nature computational science

Supplementary information

https://doi.org/10.1038/s43588-025-00863-0

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

In the format provided by the authors and unedited

Supplementary information

Section 1 Summary statistics for the reading eye-tracking and fMRI data

Supplementary Table 1. Mean and standard deviation of next-word prediction (NWP) loss from all LLMs averaged over all sentences in the test stimuli.

Model	Mean	Standard deviation
GPT-2 774M	5.268	1.671
LLaMA 7B	4.876	1.050
Alpaca 7B	5.378	1.196
Vicuna 7B	5.256	1.180
Gemma-Instruct 7B	5.947	1.701
Mistral-Instruct 7B	5.227	1.149
LLaMA 13B	4.873	1.078
Alpaca 13B	5.399	1.534
Vicuna 13B	5.781	1.320
LLaMA 30B	4.801	1.030
LLaMA 65B	4.805	1.067

Supplementary Table 2. Pairwise *t*-test results of the next-token prediction (NWP) loss from pairs of LLMs. Only pairs with a significant ($p<0.05$) or marginally significant ($p<0.1$) results are shown.

model-pair	*t*	*p*
GPT-2 774M vs. LLaMA 7B	8.688	2.73E-14
GPT-2 774M vs. Alpaca 7B	-3.819	3.55E-04
GPT-2 774M vs. Gemma-Instruct 7B	-7.398	3.07E-11
GPT-2 774M vs. LLaMA 13B	8.370	1.60E-13
GPT-2 774M vs. Alpaca 13B	-2.327	0.031
GPT-2 774M vs. Vicuna 13B	-10.323	2.27E-18
GPT-2 774M vs. LLaMA 30B	9.772	5.72E-17
GPT-2 774M vs. LLaMA 65B	9.466	3.28E-16
LLaMA 7B vs. Alpaca 7B	-17.695	3.64E-36
LLaMA 7B vs. Vicuna 7B	-15.218	1.60E-30
LLaMA 7B vs. Gemma-Instruct 7B	-13.322	4.92E-26
LLaMA 7B vs. Mistral-Instruct 7B	-11.634	9.45E-22
LLaMA 7B vs. Alpaca 13B	-6.439	4.20E-09
LLaMA 7B vs. Vicuna 13B	-20.090	1.02E-41
LLaMA 7B vs. LLaMA 30B	3.589	0.001
LLaMA 7B vs. LLaMA 65B	2.659	0.013
Alpaca 7B vs. Vicuna 7B	4.731	1.15E-05
Alpaca 7B vs. Gemma-Instruct 7B	-6.351	6.23E-09
Alpaca 7B vs. Mistral-Instruct 7B	3.208	0.003
Alpaca 7B vs. LLaMA 13B	15.111	2.51E-30
Alpaca 7B vs. Vicuna 13B	-8.947	6.49E-15

Alpaca 7B vs. LLaMA 30B	17.186	4.59E-35
Alpaca 7B vs. LLaMA 65B	15.736	1.20E-31
Vicuna 7B vs. LLaMA 13B	12.876	5.93E-25
Vicuna 7B vs. Alpaca 13B	-1.806	9.98E-02
Vicuna 7B vs. Vicuna 13B	-12.652	2.09E-24
Vicuna 7B vs. LLaMA 30B	15.047	3.08E-30
Vicuna 7B vs. LLaMA 65B	14.424	9.34E-29
Gemma-Instruct 7B vs. Mistral-Instruct 7B	8.087	7.48E-13
Gemma-Instruct 7B vs. LLaMA 13B	13.417	3.12E-26
Vicuna 7B vs. LLaMA 13B	12.876	5.93E-25
Gemma-Instruct 7B vs. Alpaca 13B	3.906	2.70E-04
Gemma-Instruct 7B vs. LLaMA 30B	14.926	5.40E-30
Gemma-Instruct 7B vs. LLaMA 65B	14.245	2.39E-28
Mistral-Instruct 7B vs. LLaMA 13B	12.935	4.62E-25
Mistral-Instruct 7B vs. Alpaca 13B	-1.786	0.087
Mistral-Instruct 7B vs. Vicuna 13B	-11.036	3.29E-20
Mistral-Instruct 7B vs. LLaMA 30B	15.733	1.20E-31
Mistral-Instruct 7B vs. LLaMA 65B	15.298	1.23E-30
LLaMA 13B vs. Alpaca 13B	-6.290	8.02E-09
LLaMA 13B vs. Vicuna 13B	-23.185	4.41E-48
LLaMA 13B vs. LLaMA 30B	4.521	2.64E-05
LLaMA 13B vs. LLaMA 65B	3.520	9.42E-04
Alpaca 13B vs. Vicuna 13B	-4.218	8.54E-05
Alpaca 13B vs. LLaMA 30B	7.171	9.96E-11
Alpaca 13B vs. LLaMA 65B	6.920	3.60E-10
Vicuna 13B vs. LLaMA 30B	22.438	8.43E-47
Vicuna 13B vs. LLaMA 65B	23.209	4.41E-48

Supplementary Table 3. One sample *t*-test results for the divergence of the attention matrices for the experimental stimuli between the base and fine-tuned LLMs.

Model-pair	*t*	*p*
Alpaca 7B vs. LLaMA 7B	36.004	9.73E-17
Vicuna 7B vs. LLaMA 7B	34.385	1.99E-17
Alpaca 13B vs. LLaMA 13B	40.278	6.12E-21
Vicuna 13B vs. LLaMA 13B	45.411	4.29E-16

Supplementary Table 4. One sample *t*-test results for the divergence of the attention matrices for stimuli sentences with and without instruction or noise prefixes.

Prefix	Model	*t*	*p*
	LLaMA 7B	40.837	6.53E-15
	Alpaca 7B	41.077	9.85E-15

instruction	Vicuna 7B	43.583	3.01E-14
	LLaMA 13B	39.315	4.90E-19
	Alpaca 13B	44.100	2.95E-20
	Vicuna 13B	44.118	1.41E-19
noise	LLaMA 7B	40.572	3.28E-16
	Alpaca 7B	40.732	6.31E-16
	Vicuna 7B	42.266	2.24E-15
	LLaMA 13B	37.969	1.09E-20
	Alpaca 13B	45.329	9.30E-21
	Vicuna 13B	47.389	1.67E-20

Supplementary Table 5. Mean regression scores of LLMs' attention matrices against trivial patterns across layers.

Model	Mean	Standard deviation
GPT-2 774M	0.436	0.236
LLaMA 7B	0.497	0.133
Alpaca 7B	0.496	0.137
Vicuna 7B	0.491	0.135
Gemma-Instruct 7B	0.483	0.166
Mistral-Instruct 7B	0.491	0.178
LLaMA 13B	0.499	0.123
Alpaca 13B	0.482	0.124
Vicuna 13B	0.477	0.131
LLaMA 30B	0.455	0.143
LLaMA 65B	0.454	0.149

Supplementary Table 6. Pairwise *t*-test results for the mean regression scores of LLMs' attention matrices against trivial patterns. Only pairs with a significant ($p<0.05$) or marginally significant ($p<0.1$) results are shown.

model-pair	*t*	*p*
GPT-2 774M vs. LLaMA 13B	-1.711	0.092
LLaMA 7B vs. LLaMA 30B	2.036	0.045
LLaMA 7B vs. LLaMA 65B	1.676	0.097
Alpaca 7B vs. LLaMA 30B	1.961	0.053
Vicuna 7B vs. LLaMA 30B	1.819	0.072
LLaMA 13B vs. LLaMA 30B	2.187	0.032
LLaMA 13B vs. LLaMA 65B	1.770	0.080

Supplementary Table 7. Mean regression scores from the best-performing layer of LLMs against eye regression numbers across participants.

Model	Mean	Standard deviation
GPT-2 124M	0.179	0.027
GPT-2 335M	0.186	0.028
GPT-2 774M	0.174	0.026

GPT-2 1.5B	0.176	0.026
LLaMA 7B	0.346	0.032
Alpaca 7B	0.352	0.039
Vicuna 7B	0.344	0.034
Gemma-Instruct 7B	0.366	0.045
Mistral-Instruct 7B	0.377	0.047
LLaMA 13B	0.364	0.032
Alpaca 13B	0.374	0.035
Vicuna 13B	0.361	0.033
LLaMA 30B	0.387	0.038
LLaMA 65B	0.406	0.044

Supplementary Table 8. One-way ANOVA results for the regression scores of the attention matrices of each base and fine-tuned LLM against human regressive eye saccade number patterns for the stimuli sentences across the 5 experimental sections.

Model	***F***	***p***
LLaMA 7B	0.333	0.856
Alpaca 7B	0.396	0.811
Vicuna 7B	0.392	0.814
LLaMA 13B	0.852	0.493
Alpaca 13B	0.332	0.856
Vicuna 13B	1.135	0.34

Supplementary Table 9. Mean normalized correlation coefficients of the ridge regression results from the best-performing layer of LLMs against fMRI data across participants.

Model	**Mean**	**Standard deviation**
GPT-2 base	0.384	0.049
GPT-2 medium	0.434	0.059
GPT-2 large	0.458	0.063
GPT-2 xlarge	0.452	0.053
LLaMA 7B	0.500	0.070
Alpaca 7B	0.501	0.069
Vicuna 7B	0.492	0.070
Gemma-Instruct 7B	0.425	0.056
Mistral-Instruct 7B	0.471	0.058
LLaMA 13B	0.583	0.063
Alpaca 13B	0.579	0.063
Vicuna 13B	0.576	0.064
LLaMA 30B	0.586	0.072
LLaMA 65B	0.624	0.068

Section 2 Additional results from listening fMRI data

Additional results from fMRI data of naturalistic listening

To verify whether our findings can generalize to a different dataset, we performed the same analysis on a fMRI dataset collected while participants listened to a 20-minute Chinese audiobook in the scanner. The dataset was collected for another project[1] and involves a total of 26 participants (15 females, mean age=23.96±2.23 years) listening to two sections of the Chinese version of "The Little Prince". All participants were right-handed native Mandarin speakers enrolled in an undergraduate or graduate program in Shanghai and had no self-reported history of neurological disorders. The fMRI data was collected in a 7.0 T Terra Siemens MRI scanner at the Zhangjiang International Brain Imaging Centre at Fudan University, Shanghai. The experimental procedures have been approved by the Ninth People's Hospital affiliated with Shanghai Jiao Tong University School of Medicine (EEG: SH9H-2019-T33-2; fMRI: SH9H-2022-T379-2). Anatomical scans were obtained using a Magnetization Prepared RApid Gradient-Echo (MP-RAGE) SAG iPAT2 pulse sequence with T1-weighted contrast (256 single-shot interleaved sagittal slices with A/P phase encoding direction; voxel size: 0.7×0.7×0.7 mm; FOV: 208 mm; TR: 3800 ms; TE: 2.32 ms; flip angle: 7°; acquisition time: 3 s; GRAPPA in-plane acceleration factor: 3). Functional imaging was conducted with T2-weighted echo-planar imaging (85 interleaved axial slices, anterior-posterior phase encoding; voxel size: 1.6×1.6×1.6 mm; FOV: 208 mm; TR: 1000 ms; TE: 22.2 ms; multiband acceleration factor: 5; flip angle: 45°). fMRI preprocessing of the was conducted using fMRIPrep (v25.0.0[2]) with all default parameters. 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 v7.3.2[3]).

We regressed the attention weights of the base and fine-tuned LLaMA3 8B and LLaMA3 70B models against fMRI data matrices at the paragraph level. This analysis extends beyond sentence-level comprehension to discourse-level processing and incorporates both a different modality (listening vs. reading) and a different language (Chinese vs. English). Specifically, we extracted the BOLD signal time-locked to each word's offset, adding five scans to account for peak hemodynamic responses within each paragraph of the stimuli. We then constructed a representational dissimilarity matrix (RDM[4]) for each paragraph using the 20 neighboring voxels around each voxel. The lower triangles of these RDMs were extracted, concatenated across all paragraphs, and used as dependent variables. For our predictors, we extracted and flattened the lower triangles from the attention matrices for all paragraphs. We then applied the same ridge regression analyses using attention matrices from all attention heads to predict the fMRI data RDM at each voxel within a bilateral language mask. The language mask covered regions including the whole temporal lobe, the inferior frontal gyrus (IFG) and the angular gyrus (AG).

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. 1a and Supplementary Table 10-12).

Additional results from fMRI data of comprehension tasks

To further investigate the impact of fine-tuning on model-brain alignment in a task-intensive setting, we regressed predictions from LLaMA3-Instruct (8B and 70B) against fMRI data while participants answered multiple-choice comprehension questions about the preceding listening session through button press in the scanner. The full set of questions is available in our OpenNeuro

dataset directory (https://openneuro.org/datasets/ds005345[5]). An example question (translated into English) is as follows:

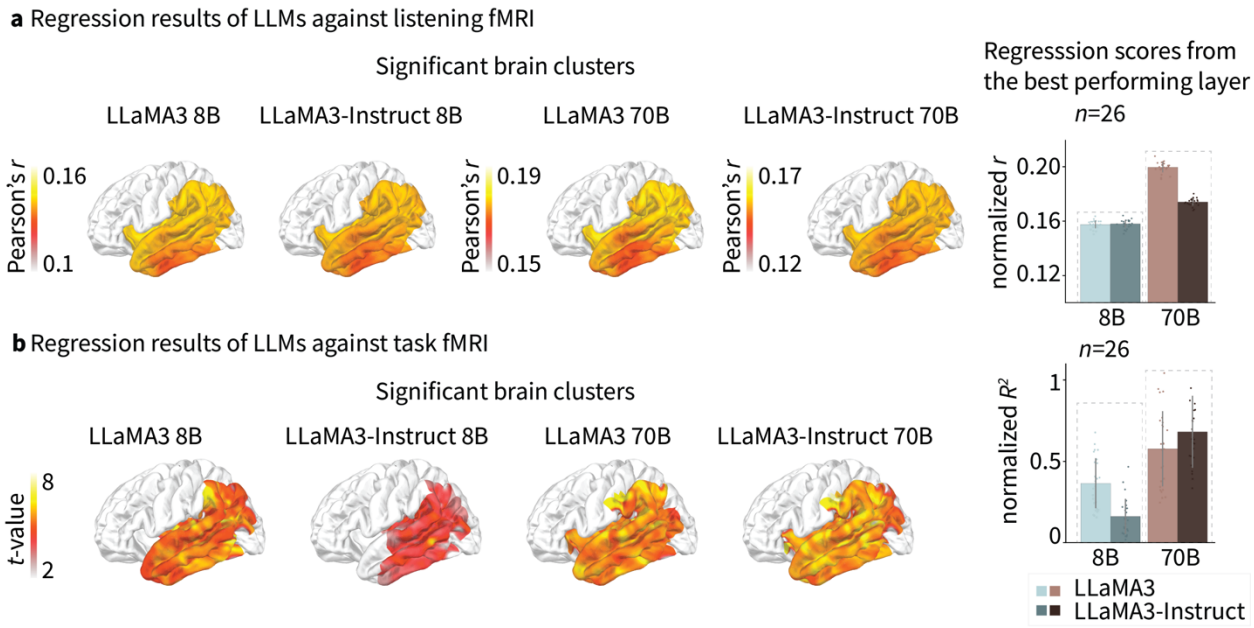
Why is it difficult to talk to the Little Prince?
a. He doesn't speak.
b. He doesn't ask questions.
c. He speaks an alien language.
d. He doesn't answer questions directly.

We prompted the fine-tuned LLMs with the text from the listening session along with the multiple-choice questions. We then extracted the models' 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 (adding 5 scans to capture the peak hemodynamic response). The regression was conducted for each voxel within a bilateral language mask in every subject. At the group level, the beta maps for each model were z-scored and assessed for statistical significance using cluster-based permutation *t*-tests [6] with 10,000 permutations.

Our results indicated that for smaller model sizes, LLaMA3 exhibited higher mean regression scores with task-based fMRI data compared to LLaMA3-Instruct. For the larger model size, LLaMA3-Instruct had a higher mean regression score across participants relative to the base model (see Supplementary Fig. 1b and Supplementary Table 13-14).

References

1. Wang, Q. *et al*. Le Petit Prince (LPP) multi-talker: Naturalistic 7 T fMRI and EEG dataset. *Sci Data* **12**, 829 (2025).
2. Esteban, O. *et al*. Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nat Protoc* **15**, 2186–2202 (2020).
3. Fischl, B. Freesurfer. *NeuroImage* **62**, 774–781 (2012).
4. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Sys. Neurosci*. (2008) doi:10.3389/neuro.06.004.2008.
5. Ma, Z., Wang, N. & Li, J. Le Petit Prince Multitalker: Naturalistic 7T fMRI and EEG Dataset. Openneuro https://doi.org/10.18112/openneuro.ds005345.v1.0.1 (2025).
6. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**, 177–190 (2007).



Supplementary Figure 1 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 normalized Pearson's *r* (y-axis) averaged over the significant brain clusters from the best-performing layer across participants (N=26). **b,** Significant brain clusters showing alignment between LLMs and fMRI data during the question-answering task. The right panel presents the averaged regression scores (y-axis) from the best-performing layer across participants (N=26). Error bars denote standard deviation.

Supplementary Table 10. Summary statistics for the significant brain clusters from the correlation coefficients of LLMs against listening fMRI.

Model	N vertices	*p*	Cohen's d
LLaMA3 8B	2504	2.00E-04	55.332
LLaMA3-Instruct 8B	2504	0	54.545
LLaMA3 70B	2504	2.00E-04	68.783
LLaMA3-Instruct 70B	2504	3.00E-04	83.215

Supplementary Table 11. Mean normalized correlation coefficients of the best-performing layer of LLMs against listening fMRI across participants.

Model	Mean	Standard deviation
LLaMA3 8B	0.158	0.003
LLaMA3-Instruct 8B	0.158	0.003
LLaMA3 70B	0.199	0.003
LLaMA3-Instruct 70B	0.174	0.002

Supplementary Table 12. Pairwise *t*-test results for the correlation coefficients from the best-performing layer of LLMs against listening fMRI data. Only pairs with a significant ($p<0.05$) or marginally significant ($p<0.1$) difference are shown.

Model-pair	*t*	*p*

LLaMA3-Instruct 8B vs. LLaMA3 8B	21.173	1.76E-17
LLaMA3 70B vs. LLaMA3 8B	2561.722	1.73E-68
LLaMA3-Instruct 70B vs. LLaMA3 8B	84.477	3.68E-32
LLaMA3 70B vs. LLaMA3-Instruct 8B	1824.727	4.17E-65
LLaMA3-Instruct 70B vs. LLaMA3-Instruct 8B	84.885	3.68E-32
LLaMA3 70B vs. LLaMA3-Instruct 70B	122.611	5.65E-36

Supplementary Table 13. Summary statistics for the significant brain clusters from the beta values of LLMs against task fMRI.

Model	N vertices	*p*	*t*	Cohen's d
LLaMA3 8B	1094	0	5.410	7.805
LLaMA3-Instruct 8B	705	0	4.323	7.366
LLaMA3 70B	1439	0	5.557	9.814
LLaMA3-Instruct 70B	1652	0	5.664	8.473

Supplementary Table 14. Mean normalized regression scores of the best-performing layer of LLMs against task fMRI across participants.

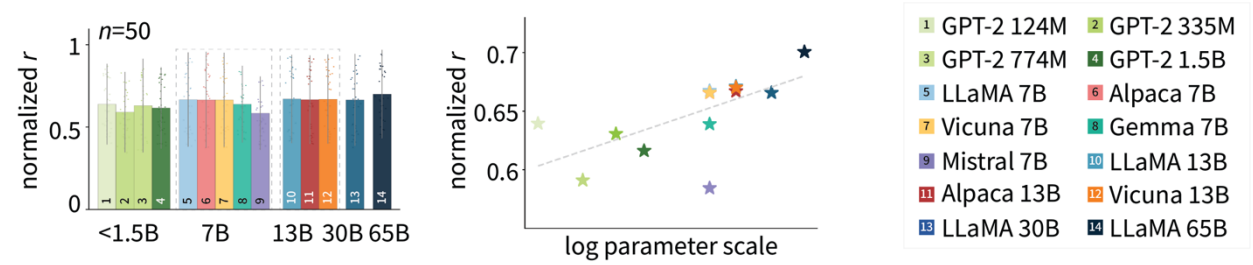
Model	Mean	Standard deviation
LLaMA3 8B	0.365	0.153
LLaMA3-Instruct 8B	0.162	0.108
LLaMA3 70B	0.578	0.231
LLaMA3-Instruct 70B	0.682	0.221

Supplementary Table 15. Pairwise *t*-test results for the beta maps of the regression results from the best-performing layer of LLMs against listening fMRI data. Only pairs with a significant ($p<0.05$) or marginally significant ($p<0.1$) difference are shown.

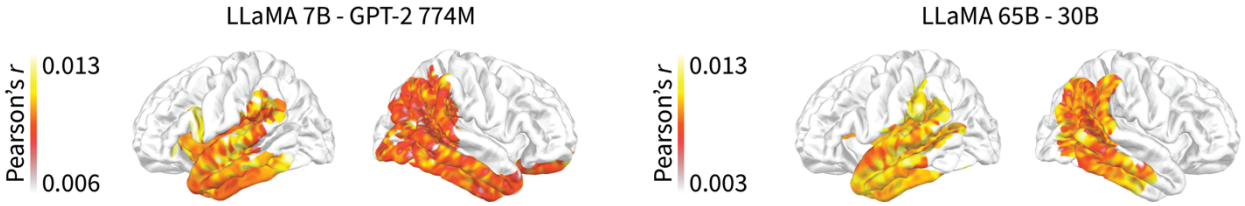
Model-pair	*t*	*p*
LLaMA3 8B vs. LLaMA3-Instruct 8B	6.874	7.47E-07
LLaMA3 70B vs. LLaMA3 8B	3.729	9.90E-04
LLaMA3-Instruct 70B vs. LLaMA3 8B	7.456	3.31E-07
LLaMA3 70B vs. LLaMA3-Instruct 8B	19.895	6.89E-16
LLaMA3-Instruct 70B vs. LLaMA3-Instruct 8B	39.124	2.16E-22
LLaMA3-Instruct 70B vs. LLaMA3 70B	4.083	5.76E-04

Section 3 Regression results for reading fMRI data using raw attention matrices (no trivial-pattern subtraction).

a Regression results of the best performing layer of different LLMs with trivial patterns against fMRI data



b Significant clusters from the contrast of the regression results between LLMs with trivial patterns



Supplementary Figure 2 Regression results using original attention matrices of the LLMs (without subtracting the trivial patterns). **a,** Regression results of the LLMs' best-performing layer on the fMRI activity patterns and their results on a logarithmic size scale. The bar plot denotes the mean of the normalized correlation coefficients of the LLMs' best-performing layer across subjects (N=50). Error bars denote standard deviation. **b,** Significant brain clusters from the contrast of the correlation coefficients of different LLMs with smaller and larger sizes.

Supplementary Table 16. Mean correlation coefficients from the best-performing layer of LLMs using their original attention matrices against fMRI data across participants.

Model	Mean	Standard deviation
GPT-2 base	0.639	0.069
GPT-2 medium	0.591	0.071
GPT-2 large	0.630	0.075
GPT-2 xlarge	0.616	0.067
LLaMA 7B	0.667	0.079
Alpaca 7B	0.665	0.080
Vicuna 7B	0.666	0.079
Gemma-Instruct 7B	0.639	0.069
Mistral-Instruct 7B	0.584	0.063
LLaMA 13B	0.671	0.076
Alpaca 13B	0.667	0.076
Vicuna 13B	0.670	0.076
LLaMA 30B	0.665	0.079
LLaMA 65B	0.700	0.077