alignment between LLMs and eye movement" and "Alignment between LLMs and fMRI data" on pp.11-13 of the manuscript.

6) On that note, Figures 3 and 4 seem a bit strange. First, the stars in Figures 3B and 4C seem to be miscolored. Additionally, none of the layers seem to exceed a normalized r-squared of 0.6 in Figure 3A, but the orange star in Figure 3B seems to exceed it.

RE29: Thank you for pointing these out! We apologize for the oversight on the color of the stars, as well as the y-axis labeling. We have now revised Fig. 3 and 4 accordingly.

7) The next-token prediction loss is not comparable across LLMs if they have different tokenizers. Instead, I would recommend reporting the average word-level surprisal or perplexity (i.e. exponentiated average word-level surprisal), which is not subject to this confound.

RE30: Thank you for pointing this out. We have now computed the mean next-word prediction (NWP) loss instead of next-token prediction (NTP) loss. We used the same method to align words and tokens as we did for the attention matrices: We aligned subwords to words by summing over the "to" tokens and averaging over the "from" tokens in a split word, as suggested by Clark et al. (2019) and Manning et al. (2020). For example, suppose the phrase "delicious cupcake" is tokenized as "del icious cup cakes", the attention score from "cupcake" to "delicious" is the sum of the attention scores from "del" to "cup" and "cake" and "icious" to "cup" and "cake", divided by 2 as there are to "to" tokens ("cup" and "cake"). We also removed the special tokens "<s>" at sentence beginning positions. We have now revised all occurrence of "NTP loss" to "NWP loss" and have added the information in the "Comparison of next-word prediction loss" section in Methods on p.10 of the manuscript. The NWP loss exhibited a trend where, for the base models, an increase in model size corresponded to a decrease in mean NWP loss. However, fine-tuned models did not improve performance on NWP for our test stimuli (see Fig. 2a and Supplementary Table 1 for the mean NWP loss for each model. Supplementary Table 2 shows the *t*-test statistics between all model pairs).

8) I think including modeling results from other GPT-2 variants (i.e. small, medium, XL) would greatly strengthen the evidence for the scaling behavior reported in this study (and be relatively lightweight to conduct).

RE31: We have now added results from GPT-2-base (124M), medium (355M), and xlarge (1.5B). We did not observe any significant differences in the fit between these models and the eye-regression patterns or fMRI patterns (see Fig. 3a and Fig. 4b). This may be because the size differences among these models are not as substantial as, for example, between 7B and 65B. We have discussed it on p.5 of the manuscript.

Suggested improvements.

Addressing the points raised in the previous section, either by including a discussion of the rationales underlying modeling decisions or re-conducting some of the experiments, will greatly improve the quality of the manuscript.

RE32: We have followed the suggestions and addressed the points raised above by including a discussion on the rationales behind our modeling decisions and re-running the experiments using all models from the GPT-2 family.

Clarity and context.

The prose is generally clear, but I have a high-level suggestion to further improve clarity. The Introduction section situates the current work by outlining studies that examine the influence of LLM size on various tasks. Currently, the authors appear to be using the term "performance" to mean many different things, which may be confusing to readers who aren't familiar with all these studies. I would recommend specifying performance on natural language processing tasks,

natural language generation tasks, modeling brain imagery, modeling human reading times, etc. where applicable.
RE33: Thank you for the suggestion! We have now added explanations for all instances of "performance" in the Introduction on pp.2-3 of the manuscript.

References.
The following three groups of studies seem relevant to the current manuscript.
1) These studies have shown that surprisal from larger LLMs yield poorer fits to word-by-word human reading times: Oh and Schuler (2023, "Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?"); Shain, Meister, Pimentel, and Levy (2024, "Large-scale evidence for logarithmic effects of word predictability on reading time").
RE44: Yes these two studies are indeed very relevant. We have now added them in the Introduction on p.2 of the revised manuscript.

2) This study shows instruction tuning LLMs does not improve surprisal's fit to reading times: Kuribayashi, Oseki, and Baldwin (2024, "Psychometric predictive power of large language models").
RE35: Yes R4 also mentioned this study and we have now added it in the Discussion on p.8 of the revised manuscript.

3) This concurrent preprint also demonstrates the 'brain scaling law:' Hong et al. (2024, "Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language").
RE36: Yes these references are indeed highly relevant, as also noted by R2. We have now incorporated these references on p.7 in the Discussion of the revised manuscript.

Your (the reviewer's) expertise.
I do not have expertise in the collection, validation, and pre-processing of fMRI data, but since this manuscript utilizes publicly available data from a previously published study, I think these aspects are less relevant for evaluating this manuscript.

Reviewer #3 (Remarks on code availability):
I only checked whether the repository was empty or not. But based on the methodology described in the manuscript, the modeling experiments seem very straightforward to replicate in any case.

Reviewer #4 (Remarks to the Author):
This paper demonstrates that the scaling law holds for alignment between LLMs and human behavioral and neural data, while instruction-tuning is inert for that purpose. I'm generally positive for endorsing this paper for publication, but the following major and minor comments should be addressed before publication.

Major comments:
- This paper states "our work is the first to systematically investigate the impact of both scaling and finetuning on the alignment...", but this is an overstatement because the previous literature have already investigated the impact of scaling/instruction tuning on the alignment with human language processing. For example, Kuribayashi et al. (2024) "Psychometric Predictive Power of Large Language Models" has shown that both scaling (as measured in perplexity) and instruction tuning have negative impacts on the alignment with human behavior. Please relax

this statement (i.e. the impact of scaling/instruction tuning on the alignment with human neural activity may be new) and, more importantly, discuss how to reconcile the results of this paper and Kuribayashi et al. (2024).

RE37: We have now relaxed the statement in the Introduction on p.3 of the revised manuscript. Compared to Kuribayashi et al. (2024), our study investigates a different set of base and fine-tuned models and incorporates both behavioral and neuroimaging data, yet both studies report a negative effect of instruction tuning on alignment with human language processing. The positive effect observed in Falcon IT-LLMs by Kuribayashi et al. (2024) is particularly interesting, suggesting that different fine-tuning techniques may play a role. We have incorporated this discussion on p.8 in the Discussion section of the revised manuscript.

- This paper sometimes performs statistical analyses on "every pair of models", but I don't think that any comparisons except the ones between models of the same sizes are meaningful. Please justify the reason why the comparisons of "every pair of models" is necessary.

RE38: Comparisons between models of different sizes are meaningful, as our goal is to demonstrate that larger models exhibit better brain encoding, which is a key conclusion of our study. We understand that you may find this comparison less meaningful due to differences in the number of regressors in the encoding model, as also noted by R3. However, we accounted for model size by including a randomized-weight control model of matching size for each LLM, ensuring control over both the number of attention heads and layers. We subtracted the regression scores of the control models from the corresponding LLM regression scores at all voxels for each subject. We have added an explanation of this approach in the “Alignment between LLMs and Eye Movement” and “Alignment between LLMs and fMRI Data” sections on p.12 of the revised manuscript.

- The result that the D_JS of model attentions was significantly larger only between LLaMA-13B and Vicuna-13B is interesting, which may be attributed to the difference between Vicuna and Alpaca in how those models are instruction-tuned. Please explain how differently Vicuna and Alpaca are fine-tuned, and discuss the results from that perspective.

RE39: Alpaca was fine-tuned on instruction-following examples, with its training focused on ensuring the model reliably adheres to given instructions (Taori et al., 2023). This approach generally results in strong performance on single-turn tasks. In contrast, Vicuna was fine-tuned using conversational data, incorporating multi-turn dialogues that capture diverse conversational contexts (Chiang et al., 2023). Consequently, Vicuna offers a more natural and context-aware dialogue experience compared to Alpaca, which may contribute to the greater divergence observed between Vicuna 13B and LLaMA 13B. However, we caution against overinterpreting this result, as neither Vicuna 13B nor Alpaca 13B exhibited improved model-eye or model-brain alignment in our dataset, suggesting that their instruction-tuning techniques may negatively impact brain encoding performance. We have added these discussion on p.4 of the manuscript.

- The results of this paper are based on the models after "we subtracted these patterns from all the attention matrices of the LLMs for the following analyses", which seems to be arbitrary and not usual in the literature. Please justify this design choice.

RE40: We consider the model’s tendencies to attend to the immediately preceding or current word “trivial patterns” because these behaviours are exhibited by all LLMs. As a result, it is not relevant to the effects of scaling or fine-tuning on LLMs’ brain encoding performance, which is the primary focus of this study. We subtracted these patterns from all the attention matrices of the LLMs for the following ridge regression analyses given that similar patterns were not observed in human eye movement data, we believe they do not reflect underlying

human cognitive processes. We have added a justification of this design choice in the "Model sensitivity to trivial patterns" section in Methods on p.11 of the revised manuscript.

- This paper performs regression analyses on human behavior and neural activity with "the attention matrices of the LLMs", but how matrices (i.e. model attentions) can predict vectors (i.e. human behavior and neural activity) is not clear. While the details seem to be described in the Methods section, please explain the details briefly in the Results section as well.

RE41: Thank you for the suggestion. We have now added a detailed description in the Results on p.5 of the manuscript, in response to R2's request as well: We extracted the lower-triangle portions (excluding the diagonal line) of the attention matrix $n_{word} \times n_{word} \times n_{head}$ from all attention heads for every sentence. The attention matrices for all sentences were concatenated to create a regressor with dimensions $7388 \times n_{head}$ for each layer, where 7388 represents the total number of elements obtained after concatenating the lower triangles of the attention matrices across all sentences in our stimuli. For the human eye saccade data, we constructed matrices for saccade number, $E_{num} \in \mathbb{R}^{n_{word} \times n_{word}}$, for each sentence. Each cell at row l and column m in E_{num} and E_{dur} represents the number of eye fixation moving from the word in row l to the word in column m, respectively. We then extracted the lower-triangle parts of the matrices which marks right-to-left eye movement. Similar to the models' attention matrices, we flattened the regressive eye saccade number matrices for all sentences and concatenated them to create 7388-length vectors for each subject. We then performed ridge regression for each model layer, using the $7388 \times n_{head}$ regressor to predict each subject's regressive eye saccade number vectors. The final R^2_{model} was normalized by the $R^2_{ceiling}$, where the ceiling model represents the mean of all subjects' regressive eye saccade number vectors.

- This paper defined fROI based on "the eye movement regression number patterns", which is not a standard practice in the literature to my knowledge. Please justify this design choice.

RE42: Yes we understand that this is not a standard practice and R2 raises a similar concern. Our rationale for using the fROI is that we constructed the fMRI data matrices using the same method as the regressive eye saccades matrices, both of which capture word-to-word relationships rather than linear sequences, like attention weights in LLMs. We believe it is appropriate to examine model-brain alignment within regions that have already demonstrated significant model-behavior alignment. Additionally, we find that this fROI largely overlaps with the language network reported in the literature (e.g., Lipkin et al., 2022; Malik-Moraleda et al., 2022).

We also conducted the same analysis across all brain voxels, as suggested, for LLaMA 7B and LLaMA 65B. The significant clusters from the contrast R^2 maps are displayed in Supplementary Fig. 1a. Notably, the significant voxels identified in the whole-brain analysis largely overlap with our fROI. Given this alignment, we opted to restrict our analyses to the fROI, as we believe that eye saccade patterns and brain signals for word-to-word relationship within a sentence should be correlated. We have added the motivation for our analyses on p.6 of the revised manuscript. The results from our additional whole-brain analysis are added to Supplementary Fig. 1a.

Minor comments:

- The notations are inconsistent for the GPT family (e.g. GPT-2 vs. GPT4).

RE43: Thank you for pointing this out. We have now removed all hyphens from the model names of the GPT family.

- What does the sentence "GPT-2 already predicts neural responses more accurately than an average human" mean?

RE44: This refers to the findings from Schrimpf et al. (2021), which shows that GPT-2 xlarge predicts more variance relative to the ceiling in some neuroimaging datasets (Fedorenko et al., 2016; Pereira et al., 2018)—defined as the mean of the neural responses from all participants in these datasets. We have revised the sentence on p.7 of the manuscript accordingly.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., … Zhou, X. (2024). *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone* (arXiv:2404.14219). arXiv.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*.

Antonello, R., Vaidya, A., & Huth, A. (2023). Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, *36*, 21895–21907.

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., Resnik, P., & Simon, J. Z. (2023). Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *eLife*, *12*, e85012.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, *30*.

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

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (arXiv:2501.12948). arXiv.

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781.

Follmer, D. J., Fang, S.-Y., Clariana, R. B., Meyer, B. J. F., & Li, P. (2018). What predicts adult readers' understanding of STEM texts? *Reading and Writing*, *31*(1), 185–214.

Gao, C., Huang, S., Li, J., & Chen, J. (2023). Roles of scaling and instruction tuning in language perception: Model vs. human attention. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 13042–13055). Association for Computational Linguistics.

Gao, C., Li, J., Chen, J., & Huang, S. (2024). Measuring meaning composition in the human brain with composition scores from large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11295–11308.

Gemma Team. (2024). *Gemma: Open models based on Gemini research and technology* (arXiv:2403.08295). arXiv.

Hong, Z., Wang, H., Zada, Z., Gazula, H., Turner, D., Aubrey, B., Niekerken, L., Doyle, W., Devore, S., Dugan, P., Friedman, D., Devinsky, O., Flinker, A., Hasson, U., Nastase, S. A., & Goldstein, A. (2024). Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language. *eLife*, *13*.

Hsu, C.-T., Clariana, R., Schloss, B., & Li, P. (2019). Neurocognitive signatures of naturalistic reading of scientific texts: A fixation-related fmri study. *Scientific Reports*, *9*(1), Article 1.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (arXiv:2310.06825). arXiv.

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

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, *15*(1), 5523.

Li, J., Lai, M., & Pylkkänen, L. (2024). Semantic composition in experimental and naturalistic paradigms. *Imaging Neuroscience*, *2*, 1–17.

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., Siegelman, M., Hoeflin, C., Pongos, A., Blank, I. A., Struhl, M. K., Ivanova, A., Shannon, S., Sathe, A., Hoffmann, M., Nieto-Castañón, A., & Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, *9*(1), 529.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*(1), 6–14.

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), Article 8.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

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

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative

modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. Https://Crfm. Stanford. Edu/2023/03/13/Alpaca. Html*, *3*(6), 7.

Wang, Q., Zhou, Q., Ma, Z., Wang, N., Zhang, T., Fu, Y., & Li, J. (2025). *Le Petit Prince (LPP) Multi-talker: Naturalistic 7T fMRI and EEG Dataset* (p. 2025.02.26.640265). bioRxiv.

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M., & Cichy, R. M. (2022). Visual category representations in the infant brain. *Current Biology*, *32*(24), 5422-5432.e6.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yu, S., Gu, C., Huang, K., & Li, P. (2024). Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science Advances*, *10*(21), eadn7744.

Dear Editors and Reviewers,

Thank you for your letter of April 10, 2025 regarding our submission to Nature Computational Science (MS#: NATCOMPUTSCI-24-2037). We sincerely appreciate your insightful feedback and constructive suggestions. In this revised version, we have carefully addressed all the points you raised, and we hope that you and the reviewers will find this version suitable for publication in your journal. We have made several significant revisions, as outlined below (with the revised text highlighted for clarity).

Editor comments:

- Starting at line 447, when describing the regression models, please explicitly state the dimensionality of both the model (X) and the outcome variable (y).

We have now explicitly stated the dimensionality of both the model inputs (X) and the outcome variable (y) on pp. 11–12 of the revised manuscript.

- Are these models fit to one long vector of 7,388 samples, or are these fit for separate models for each sentence? How is overfitting avoided, particularly for the shortest sentences, making ridge regression more necessary.

We fit a single long vector of 7,388 samples. To avoid overfitting, we now use out-of-sample prediction with a 10-fold train-test split. Each voxel's regularization parameter was selected through a grid search with nested cross-validation across 20 candidate regularization coefficients, log-spaced between 10 and 1,000.

- Please use out-of-sample prediction (cross-validation) for the core, main-text analyses.
- In the response letter, it is mentioned that "two-level GLM approach" is used in typical fMRI analyses, but this approach is typically used to compare regression coefficients assigned to individual regressors in a regression model. This approach is not valid for comparing R-squared scores for models of varying dimensionality.
- The study normalizes R-squared values and use a one-sample one-tailed t-test with a cluster-based permutation test. However, the one-sample test is problematic as the R-squared values are positive.

All of these concerns have been addressed in our revised ridge regression analyses with whole-brain data using out-of-sample prediction. We have also updated the results with the same train-test split methods for the naturalistic listening data in the Supplementary Information.

- The method of using nilearn's vol_to_surf to resample volumetric data onto the cortical surface may have a low-quality surface projection due to the different cortical anatomy of the fsaverage(5) surface template, which is aligned to MNI space. Best practice is to resample individual-subject, native-space data onto subject-specific FreeSurface reconstruction.

We used Nilearn's vol_to_surf function to resample the volumetric data in MNI152 linear space onto the fsaverage5 surface, following previous practice (e.g., Millet et al., 2022). We have now reprocessed all subjects' raw fMRI data (including the naturalistic listening and task data in the Supplementary information) using fMRIPrep and resampled individual-subject, native-space data onto the subject-specific FreeSurfer surfaces.

- The negative impact of instruction tuning on alignment with human language processing seems to be consistent with the previous literature, but the impact of scaling is unexplored.

We have added the two references suggested by R4 and discussed the impact of scaling in the Introduction (p.1).

- Please discuss how it is concluded that "the significant voxels identified in the whole-brain analysis largely overlap with our fROI".
We initially based this conclusion on visual inspection, noting that significant clusters in both results largely overlapped with the language network. We have now removed this statement altogether, as we have rerun all the ridge regression analyses with a train-test split method, following R2's suggestions.

Reviewers comments:
Reviewer #1 (Remarks to the Author):
Thank you for your thoughtful and thorough revisions to the manuscript. I have carefully reviewed the updated version and am pleased to see that all the suggested changes and comments have been appropriately addressed. The revisions have significantly improved the clarity and overall quality of the paper. I have no further comments at this stage except the one that the codebase for the newly conducted experiments is still not added to the shared code repository.

Reviewer #1 (Remarks on code availability):
The codebase for the newly conducted experiments is not added to the shared code repository.
RE1: Thank you for pointing this out. The code for extracting the attention matrices for the newly added models (Gemma, Mistral, and GPT-2) remains the same as before. We have updated the README file to include a new section explaining how to load these models. Additionally, we have added codes for calculating LLMs' next-word prediction performance and for performing ridge regression with train-test split methods for the fMRI data.

Reviewer #2 (Remarks to the Author):
The authors have put considerable effort into improving the clarity of the manuscript and bolstering their findings with additional language models. I still have some concerns about the methodology that I try to unpack below.

Starting at line 447, when describing the regression models, can the authors please explicitly state the dimensionality of both the model (X) and the outcome variable (y)? The new paragraph added at line 140 in the Results is extremely helpful and at least that level of detail should also be present in the Methods (some redundancy is warranted in this case). I had initially thought the models were quite wide, but I think now I understand that the width of X corresponds to the number of heads at a given layer—ranging from 12 to 64? The authors need to be very clear about this. For example, in the response letter, you say you "used ridge regression because our regressor consists of 7388*N columns"—this could be interpreted as a wide model with thousands of columns (typical in fMRI encoding analyses), or a relatively narrow model with only n_heads columns. My understanding is that the latter interpretation is correct: the models have a number of columns equal to the number of heads in a given layer—please confirm.
RE2: Yes, the number of columns corresponds to the number of attention heads (ranging from 12 to 64) at a given layer. The dimensionality of the model input (X) is $7388 \times N_{att\ head}$, where 7388 represents the length of the concatenated vector formed by flattening the lower-triangular part of the attention matrix for each sentence at each layer, and N is the number of attention heads at that layer. We did not average across attention heads to obtain a single attention matrix per sentence, as each head is known to capture distinct relationships among words in a sentence (Manning et al., 2020). The dependent variable (y) is a 7388-dimensional vector, where 7388 again reflects the length of the concatenated and flattened lower-triangular part of the BOLD response matrix for each sentence. Since we constructed the BOLD matrix for each sentence

based on the regressive eye saccade patterns, each BOLD matrix of each sentence is of size $N_{word} \times N_{word}$, its dimension matches the dimension of model's attention matrix for each sentence. We have added more description of the dimensionality of the model (X) and the outcome variable (y) in the "Alignment between LLMs and eye movement" and the "Alignment between LLMs and fMRI data" sections in Methods on pp.11-12 of the revised manuscript. We have also updated Fig.1c to better visualize our regression model with the new train-test split method.

This raises a question: (1) do you fit these models to one long vector of 7,388 samples, or (2) do you fit separate models for each sentence? If the answer is 1, then I don't think you need to worry about ridge regression at all, OLS would probably be fine. If the answer is 2, then you might encounter overfitting, particularly for the shortest sentences, making ridge regression more necessary. It's important for readers to understand how the regression models are fit at each stage of analysis.

RE3: We fit a long vector of 7,388 samples $\times N_{att\_head}$ for each voxel. Since most models have 32 attention heads, with the maximum being 64, we believe that applying ridge regression is preferable to mitigate collinearity among the regressors and enhance prediction accuracy. We have added a more detailed explanation of this approach in the Methods section on p.11 of the revised manuscript.

Regarding ridge, there is no reason to think that including a ridge penalty (especially with an arbitrary value of 1) will adequately mitigate overfitting in a training set (i.e. when the R-squared is computed "in-sample"). The appropriate ridge parameter is non-trivially related to the structure of the data/model (e.g., variance, dimensionality) and is typically determined empirically; e.g. using grid search with nested cross-validation. The "correct" ridge penalty is the one that yields the best out-of-sample prediction performance. Note that, when fitting wide models to fMRI data, ridge penalties vary wildly (e.g., Huth et al., 2016, use a grid of ridge coefficients ranging from 10 to 1,000 and settle on coefficients on the scale of ~200).

RE4: As discussed in RE2 and RE3, our regression models are not wide, with the maximum number of regressors being 64. We have now revised all ridge regressions using a grid search with nested cross-validation across 20 candidate regularization coefficients (log-spaced between 10 and 1,000). The results revealed similar scaling effects, with larger models exhibiting higher activity in a broad network covering the temporal and parietal regions in both the left and right hemispheres (see Fig. 4). Although the effect size as measured by the Cohen's d is larger in the left hemisphere than the right hemisphere (see Table 2). No significant cluster was found for the contrasts between the finetuned and base LLMs of the same sizes. We have revised the results section on p.5 of the revised manuscript and also updated Supplementary Tables 10-11.

Whether the authors need to use ridge or not, I would strongly encourage the authors to use out-of-sample prediction (cross-validation) for their core, main-text analyses. The authors try to maneuver around this in several ways, but it ends up complicating the analysis pipeline quite a bit. In the response letter, the authors mention the "two-level GLM approach" used in typical fMRI analyses, but this approach is typically used to compare regression coefficients (not R-squared values!) assigned to individual regressors in a regression model. Respectfully, I do not think that this approach is valid for comparing R-squared scores for models of varying dimensionality. As the authors acknowledge (line 466), the problem is that bigger models (i.e., with a larger number of heads) will obtain trivially higher fits in a training set based on their dimensionality. They will overfit noise in the training set by definition, and this is equally true of the behavioral and fMRI analyses (see Yarkoni & Westfall, 2017, for a great discussion of

this issue). I made a little demo showing that R-squared increases monotonically with model dimensionality, even with completely random model features and target variables, and that using a ridge penalty in-sample does not really help at all:
https://github.com/snastase/tutorials/blob/main/overfitting.ipynb
The authors try to get around this problem by using a control (line 488): "we included a control model of matching size for each LLM, where the model weights were randomized to address this potential confound" and "we then subtracted the regression scores of randomly weighted LLMs from the corresponding LLM regression scores for each subject." To me, this raises more questions than it resolves. When you say the model weights were randomized, do you mean you used a randomly-initialized untrained language model? Or that you randomized the encoding weights and re-generated predictions? Note that the variance of possible R-squared values across random models also increases with increasing dimensionality, meaning that this subtraction method could yield larger differences for larger models based on the randomization alone (see the demo). After this normalization step, the authors then feed these R-squared values into a "one-sample one-tailed t-test with a cluster-based permutation test". It's not clear to me what is being permuted in this one-sample test (maybe sign-flipping?). There's an issue with feeding in-sample R-squared values into a one-tailed test, which is that the R-squared values will be positive by definition: a parametric test against mean zero will be trivially "significant". The authors try to get around this by z-scoring the R-squared values, which will forcibly mean-center them… but z-score them across what, sentences? If you forcibly mean-center the values (with z-score), I don't understand how you can then expect to find a significant effect of the mean being greater than zero (rejecting null hypothesis mean = 0) by means of any statistical test, parametric or otherwise.
RE5: We thank you very much for the helpful demo. We agree that model performance (R^2) tends to increase monotonically with model dimensionality. To account for this, we introduced a control method in which we subtracted the R^2 values obtained from control models of the same size (untrained models with randomly initialized weights). Additionally, we z-scored the R^2 values across voxels within each subject to identify voxels that consistently exhibited higher regression performance compared to all voxels' mean R^2 values across subjects.

We acknowledge that our original approach may have introduced certain issues. To address this, we have revised all ridge regression procedures to adopt a train-test split method, following Huth et al. (2016). Specifically, we adopted a 10-fold cross-validation, using 90% of the fMRI data (133 out of 148 sentences) to fit the ridge regression models and evaluating performance by computing the correlation between the predicted and observed fMRI time courses on the remaining 10% of the data (15 out of 148 sentences). For each voxel, a p-value for the correlation coefficient (r) was obtained by permuting the predicted time course 10,000 times and comparing the observed r to the distribution of permuted r values. The updated results continue to show similar scaling effects, with larger models exhibiting stronger activity within a broad temporal-parietal network bilaterally (see Fig. 4 and Table 2). We have updated the corresponding Methods section on p.12 of the revised manuscript.

Line 536: "Since it is unlikely that ridge regression would cause overfitting at the same voxels in all participants"—this should be removed; when evaluated in-sample, ridge regression will very likely yield overfitting in *all* voxels and participants.
RE6: We have now removed this sentence from the revised manuscript.

In the revision, they include a supplementary figure (Fig. S1b) where they use a train-test split; what kind of train-test split? Split-half? K-fold? I would use e.g. 5- or 10-fold splitting across sentences.

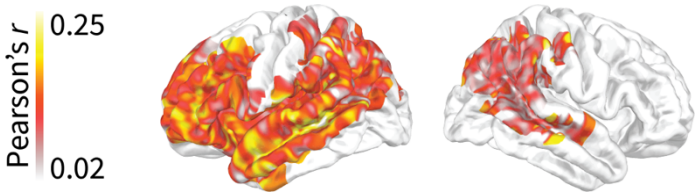
RE7: We used 10-fold splitting, as described on p.12 of the previous manuscript "where we used 90% of the fMRI data to fit the ridge regression and then computed the correlation between the predicted and observed time courses for the remaining 10% of the data." We have made this more explicit on p.12 of the revised manuscript.

This analysis yields a quite different-looking brain map than the in-sample model fits—why? To me, this brain map actually looks to better match the language network than the others (although it looks strangely spotty/patchy, see the comment on surface sampling below). I'm truly sorry to be annoying about this, but I am still very hung up on these brain visualizations. I understand that I might have different priors than the authors, but to my mind, there appears to be something wrong. The authors point to Li et al., 2024, for similar looking brain maps—but those maps are derived from MEG, not fMRI. (Really no right hemisphere results for any of these fMRI analyses?) I'm willing to concede that source-reconstructed MEG data may yield maps like this, but they do not seem typical for fMRI. The authors also cite Maris & Oostenveld, 2007, and Brodbeck et al., 2023, for their clustering analysis—but both of these toolboxes are designed for EEG/MEG data, not fMRI data. I trust the authors to coerce their fMRI data into a format that EEG/MEG software can accommodate (line 506), but this is one place I would have a close look when trying to figure out why the maps look so strange. I had a quick look through the code but couldn't find where the fMRI resampling, statistical clustering, etc is happening. For example, does MNE properly deal with the neighborhood structure of the surface vertices for computing clusters? Does it properly deal with the two separate hemispheres? Maybe using a simpler method of correction for multiple tests would be helpful here? Regardless of the cluster coordinate table the authors report in the response letter, the brain maps seem uncomfortably centered around the insula (especially Fig. S2b) and do not very closely reflect the functional anatomy we might expect for language processing (including Lipkin et al., 2022, and Malik-Moraleda et al., 2022, that the authors cite).

RE8: We have now adopted the train-test split methods with statistical maps derived from permuting the predicted time course 10,000 times and comparing the observed r to the distribution of permuted r values. The results still show that scaling improved the model-brain alignment whereas instruction-tuning had less effect (see the updated Fig.4, Table 2 and Supplementary Table 10-11).

Another way to build confidence here would be to run a simple "sanity check" analysis to ensure the brain maps look sensible. Here's an example that might work: For each subject, average the fMRI volumes across TRs within each sentence, resulting in an n_sentences-long vector per voxel. Now, for each voxel, compute the intersubject correlation of these sentence-level time series. If there's reliable variance across sentences—which could be due to interesting things like meaning, or uninteresting things like word rate—I would expect to see ISC in the typical language network.

RE9: We have conducted ISC analyses on our newly preprocessed surface data, following your suggestion. Statistical significance of the ISC value was determined by a bootstrap hypothesis test (Chen et al., 2016). The results show significant ISC values in the frontal-temporal-parietal regions bilaterally (see the figure below), which we think largely overlap with the language network (e.g., Malik-Moraleda et al., 2022). We believe the results from the newly preprocessed surface data are both consistent and interpretable.



Taking a closer look at the methods, another possible issue here is in using nilearn's vol_to_surf to resample volumetric data onto the cortical surface. The volumetric data are aligned to MNI space using SPM—but we need to know which MNI template; e.g., "MNI 152 linear", "MNI 152 nonlinear 6th generation", "MNI 152 nonlinear 2009"? (See the Lead-DBS website for details: https://www.lead-dbs.org/about-the-mni-spaces/). The potential issue here is that the fsaverage(5) surface template will have different cortical anatomy and is aligned to MNI 305 space (see, e.g., https://surfer.nmr.mgh.harvard.edu/fswiki/CoordinateSystems). This will result in a fairly low-quality surface projection. This is why best practice is to resample individual-subject, native-space data onto the subject-specific FreeSurface reconstruction (as in, e.g., fMRIPrep). You might be able to get a better volume-to-surface projection from neuromaps: https://netneurolab.github.io/neuromaps/generated/neuromaps.transforms.mni152_to_fsaverage.html#neuromaps.transforms.mni152_to_fsaverage

RE10: We thank you very much for pointing out these important references. We used Nilearn's vol_to_surf function to resample the volumetric data in MNI152 linear space onto the fsaverage5 surface, following previous practice (Millet et al., 2022). We have now reprocessed all subjects' raw fMRI data using fMRIPrep (v25.0.0) with default parameters and generated subject-specific FreeSurfer surfaces: Anatomical images were corrected for intensity non-uniformity (N4BiasFieldCorrection), skull-stripped using ANTs-based extraction (OASIS30ANTs template), and segmented into tissue classes using FSL's 'fast'. Brain surfaces were reconstructed using 'recon-all' (FreeSurfer 7.3.2). The T1-weighted images were normalized to MNI152NLin2009cAsym:res-2 space via nonlinear registration (ANTs). For each BOLD run, head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using 'mcflirt' (FSL 5.0.9) and slice timing correction was applied using 3dTshift (AFNI 20160207). The BOLD reference was then co-registered to the T1w reference using bbregister which implements boundary-based registration. Co-registration was configured with six degrees of freedom. No susceptibility distortion correction was applied. Confound regressors included motion parameters (and their derivatives/quadratics), framewise displacement (FD), DVARS, global signals, and t/aCompCor components computed from white matter and CSF after high-pass filtering (128s cutoff). Volumes exceeding FD>0.5 mm or standardized DVARS>1.5 were flagged as motion outliers. Final resampling to MNI space and fsaverage5 surface was performed in a single interpolation step using 'antsApplyTransforms' and 'mri_vol2surf'. We have added this information on p.8 of the revised manuscript.

I apologize for putting the authors through the ringer here, and I hope that this manuscript can be published soon—but I really think it will improve the impact of manuscript to make sure the modeling procedure and brain visualizations are as convincing as possible.

RE11: We sincerely appreciate your efforts in reviewing our paper and believe that your comments and suggestions have significantly strengthened the rigor of the study.

Reviewer #3 (Remarks to the Author):

I thank for authors for addressing all of my methodological concerns. Expanding the experiments to include additional models/datasets and re-writing parts of the prose further improved the quality of the manuscript. I agree with the other reviewers that this work makes a nice contribution in investigating the impact of scaling and instruction-tuning on brain-LLM alignment.

I have another set of minor suggestions:

1) I don't think it's ever stated explicitly that each sentence is provided separately as input to the LLMs (cf. providing the entire passage in one context window). I think this is an important modeling decision that should be mentioned, maybe in the "LLMs" paragraph under the "Methods" section.
RE12: Each sentence was input into the LLMs individually, consistent with how sentences were presented separately to participants on the screen during fMRI scanning. Furthermore, regressive eye saccade information was available only at the sentence level. We have added an explanation of this in the "Alignment between LLMs and eye movement" section in Methods on p.10 of the revised manuscript.

2) Lines 5, 38: Very minor, but "hundreds" could be replaced with "thousands" for more rhetorical effect.
RE13: We have revised them to be "thousands".

3) Line 42: In "language comprehension processing," "comprehension" and "processing" seem redundant.
RE14: Thank you very much for pointing this out! We have revised the wording to "language processing" on p.1 of the revised manuscript.

4) Lines 459-465: It seems like this was written to describe model-brain alignment, when the paragraph is about "Alignment between LLMs and eye movement."
RE15: Yes, this paragraph was indeed misplaced. We have removed it altogether, as we have now adopted a grid search method for selecting the regularization penalty coefficients, following R2's suggestion.

Reviewer #3 (Remarks on code availability):
I only checked whether the repository was empty or not. But based on the methodology described in the manuscript, the modeling experiments seem very straightforward to replicate in any case. The code was helpful for me to understand that the stimuli sentences were presented separately to the LLMs.

Reviewer #4 (Remarks to the Author):
All the comments have been sufficiently addressed, so that I would be happy to recommend this paper for publication provided that the following major and minor comments are resolved.

Major comments:
- The negative impact of instruction tuning on alignment with human language processing seems to be consistent with the previous literature, but how about scaling? The previous literature have reported the positive impact of scaling on alignment with human language processing (Kuribayashi et al. 2021, Oh & Schuler 2023, etc.).
RE16: We have carefully reviewed the two references, Millet et al., (2023) and Oh & Schuler (2023). It seems that both papers suggest that language models with the lowest perplexity do not necessarily produce the best fits to human reading times. We assume you are referring to the potential *negative* effects of scaling on human behavioral data? We have now incorporated these references into the Introduction, where we discuss prior literature on scaling effects for model-brain alignment, on p. 1 of the revised manuscript.

- Now I understand that fROI is defined from human eye movement to identify "regions that have already demonstrated significant model-behavior alignment", but how did you conclude

"the significant voxels identified in the whole-brain analysis largely overlap with our fROI"? Any quantification, or just visual inspection?
RE17: We have now removed this statement which was based on visual inspection. We have now rerun all analyses using whole brain data instead of fROI, following the suggestions of R2.

- The "trivial patterns" seem to be subtracted because the authors "believe they do not reflect underlying human cognitive processes", but in order to empirically test whether this belief is true, please perform the regression analyses without any subtraction of "trivial patterns" and show the results as supplementary information.
RE18: We agree and we have now rerun all the analyses with the original attention matrices. The results continued to demonstrate a scaling effect, with larger models showing higher model-brain alignment within the language network. However, no significant brain clusters were observed for the contrasts between the mid-sized LLaMA pairs 13B vs. 7B and 30B vs. 13B (see Supplementary Fig. 2 and Supplementary Tables 18–19). This may be due to smaller and mid-sized models being more susceptible to capturing trivial patterns, leading to similar brain responses. We have added these results on p.10 of the revised manuscript.

Minor comments:
- l.57: alignment between -> alignment with
RE19: Revised in l.58.

- l.96: rephrase/delete "as the size increases", because D_JS does not increase with the model size
RE20: Revised in l.97.

- l.202: relationship -> relationships
RE21: This whole paragraph was deleted as we now performed the ridge regression with whole-brain data.

- l.240: applies -> applies to
RE22: Revised in l.230.

- l.254: has -> have
RE23: Revised in l.244.

- l.282: instruction tuning and prompting -> instruction-tuned and prompted LLMs
RE24: Revised in l.272.

- l.283: delete "direct probability measurements from", because API vs. prompting is not relevant here
RE25: Revised in l.273.

- l.287: IT -> instruction-tuned
RE26: Revised in l.277.

- l.443: the new sentence starting with "We subtracted..." seems to be ungrammatical
RE27: Revised in l.441-442.

Reviewer #4 (Remarks on code availability):

The code provides a README file with enough instructions for installing and running the application.

References

Chen, G., Shin, Y. W., Taylor, P. A., Glen, D. R., Reynolds, R. C., Israel, R. B., & Cox, R. W. (2016). Untangling the relatedness among correlations, part I: Nonparametric approaches to inter-subject correlation analysis at the group level., Untangling the Relatedness among Correlations, Part I: Nonparametric Approaches to Inter-Subject Correlation Analysis at the Group Level. *NeuroImage*, *142*, 248, 248–259.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5203–5217.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022, October 31). Toward a realistic model of speech processing in the brain with self-supervised learning. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Oh, B.-D., & Schuler, W. (2023). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1915–1921.