Instruction-tuned large language models misalign with natural language comprehension in humans

Authors:
Changjiang Gao (first author), Department of Computer Science and Technology, Nanjing University; Department of Linguistics and Translation, City University of Hong Kong
gaocj@smail.nju.edu.cn

Zhengwu Ma (co-first author), Department of Linguistics and Translation, City University of Hong Kong
zhengwuma2-c@my.cityu.edu.hk

Jiajun Chen, Department of Computer Science and Technology, Nanjing University
chenjj@nju.edu.cn

Ping Li, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
pi2li@polyu.edu.hk

Shujian Huang (corresponding author), Department of Computer Science and Technology, Nanjing University, huangsj@nju.edu.cn

Jixing Li (corresponding author), Department of Linguistics and Translation, City University of Hong Kong
jixingli@cityu.edu.hk

Author contributions: Changjiang Gao designed the study and analysed the models and eye-tracking data; Zhengwu Ma analysed the fMRI data and plotted the results; Jiajun Chen and Shujian Huang designed the research; Ping Li collected the eye-tracking and fMRI data; Jixing Li designed the research, analyzed data, and wrote the paper.

Funding: This work was supported by the CityU Start-up Grant 7020086 and CityU Strategic Research Grant 7200747 (Li, J.).

Conflict of interest: The authors declare no competing interests.

Abstract
Transformer-based language models have significantly advanced our understanding of meaning representation in the human brain. Prior research utilizing smaller models like BERT and GPT-2 suggests that "next-word prediction" is a computational principle shared between machines and humans. However, recent advancements in large language models (LLMs) have highlighted the effectiveness of instruction tuning beyond next-word prediction. It remains to be tested whether instruction tuning can further align the model with language processing in the human brain. In this study, we evaluated the self-attention of base and finetuned LLMs of different sizes against human eye movement and functional magnetic resonance imaging (fMRI) activity patterns during naturalistic reading. Our results reveal that increases in model size significantly enhance the alignment between LLMs and brain activity, whereas instruction tuning does not. These findings confirm a scaling law in LLMs' brain-encoding performance and suggest that "instruction-following" may not mimic natural language comprehension in humans.

Introduction
Autoregressive Transformers are increasingly used in cognitive neuroscience for language processing studies, enhancing our understanding of meaning representation and composition in the human brain (Caucheteux & King, 2022; Goldstein et al., 2022; Huth et al., 2016; Schrimpf et al., 2021; Yu et al., 2024). For instance, Goldstein et al. (2022) found that the probability of words given a context significantly correlates with human brain activity during naturalistic listening, suggesting that language models and the human brain share some computational principles for language processing. Additionally, pre-trained Transformers are essential for decoding speech or text from neuroimaging data (e.g., Millet et al., 2022; Tang et al., 2023). They provide embeddings for training encoding models that map words to neural data and generate continuations as decoding candidates (Tang et al., 2023). However, those studies mostly adopted smaller pre-trained language models such GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019), whereas recent large language models (LLMs) such as GPT4 (OpenAI et al., 2024) and LLaMA (Touvron et al., 2023) are significantly larger in terms of parameter size and training data. It has been demonstrated that as model size, training dataset and compute budget increase, so does performance, following a power-law scaling law (Henighan et al., 2020; Hestness et al., 2017; Kaplan et al., 2020). These newer LLMs have already been adopted in recent studies to understand language processing in the human brain (e.g., Gao et al., 2024), but whether these LLMs better resemble human language processing remains debated. On the one hand, Antonello et al. (2023) demonstrated that larger models enhance brain encoding performance, mirroring the scaling law in other deep learning contexts. On the other hand, larger models have been questioned as valid cognitive models due to their extensive training data and their ability to access context hundreds of words long, which far exceeds human capabilities (e.g., Warstadt et al., 2023). Research has shown that limiting the context access of language models (Kuribayashi et al., 2022; Yu et al., 2024) can improve their simulation of language comprehension processing in humans.

In addition to scaling, finetuning LLMs post-training has also been shown to enhance model performance and generalization to new tasks (Chung et al., 2024; Ouyang et al., 2022; Sanh et al., 2022; Wei et al., 2021). 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. They argued that the typical language modeling objective of predicting the next token is "misaligned" with user intentions, resulting in unintended model

behaviors. However, the "next-word prediction" mechanism, rooted in the information-theoretic approaches to efficient message transmission (e.g., Oliver, 1952; Shannon, 1948), has a long history in cognitive science. It is regarded as a fundamental computational principle shared by both machine and human language processing (Goldstein et al., 2022; Hasson et al., 2020; Ryskin & Nieuwland, 2023). Additionally, the "next-sentence prediction" mechanism can also be implemented in LLMs and has been shown to simulate human language processing patterns (Yu et al., 2024). The superior performance of finetuned LLMs over base LLMs raises the question of whether "instruction-following" also has some cognitive reality in human language processing.

In this work, we systematically compared the self-attention of base and finetuned LLMs of varying sizes against human eye movement and functional magnetic resonance imaging (fMRI) activity patterns during naturalistic reading (Li et al., 2022). While previous research has demonstrated that scaling enhances brain encoding performance (Antonello et al., 2023), our work is the first to systematically investigate the impact of both scaling and finetuning on the alignment of LLMs with human behavior and neural activity during naturalistic reading. We show that as model size increases from 774M to 65B, the alignment between human eye movement and fMRI activity patterns also significantly improves, adhering to a scaling law. Instruction tuning, on the other hand, does not affect this alignment. Model analyses show that base and finetuned LLMs diverged the most when instructions were added to the stimuli sentences, suggesting that finetuned LLMs are sensitive to instructions in ways that naturalistic human language processing may not be.

Results

Model performance on next-token prediction. We used the publicly available Reading Brain dataset from OpenNeuro (Li et al., 2022) to investigate the effects of scaling and instruction tuning on the alignment between LLMs and human eye movement and neural data. The dataset includes concurrent eye-tracking and fMRI data collected from 50 native English speakers (25 females, mean age = 22.5 ± 4.1 years) as they read five English STEM articles inside an fMRI scanner. Each article contains an average of 29.6 ± 0.68 sentences, with each sentence comprising approximately 10.33 ± 0.15 words. Participants read each article sentence-by-sentence in a self-paced manner, pressing a response button to advance to the next sentence. We regressed the self-attention of base and finetuned LLMs of varying sizes against the eye movement and functional fMRI activity patterns of each sentence (see Figure 1 for the experimental procedure and the analyses pipeline). The LLMs employed in this study include GPT2-large, 4 different sizes of LLaMA (7B, 13B, 30B, and 65B) and two fine-tuned versions of LLaMA (Alpaca and Vicuna) in 7B and 13B configurations (see Table 1 for the detailed configurations of the LLMs).

Before comparing the LLMs with human behavioral and neural patterns, we first conducted a series of analyses to evaluate the internal states and outputs of the LLMs using our experimental stimuli. To test how much the LLMs differ in predicting the next word given the test stimuli, we calculated the averaged next-token prediction (NTP) loss of all the LLMs on every sentence of the articles in the Reading Brain experiment, and performed a two-sided related samples *t*-test for every pair of models. The mean NTP loss of all sentences in the test stimuli for the 9 models are as follows: GPT2-large: 5.27 ± 1.67, LLaMA-7B: 4.88 ± 1.05, LLaMA-13B: 4.87 ± 1.08, LLaMA-30B: 4.8 ± 1.03, LLaMA-65B: 4.81 ± 1.07, Alpaca-7B: 5.38 ± 1.2, Alpaca-13B: 5.75 ± 1.64, Vicuna-7B: 5.26 ± 1.18, and Vicuna-13B: 5.78 ± 1.32. The differences in NTP losses between most model pairs are significant, except for the following four pairs: GPT2-large vs. Vicuna-7B, LLaMA-7B vs. LLaMA-13B, LLaMA-30B vs. LLaMA-65B, and Alpaca-7B vs. Alpaca-13B (see Figure 2A for the mean NTP losses for all models and Supplementary Table 1 for the detailed *t*-

test statistics between all model pairs). These results suggest that for non-finetuned models, as model size increases, the mean NTP loss decreases. However, finetuned models like Alpaca and Vicuna did not improve performance on NTP for our test stimuli.

Comparison of model attentions. To examine the effect of scaling and instruction tuning on LLMs' attention matrices, we calculated the mean Jensen-Shannon (J-S) divergence (D_{JS}) for each pair of LLM's attention matrices over all attention heads at each model layer. For LLMs with the same number of layers, we computed the D_{JS} layerwise. For LLMs with different numbers of layers, we averaged the attention matrices for every quarter of layers and computed the D_{JS} for each quarter-layer. Figure 2B shows the results of the divergence analyses. We observed that for both the base (LLaMA) and finetuned models (Alpaca and Vicuna), as the size increases from 7B to 13B, the D_{JS} of model attentions linearly increases from the first quarter to the last quarter of the model layers. However, when comparing the base and finetuned models of the same sizes, the D_{JS} of model attentions remains small across all layers for LLaMA-7B vs. Alpaca-7B, LLaMA-7B vs. Vicuna-7B, and LLaMA-13B vs. Alpaca-13B. Notably, the D_{JS} of model attentions between Vicuna-13B and LLaMA-13B are significantly larger than the other three pairs, especially at the higher layers (see Supplementary Table 2 for the detailed *t*-test statistics). This suggests that fine-tuning may affect the weights of model attention, but only when the model size is large enough.

Sensitivity of model attention to instructions. Since the Alpaca and Vicuna models were finetuned with instructions, we also compared the sensitivities of model attention to instructions. We added two instructions before each sentence in our text stimuli: "Please translate this sentence into German:", and "Please paraphrase this sentence:". As a control, we introduced a noise prefix composed of five randomly sampled English words, such as "Cigarette first steel convenience champion." We then extracted the attention matrices for the original sentence spans and calculate the D_{JS} of attentions between each model pair layerwise. Our results showed a significantly larger divergence in the attention matrices for the fine-tuned models when processing plain versus instructed texts, for both the 7B and 13B sizes. In contrast, the LLaMA models did not show sensitivity to instructions at either size. No significant difference was found for the D_{JS} of attentions across all layers between the base and finetuned models for plain versus noise-prefixed text (see Figure 2C and Supplementary Table 3 for the detailed *t*-test statistics).

Sensitivity of model attention to trivial patterns. Prior studies have highlighted certain patterns in LLMs' attention matrices, such as a tendency to focus on the first word of a sentence, the immediately preceding word (Vig & Belinkov, 2019), or on the word itself (Clark et al., 2019). To examine how scaling and fine-tuning influence the models' sensitivity to these trivial patterns, we constructed a binary matrix for each sentence in the test stimuli, marking cells that exhibited these trivial relationships. We then regressed each model's attention matrix for each sentence at each layer against the corresponding trivial patterns. Our findings showed that for the LLaMA series and their fine-tuned versions, as the model size increases from 7B to 65B, the average regression score for predicting the trivial patterns across layers also increases (see Figure 2). Pairwise independent *t*-tests confirmed significant differences between the regression scores for LLaMA-7B and LLaMA-65B (t=-3.19, p=0.011), Alpaca-7B and LLaMA-30B (t=-2.56, p=0.044), Alpaca-7B and LLaMA-65B (t=-3.72, p=0.004), and Vicuna-7B and LLaMA-65B (t=-3.45, p=0.006). No significant differences were observed between the LLaMA models and their fine-tuned versions (see Supplementary Table 4 for the detailed *t*-test statistics). Interestingly, the GPT2 model, despite

having the smallest parameter size, had the highest regression scores for predicting these trivial patterns. Since these trivial patterns do not capture relationships between non-adjacent words, we subtracted these patterns from all the attention matrices of the LLMs for the following analyses (see "Model sensitivity to trivial patterns" in Methods for details).

Effects of scaling versus finetuning on model-behavior alignment. To assess the alignment of different LLMs with human eye movement patterns during naturalistic reading, we regressed the attention matrices of the LLMs against eye regression number and duration patterns for all stimuli sentences (detailed in the "Alignment between LLMs and eye movement" section in Methods). Our findings show that for the LLaMA series, as model size increases from 7B to 65B, the regression scores also increase across layers. GPT2, which has the smallest parameter size, exhibits the lowest regression scores, particularly in the higher layers. In contrast, base and finetuned models of the same sizes exhibit no difference in their regression scores when aligned with human eye movement patterns, suggesting that scaling, rather than finetuning, enhances the alignment between LLMs and human reading behaviors. These trends hold for both regression number and duration patterns, although eye regression number showed higher overall regression scores across layers than eye regression duration (see Figure 3A). We further plotted the maximum regression scores from all model layers against different LLMs and the logarithmic scale of parameter size, illustrating a clear scaling law of model-behavior alignment (see Figure 3B).

Given that participants answered 10 comprehension questions after reading each article, there is a possibility that their reading behavior shifted from naturalistic reading to a more focused approach aimed at solving questions as the experiment progressed. This could mean that LLMs with instruction tuning might increasingly align with human behavior later in the experiment. To test this hypothesis, we performed the same regression analyses separately for each section of the experiment. Our results revealed no significant difference in the regression scores for base and finetuned LLMs over time, suggesting that human reading behaviors during naturalistic reading are not influenced by the subsequent comprehension questions (see Figure 3C). Supplementary Table 5 lists the F statistics from one-way analysis of variance (ANOVA) for each base and finetuned LLM across the 5 experimental sections.

Effects of scaling versus finetuning on model-brain alignment. We next aligned the eye movement regression number patterns with the fMRI activity patterns. The results revealed a large cluster (N vertices=1016, t=0.59, p=0.0001, Cohen's d=1.9) spanning the left frontal-temporal regions (see Figure 4A), aligning with previous research on language network in the brain (e.g., Malik-Moraleda et al., 2022). We extracted this cluster as a functional region-of-interest (fROI) and regressed the attention matrices of different LLMs against the fMRI activity patterns for all our test stimuli within this fROI (see "Alignment between LLMs and fMRI data" in Methods for details). As shown in Figure 4B, the average regression score across subjects increases with model size, particularly in the middle and higher layers. In contrast, base and finetuned LLMs of the same sizes did not show differences in their average regression scores across subjects at each layer. We also plotted the regression scores from the best-performing layer of each model on a logarithmic scale, demonstrating a clear scaling effect where larger models better explained the fMRI activity patterns during naturalistic reading (see Figure 4C).

We averaged the regression scores across model layers for different LLMs and compared them between four model pairs with varying sizes: GPT2-large vs. LLaMA-7B, LLaMA-7B vs. LLaMA-13B, LLaMA-13B vs. LLaMA-30B, and LLaMA-30B vs. LLaMA-65B. Our findings

revealed large clusters in the fROI where larger LLMs consistently exhibited significantly higher regression scores than their smaller counterparts (see Figure 4D and Table 2). Additionally, we compared the regression scores of 4 pairs of base and finetuned models: LLaMA-7B vs. Alpaca-7B, LLaMA-7B vs. Vicuna-7B, LLaMA-13B vs. Alpaca-13B, and LLaMA-13B vs. Vicuna-13B. However, these comparisons did not reveal any significant brain clusters.

Discussion

Scaling and finetuning are two key factors behind the enhancements of recent LLMs compared to their predecessors, such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019). While prior work suggests that the scaling law also applies LLMs' brain encoding performance with extensive fMRI data during naturalistic language comprehension (Antonello et al., 2023), the efficacy of scaling in model-brain alignment with shorter texts remains uncertain. Given that GPT2 already predicts neural responses more accurately than an average human, often considered the ceiling level (Schrimpf et al., 2021), smaller models may have reached their performance limits on next-word prediction for simpler texts. Moreover, current LLMs far exceed human capabilities in terms of data input during training and memory resources for accessing contextual information during comprehension. It has been argued that larger models increasingly diverge from human language processing patterns (Kuribayashi et al., 2022). In this study, we evaluated the alignment between LLMs of varying sizes and human eye movement and fMRI activity patterns during naturalistic reading. Despite using experimental stimuli and fMRI data much smaller in size compared to previous studies (Antonello et al., 2023), we observed consistent improvements in alignment as model size increased from 774M to 65B, without any apparent diminishing returns. This indicates that the scaling law of model-brain alignment holds even with shorter text stimuli and smaller fMRI data.

While the largest LLMs today still do not match the human brain in terms of synapse count, training and operating such large LLMs pose significant computational challenges, especially in academic settings with limited computing budgets. Finetuning LLMs with instructions offers a viable approach to enhance the performance and usability of pre-trained language models without expanding their size (Chung et al., 2024; Ouyang et al., 2022; Sanh et al., 2022; Wei et al., 2021). Ouyang et al. (2022) noted that the typical "next-token prediction" training objective of language models often diverges from user intentions, leading to outputs that are less aligned with user preferences. Although there is ample evidence that human language processing involves "next-word prediction" (Goldstein et al., 2022; Hasson et al., 2020; Ryskin & Nieuwland, 2023), research also showed that finetuning language models for tasks like narrative summarization can enhance model-brain alignment, especially in understanding characters, emotions, and motions (Aw & Toneva, 2022). It is possible that "instruction-following" plays a role in human language learning and that finetuned models might contain richer discourse and pragmatic information beyond basic meaning representation.

However, our regression results with human behavioral and neural patterns did not reveal any significant improvement in alignment for finetuned LLMs compared to base models of the same sizes. We also examined whether finetuned models exhibited better alignment to eye movement patterns as participants completed more comprehension questions over time, but no significant differences were found in the regression scores. Since instruction tuning effectively realigned the model weights in response to instructions, our results suggest that human language processing in a naturalistic setting is not sensitive to instructions. However, these realigned model weights may better fit brain activity patterns during tasks involving instructions.

In summary, we compared base and finetuned LLMs of varying sizes against human eye movement and fMRI activity patterns during naturalistic reading. Our results highlighted a significant impact of scaling on model-brain alignment, whereas instruction tuning showed no such effect. These results serve as a reference for researchers selecting LLMs for brain encoding and decoding studies. One limitation of the study is the lack of comparisons between base and finetuned LLMs against human behavioral and neural data during experimental tasks with instructions. This gap leaves the potential for future research to explore the impact of instruction tuning on model-brain alignment in controlled experimental settings.

Methods

Eye-tracking and fMRI data. We used the openly available Reading Brain dataset (Li et al., 2022) on OpenNeuro. This dataset includes concurrent eye-tracking and fMRI data collected from 50 native English speakers (25 females, mean age = 22.5 ± 4.1 years) as they read five English STEM articles inside an fMRI scanner. Each article contains an average of 29.6 ± 0.68 sentences, with each sentence comprising approximately 10.33 ± 0.15 words. Participants read each article sentence-by-sentence in a self-paced manner, pressing a response button to advance to the next sentence. If there was no response within 8000 ms, the screen would automatically progress to the next sentence. The sequence in which the five texts were presented was randomized across participants to control for potential order effects. At the end of each article, participants answered 10 multiple-choice questions to ensure their comprehension. The whole experiment, including preparation time, lasted for about one hour (see Figure 1A for the experimental procedure).

All imaging and eye-tracking data were acquired at the Center for NMR Research at the Pennsylvania State University Hershey Medical Center in Hershey, Pennsylvania. The anatomical scans were acquired using a Magnetization Prepared RApid Gradient-Echo (MP-RAGE) pulse sequence with T1 weighted contrast (176 ascending sagittal slices with A/P phase encoding direction; voxel size = 1mm isotropic; FOV = 256 mm; TR = 1540 ms; TE = 2.34 ms; acquisition time = 216 s; flip angle = 9°; GRAPPA in- plane acceleration factor = 2; brain coverage is complete for cerebrum, cerebellum and brain stem). The functional scans were acquired using T2* weighted echo planar sequence images (30 interleaved axial slices with A/P phase encoding direction; voxel size = 3 mm × 3mm × 4 mm; FOV = 240 mm; TR = 400 ms; TE = 30 ms; acquisition time varied on the speed of self-paced reading, maximal 5.1 minutes per run; multiband acceleration factor for parallel slice acquisition = 6; flip angle = 35°; where the brain coverage missed the top of the parietal lobe and the lower end of the cerebellum). A pair of spin echo sequence images with A/P and P/A phase encoding direction (30 axial interleaved slices; voxel size=3mm× 3mm× 4mm; FOV=240mm; TR=3000ms; TE=51.2 ms; flip angle = 90°) were collected to calculate distortion correction for the multiband sequences (Glasser et al., 2013). fMRI preprocessing and analysis were performed in SPM12 v6906 (http://www.fil.ion.ucl.ac.uk/spm). The preprocessing steps involves the correction of field inhomogeneity artefacts, motion correction, coregistration and normalization to the Montreal Neurological Institute (MNI) atlas. Images were normalized with the 4th degree B-Spline Interpolation algorithm and further smoothed with a Gaussian kernel of 8 mm full-width-at-half-maximum (FWHM; Hsu et al., 2019). We further projected the volume-based data for each run onto a brain surface by using the nilearn's vol_to_surf function with defaults parameters (Abraham et al., 2014) and a "fsaverage5" template (Fischl, 2012).

Participants' eye movements were simultaneously recorded using an MRI-compatible EyeLink 1000 Plus eye tracker (SR Research, 2016) with a sampling rate of 1000 Hz. The eye tracker was mounted at the rear end of the scanner bore and captured eye movements via a

reflective mirror positioned above the MRI's head coil. This setup included the following parameters: distance between the reader's eyes and the presenting screen = 143 cm; distance between the camera and the participant's eyes via the reflective mirror was 120 cm; presenting screen size = 35.7 cm × 57.2 cm; average word length on the screen = 3.08 cm; average distance between words = 0.95 cm. On average, a reader's visual angle when fixating on a word is 1°14¢. Recording was performed monocularly from the right eye, and the participant's head was stabilized using the head coil. A 13-point calibration routine was conducted at the beginning of the experiment, followed by a validation process during the first run. For subsequent runs, validation checks were performed regularly, and if the error exceeded 1°, recalibration was conducted. Participants initially viewed a fixation cross for 500 ms on the left side of the screen before each sentence was displayed, helping them to anticipate the text presentation. The initial fixation cross lasted 6000 ms to allow participants ample time to prepare and for the blood-oxygen-level-dependent (BOLD) signal to stabilize. Subsequent fixation crosses between sentences were displayed for 500 ms each. Due to fixation drifting caused by the declining accuracy of calibration over time, we manually adjusted fixations falling outside the range of predefined target regions where sentences are presented using the Data Viewer™ software from SR Research (SR Research, 2016). Rather than using an auto-adjustment that aligns all fixations to a single horizontal line, we applied trial-by-trial corrections only along the y-axis. This approach preserved the original fixation patterns of the readers. Approximately 15% of the data required this type of manual correction.

LLMs. To investigate the effects of scaling and instruction tuning on the alignment of LLMs with human behavior and neural data, we utilized the open-source LLaMA model (Touvron et al., 2023) and its instruction-tuned variants, Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023), available in various sizes. LLaMA is a series of pre-trained causal language models trained on over 1 trillion publicly accessible text tokens, primarily in English. It achieved state-of-the-art performance on most LLM benchmarks (Touvron et al., 2023). We employed all four sizes of LLaMA: 7B, 13B, 30B, and 65B. Additionally, we included the 774M GPT-2 large (Radford et al., 2019) to represent smaller pre-trained language models.

Alpaca (Taori et al., 2023) was fine-tuned from the 7B LLaMA model and was trained on 52K English instruction-following demonstrations generated by GPT-3 (Brown et al., 2020) using the self-instruct method (Wang et al., 2023). We also developed a 13B version of Alpaca using the same training data and strategy. Our 13B Alpaca model achieved accuracy scores of 43.9 and 46.0 on the MMLU dataset (Hendrycks et al., 2020) in zero-shot and one-shot settings, respectively, outperforming the original 7B model's scores of 40.9 and 39.2. Vicuna versions 7B and 13B (Chiang et al., 2023) were fine-tuned from the respective 7B and 13B LLaMA models, using 70K user-shared conversations with ChatGPT (OpenAI et al., 2024). This dataset includes instruction and in-context learning samples across multiple languages.

Comparison of next-token prediction loss. To examine the effects of scaling and finetuning model's performance in next-token prediction, we calculated the per-token prediction loss (the negative log-likelihood loss normalized by sequence lengths) of the models employed in this study on every sentence of the articles in the Reading Brain dataset. The losses for all sentences were z-scored model-wise and the contrasts of the z-scored losses for two models (e.g., LLaMA-7B vs. Alapaca-7B) were tested using a two-sample two-tailed related *t*-test. The false discovery rate (FDR) was applied to correct for multiple comparisons across layers.

Comparison of model attentions. The self-attention matrices of different LLMs given the same input were compared using their mean Jensen-Shannon (J-S) divergence across all layers. For every sentence in our stimuli, we extracted the attention matrices A and B from one attention head and one layer of two target LLMs ($A, B \in \mathbb{R}^{n_{word} \times n_{word}}$), and their J-S divergence $D_{JS}(A, B)$ is computed as $D_{JS}(A, B) = \frac{1}{2}\sum_{i=1}^{n_{word}}[D_{KL}(A_i||B_i) + D_{KL}(B_i||A_i)]$, where A_i and B_i are the i-th rows in the two matrices and D_{KL} is the Kullback–Leibler (K-L) divergence (Kullback & Leibler, 1951). The attention matrices were normalized such that each row sums to 1, and the final D_{JS} for each layer was averaged across attention heads. Since LLMs use subword tokenization (Kudo and Richardson, 2018), 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. For models with different numbers of layers, we divided their layers into four quarters and averaged the D_{JS} quarter-wise. We compared each model-pair's D_{JS} for each layer or each quarter-layer using a two-sided two-sample related *t*-test with FDR correction.

Model sensitivity to instructions. We compared the models' attention matrices for each stimuli sentence when prefixed with two instructions: "Please translate this sentence into German:", and "Please paraphrase this sentence:". As a control, we introduced a noise prefix composed of five randomly sampled English words, such as "Cigarette first steel convenience champion." We then extracted the attention matrices for the original sentence spans. We calculated the D_{JS} between the prefixed and original sentences across different models to assess each model's sensitivity to instructions.

Model sensitivity to trivial patterns. Prior studies have highlighted certain trivial patterns in the attention matrices within a given context, such as a tendency to focus on the first word of a sentence, the immediately preceding word (Vig & Belinkov, 2019), or the word itself (Clark et al., 2019). To examine whether scaling and finetuning will change the models' reliance on these trivial patterns, we constructed a binary matrix for each sentence in the test stimuli, marking cells that exhibit these trivial attention relationships with a 1. We then flattened the lower-triangle parts of these matrices to create trivial pattern vectors. Subsequently, we performed ridge regressions using each model's attention vectors for each sentence at each quarter-layer to predict the corresponding trivial pattern vectors. The resulting regression scores were averaged across model layers and were z-scored and assessed for statistical significance using two-tailed one-sample *t*-tests with FDR corrections.

Alignment between LLMs and eye movement. Since our auto-regressive LLMs utilize right-to-left self-attention, we extracted the lower-triangle portions of the attention matrix from each layer and each attention head for every sentence. These matrices were flattened and concatenated to form the attention vector $v_{model}^{j,k}$ for all sentences at head k in layer j. We stacked these vectors along the attention heads to create a matrix $v_{model}^{j,k}$ for the j-th layer. For the human eye saccade data, we constructed matrices for saccade number and duration, $E_{num}, E_{dur} \in \mathbb{R}^{n_{word} \times n_{word}}$, for

each sentence. Each cell at row l and column m in E_{num} and E_{dur} represents the number of times and the duration 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 regression. Like the models' attention matrices, we flattened the eye regression number and duration matrices for all sentences and concatenated them to get the eye regression number and duration vectors v_{num}^{i} and v_{dur}^{i} for each subject i. We then conducted a ridge regression using the model attention matrix V_{model}^{j} at each layer j to predict each subject's eye regression and duration vectors. The regularization coefficient was kept at 1. The final R_{model}^{2} was normalized by the $R_{ceiling}^{2}$, where the ceiling model represents the mean of all subjects' eye regression number and duration vectors. Given that larger models also possess more attention heads, which could introduce confounding variables due to the increased number of regressors, we included a control model of matching size for each LLM, where the model weights were randomized to address this potential confound. At the group level, the significance of the contrast of the z-scored R_{model}^{2} for every model pair at every layer was examined using a two-tailed one-sample related *t*-test, with FDR corrections for multiple comparisons across layers.

To test this hypothesis that participants' reading behavior shifted from naturalistic reading to a more focused approach aimed at solving questions as the experiment progressed, we performed the same regression analysis separately for each section of the experiment. We then compared the regression scores of each LLM across different times using analysis of variance (ANOVA) to assess changes in model fit over the course of the experiment.

Alignment between LLMs and fMRI data. For each voxel of the fMRI data for each subject, we constructed a BOLD matrix $B \in \mathbb{R}^{n_{word} \times n_{word}}$ for each sentence. The value at row l and column m in B represents the sum of the BOLD signals at the timepoints where the eye fixation moves from the word in row l to the word in column m. We extracted the lower-triangle parts of the B matrices for all sentences and concatenated them to form the BOLD vector v_{B}^{i} for each subject i. Next, we conducted a ridge regression using each subject's eye regression and duration vectors v_{num}^{i} and v_{dur}^{i} to predict each subject's BOLD vector v_{B}^{i} at each voxel. This analysis is to identify the significant brain clusters associated with eye regression patterns, serving as our functional regions of interest (fROI) for aligning LLMs with the fMRI data. 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, and normalized the regression scores by the ceiling model, which is the mean of all subjects' BOLD vectors. The normalized regression scores for all voxels of each subject were z-scored, and all subjects' z-maps underwent a one-sample one-tailed *t*-test with a cluster-based permutation test (Maris & Oostenveld, 2007) involving 10,000 permutations. Clusters were formed from statistics corresponding to a p-value less than 0.05, and only clusters spanning a minimum of 50 vertices were included in the analysis.

After aligning the eye-tracking and fMRI data, we extracted the significant clusters as our fROI. We then performed ridge regressions using the attention matrix V_{model}^{j} at each layer j from each LLM to predict each voxel's BOLD vector v_{B}^{i} within the fROI for each subject. The regression scores for each LLM were normalized and z-scored within subjects. We averaged the regression scores across model layers and compared the z-maps for each model pair using the same cluster-based permutation *t*-tests (Maris & Oostenveld, 2007) with10,000 permutations. Clusters were formed from statistics corresponding to a p-value less than 0.05, and only clusters spanning a minimum of 50 vertices were included in the analysis. All our analyses were performed using

custom python codes, making heavy use of the torch (v2.2.0), mne (v.1.6.1; Gramfort et al., 2014), eelbrain (v.0.39.8; Brodbeck et al., 2023) and scipy (v1.12.0) packages. Obtaining the model attention scores requires around 0.5 GPU hours for each model on a platform with 20 Intel Xeon Gold 6248 CPUs, 216 GB ROM, and 4 Nvidia Tesla v100 32 GB GPUs. Running each regression requires around 0.5 hours per model layer for each subject's eye movement and fMRI data on a platform with 112 AMD EPYC 7522 CPUs and 512 GB ROM.

Data availability. The attention matrices of all the LLMs for our experimental stimuli are available at https://github.com/RiverGao/scaling_finetuning. The eye-tracking and fMRI dataset is available at https://openneuro.org/datasets/ds003974/versions/3.0.0.

Code availability. All codes are available at https://github.com/RiverGao/scaling_finetuning.

References

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.

Aw, K. L., & Toneva, M. (2022). Training language models to summarize narratives improves brain alignment. *International Conference on Learning Representations*.

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.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*, 134.

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*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, *25*, 1–53.

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.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

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

Gao, C., Li, J., Chen, J., & Huang, S. (2024). *Measuring meaning composition in the human brain with composition scores from large language models* (arXiv:2403.04325). arXiv.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*, Article 3.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*, 416–434.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *International Conference on Learning Representations*.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., & McCandlish, S. (2020). *Scaling laws for autoregressive generative modeling* (arXiv:2010.14701). arXiv.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., & Zhou, Y. (2017). *Deep learning scaling is predictable, empirically* (arXiv:1712.00409). arXiv.

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.

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*, 453–458.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* (arXiv:2001.08361). arXiv.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10421–10436). Association for Computational Linguistics.

Li, P., Hsu, C.-T., Schloss, B., Yu, A., Ma, L., Scotto, M., Seyfried, F., & Gu, C. (2022). *The Reading Brain project L1 adults* [Data set]. OpenNeuro.

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.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, *35*, 33428–33443.

Oliver, B. M. (1952). Efficient coding. *The Bell System Technical Journal*, *31*(4), 724–750.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv.

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.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*, 1032–1052.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., … Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization. *International Conference on Learning Representations*.

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*, e2105646118.

Shannon, C. E. (1948). *A mathematical theory of communication*. 55.

SR Research. (2016). *EyeLink 1000 Plus long range MRI-compatible eye-tracker*.

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.

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*, 1–9.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-following LLaMA model. In *GitHub repository*. GitHub.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models* (arXiv:2302.13971). arXiv.

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. 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. 63–76). Association for Computational Linguistics.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13484–13508). Association for Computational Linguistics.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 1–6.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

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*, eadn7744.

Model	Size	Layer	Attention head	Training data size	Fine-tuning
GPT2	774M	36	20	8B	
LLaMA	7B	32	32	1T	None
	13B	40	40		
	30B	60	52		
	65B	80	64		
Alpaca	7B	32	32	52K	Instruction
	13B	40	40		
Vicuna	7B	32	32	70K	Conversation
	13B	40	40		

Table 1. Configurations of the LLMs employed in the study.

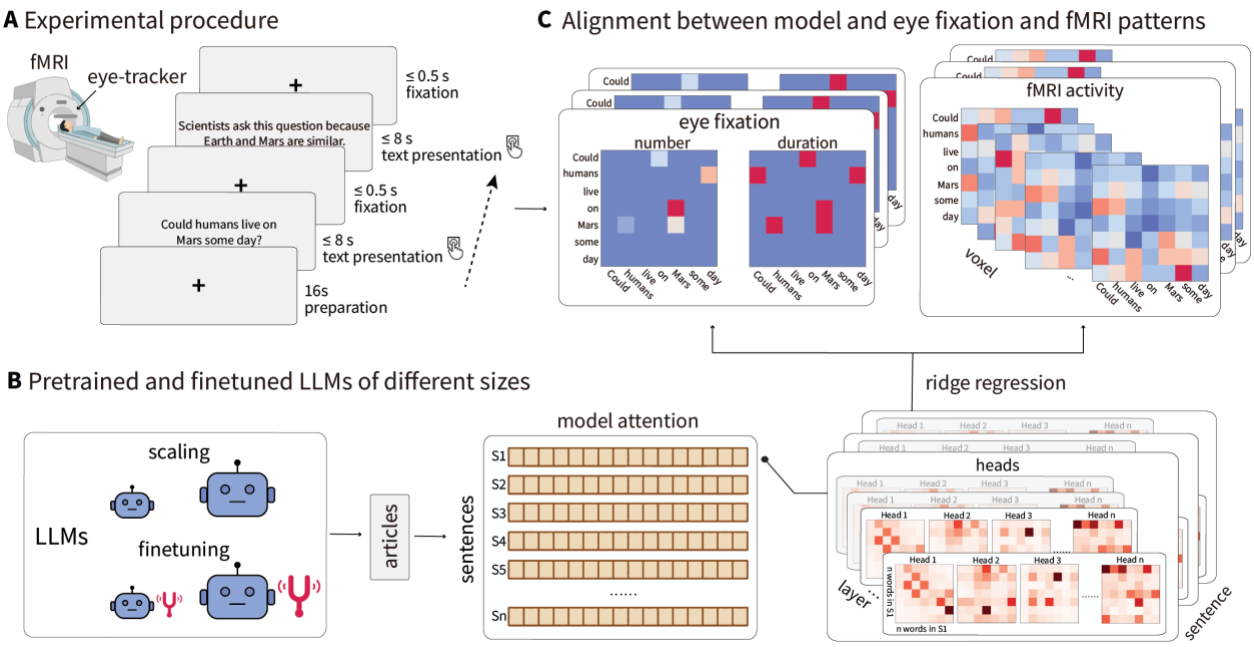

Figure 1. Experimental procedure of the dataset and the analyses pipeline. **A.** Experiment procedure. Participants read five English articles sentence by sentence inside the fMRI scanner with concurrent eye-tracking. **B.** LLMs of different sizes with and without finetuning are employed in the study. **C.** Analysis pipeline. The attention matrices of each layer of the LLMs for each sentence in the experimental stimuli were averaged over attention heads and aligned with eye movement and fMRI activity patterns for each sentence using ridge regression.

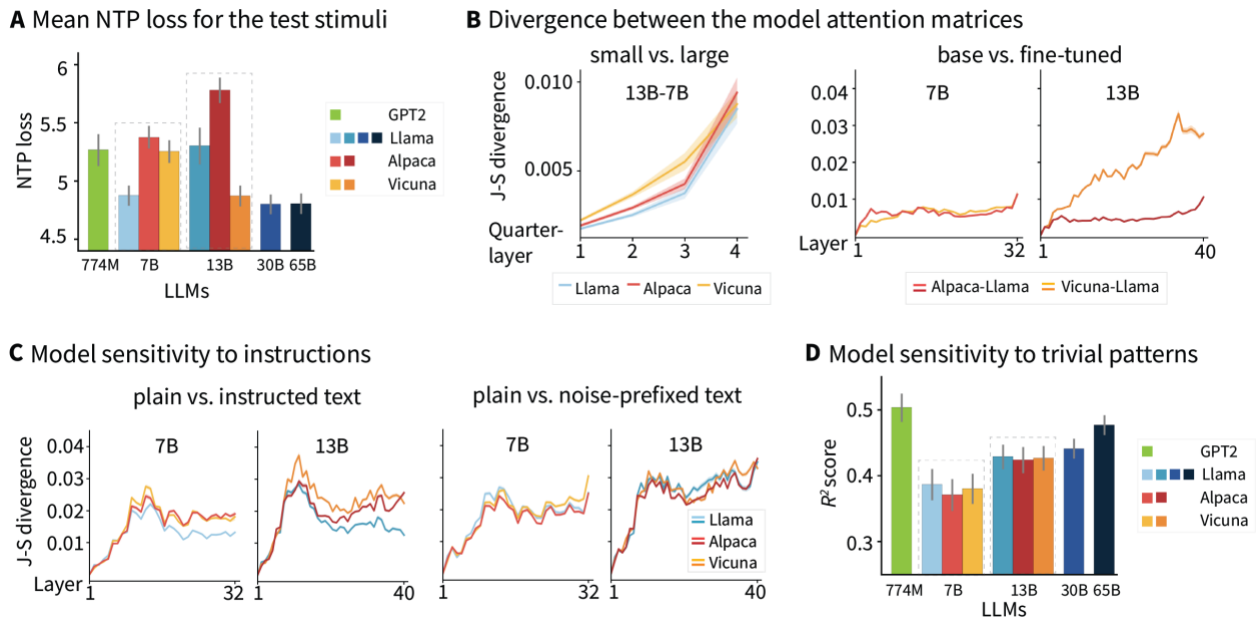

Figure 2. Comparison between the attention matrices of different LLMs. **A.** The mean next-token prediction (NTP) loss for the test stimuli of all the LLMs. **B.** J-S divergence between the attention matrices of different LLMs at each layer or quarter-layer. **C.** The effect of scaling and finetuning on LLMs' sensitivity to instructions. Shaded regions denote standard deviation. **D.** The effect of scaling and finetuning on LLMs' sensitivity to trivial patterns between words in a sentence. Error bars denote standard deviation.

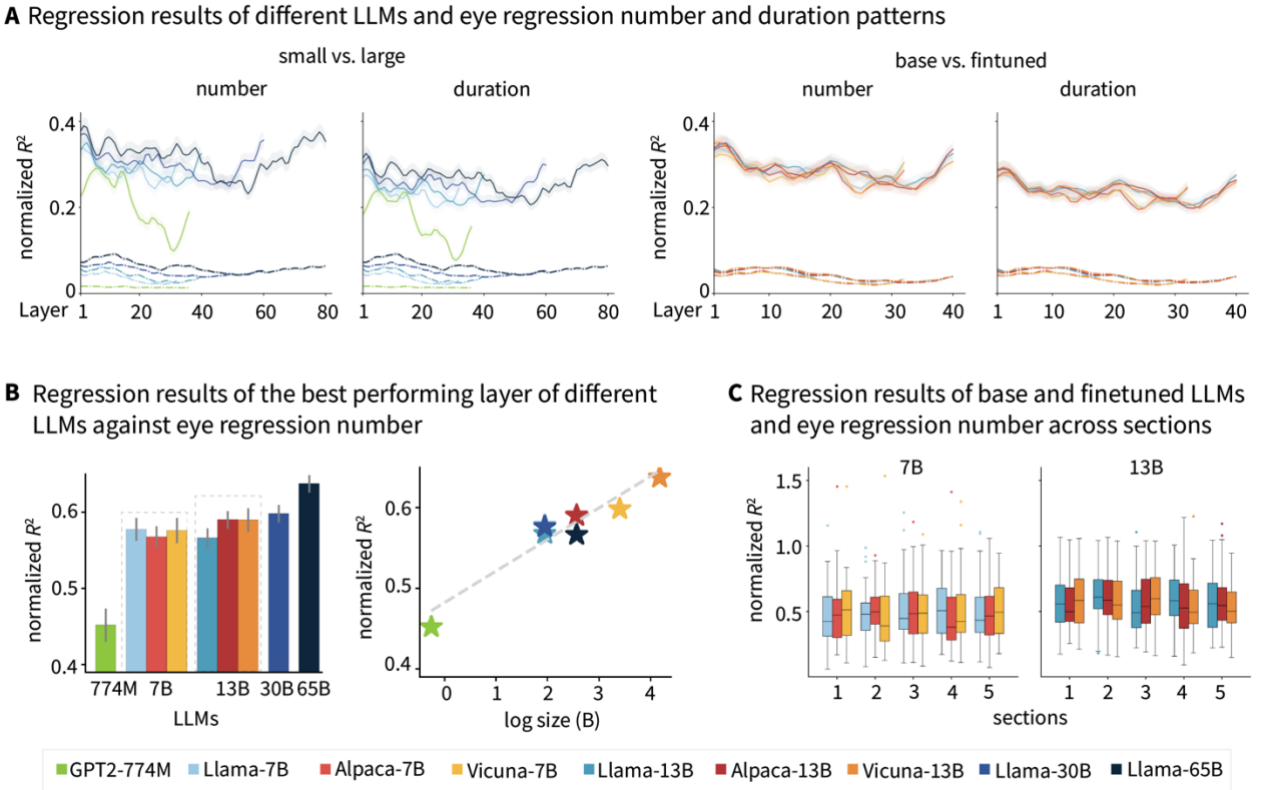

Figure 3. Effects of scaling and finetuning on the alignment between LLMs and human eye regression patterns during naturalistic reading. **A.** Regression results of the attention matrices of different LLMs and the eye regression number and duration matrices for all the stimuli sentences. Dashed lines denote the results from random models of the same sizes. Shaded regions denote standard deviation. **B.** Regression results of the LLMs' best performing layer on the eye regression number patterns and their results on a logarithmic size scale. Error bars denote standard deviation. **C.** Regression results of different LLMs and eye regression number patterns across experimental sections.

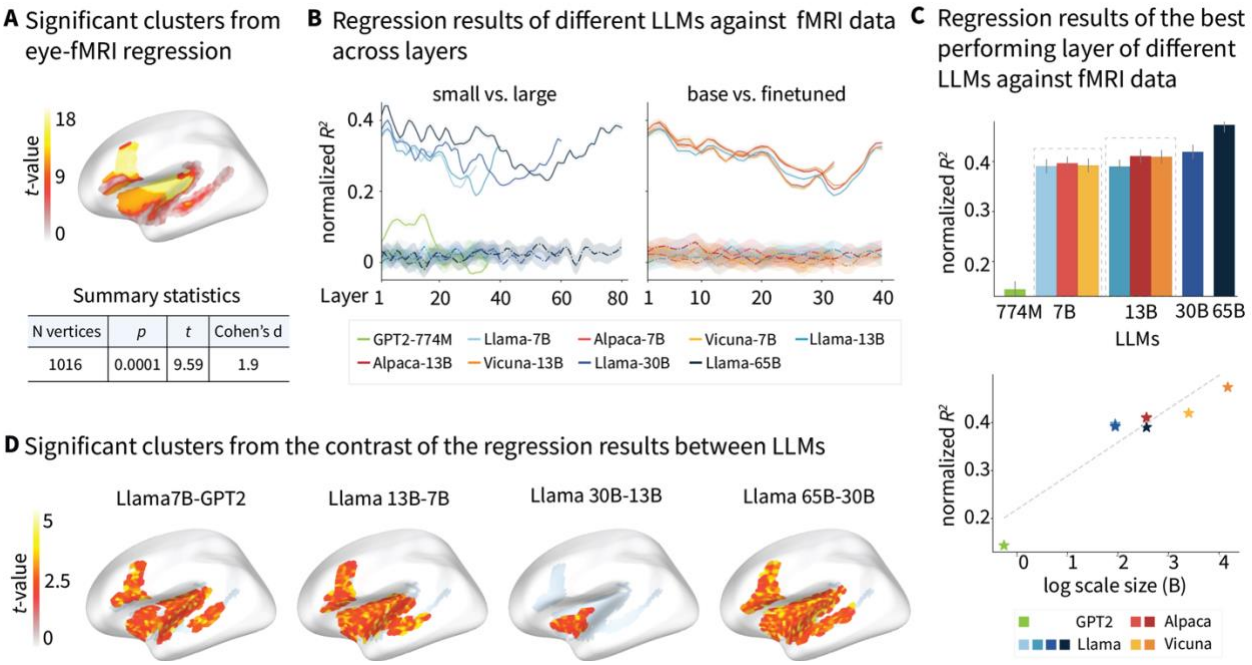

Figure 4. Effects of scaling and finetuning on the alignment between LLMs and human fMRI activity patterns during naturalistic reading. **A.** Significant clusters from the regression analyses of eye regression number and fMRI activity patterns and the summary of the statistics from a cluster-based permutation test. **B.** Averaged regression scores of LLMs and the random controls across subjects at each model layer. Shaded regions denote standard deviation. **C.** Regression results of the LLMs' best-performing layer on the fMRI activity patterns and their results on a logarithmic size scale. Error bars denote standard deviation. **D.** Significant brain clusters from the contrast of regression scores of different LLMs with smaller and larger sizes.

Model 1	Model 2	N vertices	*p*	*t*	Cohen's d
LLaMA-7B	GPT2-large	436	< 0.001	2.53	3.75
LLaMA-13B	LLaMA-7B	477	< 0.001	2.52	3.84
LLaMA-30B	LLaMA-13B	63	0.001	2.35	4.43
LLaMA-65B	LLaMA-30B	576	0.001	2.63	3.87

Table 2. Summary statistics for the significant brain clusters from the contrast of the regression scores of model pairs.

Supplementary information

model-pair	*t*	*p*
GPT2-large vs. LLaMA-7B	8.041	< 0.0001
GPT2-large vs. Alpaca-7B	-4.319	< 0.0001
GPT2-large vs. Vicuna-7B	-1.844	0.071
GPT2-large vs. LLaMA-13B	8.049	< 0.0001
GPT2-large vs. Alpaca-13B	-4.743	< 0.0001
GPT2-large vs. Vicuna-13B	-10.875	< 0.0001
GPT2-large vs. LLaMA-30B	9.349	< 0.0001
GPT2-large vs. LLaMA-65B	9.101	< 0.0001
LLaMA-7B vs. Alpaca-7B	-15.457	< 0.0001
LLaMA-7B vs. Vicuna-7B	-14.804	< 0.0001
LLaMA-7B vs. Alpaca-13B	-9.033	< 0.0001
LLaMA-7B vs. Vicuna-13B	-21.7	< 0.0001
LLaMA-7B vs. LLaMA-13B	0.209	0.835
LLaMA-7B vs. LLaMA-30B	3.178	0.002
LLaMA-7B vs. LLaMA-65B	2.115	0.041
Alpaca-7B vs. Vicuna-7B	4.814	< 0.0001
Alpaca-7B vs. Alpaca-13B	-2.287	0.028
Alpaca-7B vs. Vicuna-13B	-8.399	< 0.0001
Vicuna-7B vs. Vicuna-13B	-12.764	< 0.0001
LLaMA-13B vs. Alpaca-7B	-14.305	< 0.0001
LLaMA-13B vs. Vicuna-7B	-12.452	< 0.0001
LLaMA-13B vs. Alpaca-13B	-9.235	< 0.0001
LLaMA-13B vs. Vicuna-13B	-24.029	< 0.0001
LLaMA-13B vs. LLaMA-30B	3.222	0.002
LLaMA-13B vs. LLaMA-65B	2.176	0.036
Alpaca-13B vs. Vicuna-7B	4.017	< 0.0001
Alpaca-13B vs. Vicuna-13B	-2.069	0.044
LLaMA-30B vs. Alpaca-7B	-16.646	< 0.0001
LLaMA-30B vs. Vicuna-7B	-15.413	< 0.0001
LLaMA-30B vs. Alpaca-13B	-9.918	< 0.0001
LLaMA-30B vs. Vicuna-13B	-24.469	< 0.0001
LLaMA-30B vs. LLaMA2-65B	-0.505	0.632
LLaMA-65B vs. Alpaca-7B	-15.487	< 0.0001
LLaMA-65B vs. Vicuna-7B	-14.498	< 0.0001
LLaMA-65B vs. Alpaca-13B	-9.754	< 0.0001
LLaMA-65B vs. Vicuna-13B	-24.441	< 0.0001

Supplementary Table 1. Pairwise *t*-test results of the next-token prediction (NTP) loss from all pairs of LLMs.

Model-pair	*t*	*p*
Alpaca-7B vs. LLaMA-7B	36.004	< 0.0001
Vicuna-7B vs. LLaMA-7B	34.385	< 0.0001
Alpaca-13B vs. LLaMA-13B	40.278	< 0.0001
Vicuna-13B vs. LLaMA-13B	45.411	< 0.0001

Supplementary Table 2. One sample *t*-test results for the divergence of the attention matrices for the experimental stimuli between the base and finetuned LLMs.

Prefix	Model	*t*	*p*
instruction	LLaMA-7B	40.837	< 0.0001
	Alpaca-7B	41.077	< 0.0001
	Vicuna-7B	43.583	< 0.0001
	LLaMA-13B	39.315	< 0.0001
	Alpaca-13B	44.100	< 0.0001
	Vicuna-13B	44.118	< 0.0001
noise	LLaMA-7B	40.572	< 0.0001
	Alpaca-7B	40.732	< 0.0001
	LLaMA-7B	42.266	< 0.0001
	LLaMA-13B	37.969	< 0.0001
	Alpaca-13B	45.329	< 0.0001
	Vicuna-13B	47.389	< 0.0001

Supplementary Table 3. One sample *t*-test results for the divergence of the attention matrices for stimuli sentences with and without instruction or noise prefixes.

Model-pair	*t*	*p*
GPT2-large vs. LLaMA-7B	3.646	0.005
GPT2-large vs. Alpaca-7B	4.095	0.004
GPT2-large vs. Vicuna-7B	3.906	0.004
GPT2-large vs. Alpaca-13B	2.627	0.041
GPT2-large vs. Vicuna-13B	2.723	0.038
GPT2-large vs. LLaMA-13B	2.710	0.038
LLaMA-7B vs. LLaMA-65B	-3.188	0.011
Alpaca-7B vs. LLaMA-30B	-2.561	0.044
Alpaca-7B vs. LLaMA-65B	-3.721	0.004
Vicuna-7B vs. LLaMA-65B	-3.445	0.006

Supplementary Table 4. Pairwise t-test results for the mean regression scores of trivial patterns against LLMs' attention matrices across layers.

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	0.340

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