Improve Language Model and Brain Alignment via Associative Memory

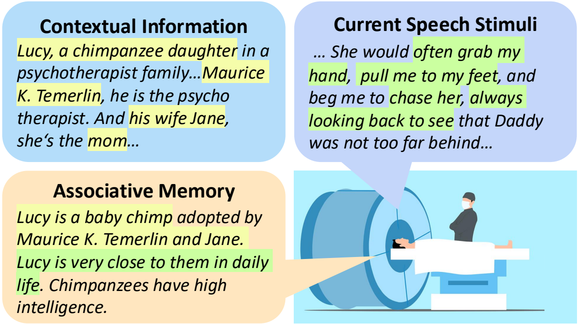
Congchi Yin Yongpeng Zhang Xuyun Wen Piji Li[*]
[1] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics
[2] The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education
{congchiyin,pjli}@nuaa.edu.cn

Abstract

Associative memory engages in the integration of relevant information for comprehension in the human cognition system. In this work, we seek to improve alignment between language models and human brain while processing speech information by integrating associative memory. After verifying the alignment between language model and brain by mapping language model activations to brain activity, the original text stimuli expanded with simulated associative memory are regarded as input to computational language models. We find the alignment between language model and brain is improved in brain regions closely related to associative memory processing. We also demonstrate large language models after specific supervised fine-tuning better align with brain response, by building the *Association*[1] dataset containing 1000 samples of stories, with instructions encouraging associative memory as input and associated content as output.

1 Introduction

Human language comprehension is a complicated process widely involving multiple brain functions (GESCHWIND, 1965; Aboitiz and Garcıéa V., 1997). Previous studies (Dronkers et al., 2007; Binder, 2015) have confirmed that Wernicke's area and Broca's area are essential in speech comprehension and language production. More relevant regions are found and subdivided to match corresponding functions through functional Magnetic Resonance Imaging (fMRI) scans in later work (Poremba et al., 2004; Gourévitch et al., 2008; Chang et al., 2011). Among all the functions related to language comprehension, *associative memory* (Anderson and Bower, 2014) plays an indispensable role, serving as the key to linking together related concepts and pieces of information.



Lucy, a chimpanzee daughter in a psychotherapist family…Maurice K. Temerlin, he is the psycho therapist. And his wife Jane, she's the mom…

… She would often grab my hand, pull me to my feet, and beg me to chase her, always looking back to see that Daddy was not too far behind…

Lucy is a baby chimp adopted by Maurice K. Temerlin and Jane. Lucy is very close to them in daily life. Chimpanzees have high intelligence.

Figure 1: An example of how associative memory works when subject listens to speech.

The human associative memory system (Mayes et al., 2007; Eichenbaum, 2017) is responsible for encoding, monitoring, and retrieving diverse components of information, including basic perceptions (i.e. semantic memory (Binder and Desai, 2011)), personal experiences (i.e. episodic memory (Tulving et al., 1972)), and contextual details (i.e. working memory (Baddeley, 2003)). While associative memory integrates multiple outside stimuli (visual, auditory, sensory, etc.), we primarily focus on associative memory during human language comprehension in this study, particularly in a task involving passively listening to continuous speech. The core effect of associative memory in language comprehension is to help build connections between diverse concepts. Human is capable of retrieving relative concepts that facilitate language understanding instinctively (Schank, 1972). Although associative memory involves complex formation and interaction within the biological brain, at a high level it can be considered as the integration of associated concepts (McNamara and Magliano, 2009). For example, as shown in Figure 1, the orange box is what associative memory involves. The content in the purple box indicates the speech stimuli that the subject is receiving. The blue box represents previous phonetic stimuli that the subject heard a few moments ago.

[*] Corresponding author.
[1] https://github.com/lemonsis/Association

Findings of the Association for Computational Linguistics: ACL 2025, pages 986–999
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Large language models like GPT-4 (Achiam et al., 2023) have shown remarkable natural language understanding and generation ability. They follow a next word prediction pattern, which is similar to human's manner of processing text information (Caucheteux et al., 2023; Antonello and Huth, 2024). Previous research (Jain and Huth, 2018; Toneva and Wehbe, 2019; Caucheteux and King, 2020; Goldstein et al., 2022) has confirmed the activations of language models can be linearly mapped to the activity of human brain when receiving the same text stimuli. This finding provides a powerful tool for investigating the alignment between biological brain and language models. For example, Caucheteux et al. (2023) showed such alignment can be improved by introducing future words prediction. Moussa et al. (2024) fine-tuned speech model with brain-relevant semantics to improve its alignment to brain activity. However, as far as we know, few studies explore the associative memory in human language processing.

In this paper, we investigate two research questions: (1) Will simulating associative memory in brain language processing improve the alignment between language models and the human brain? (2) Can we improve the alignment between language model and brain by instructing language models to generate associative content? We design the following experiment steps to answer these questions. First, the alignment between language model activation and human brain activity (i.e. brain score) is evaluated when they receive the same text stimuli as input. Following previous studies (Caucheteux and King, 2020), traditional language model GPT-2 (Radford et al., 2019) is selected. We also try large language model LLaMA-2 (Touvron et al., 2023) for comparison. For the first research question, associative memory is considered as the integration of context and associated knowledge, as the example shown in Figure 1. Data augmentation with simulated associative memory is performed to the original text stimuli. The activation of language model with augmented sentences as input is mapped to the original brain activity. The brain score of regions related to associative memory (e.g. medial temporal lobe (MTL)) is recorded in comparison to the original brain score. For the second research question, we build an instruction tuning dataset *Association* containing 1000 samples with story paragraphs and instructions encouraging associative memory as input, associated content as output. *Association* is applied in the supervised fine-tuning (SFT) of base large language model. The brain score of language model after SFT is re-evaluated to see whether the score of relevant regions is improved. Such improvement will become strong evidence suggesting the alignment between large language models and human brain can be improved by instructing language models to generate associative content.

Our contributions can be summarized as follows:

- We find associative memory simulation via data augmentation is capable of improving language model and brain alignment.
- We release *Association*, an instruction tuning dataset containing 1000 samples for investigating associative memory by supervised fine-tuning large language models.
- We demonstrate that fine-tuning a large language model with instructions that promote associative memory can enhance its alignment with brain activity.

2 Related work

Language Models and Brain Alignment Previous studies have mapped word-level embeddings to fMRI or MEG signals (Mitchell et al., 2008; Huth et al., 2012, 2016). Jain and Huth (2018); Toneva and Wehbe (2019); Goldstein et al. (2022); Caucheteux et al. (2022) indicated that human brain combines information of previous words to predict next words and such prediction is increasingly contextual along the hierarchy by extracting activations from different layers in language models. Such prediction has also been proven to span multiple timescales(Goldstein et al., 2020; Caucheteux et al., 2023). Antonello et al. (2024) further analyzed the mapping of large language models to the human brain. Some studies seek to find the reasons behind the alignment between language models and human brain. Caucheteux et al. (2021) factorized language model activations into lexical, compositional, syntactic and semantic representations. Wehbe et al. (2014); OOTA et al. (2024) investigated the specific linguistic properties and brain regions that contribute to such alignment.

Associative Memory It's a fundamental cognitive process enabling the linking of related information (Anderson and Bower, 2014). Early research (Marr et al., 1991) laid the groundwork by proposing theoretical models that describe how the hippocampus could facilitate the storage and retrieval

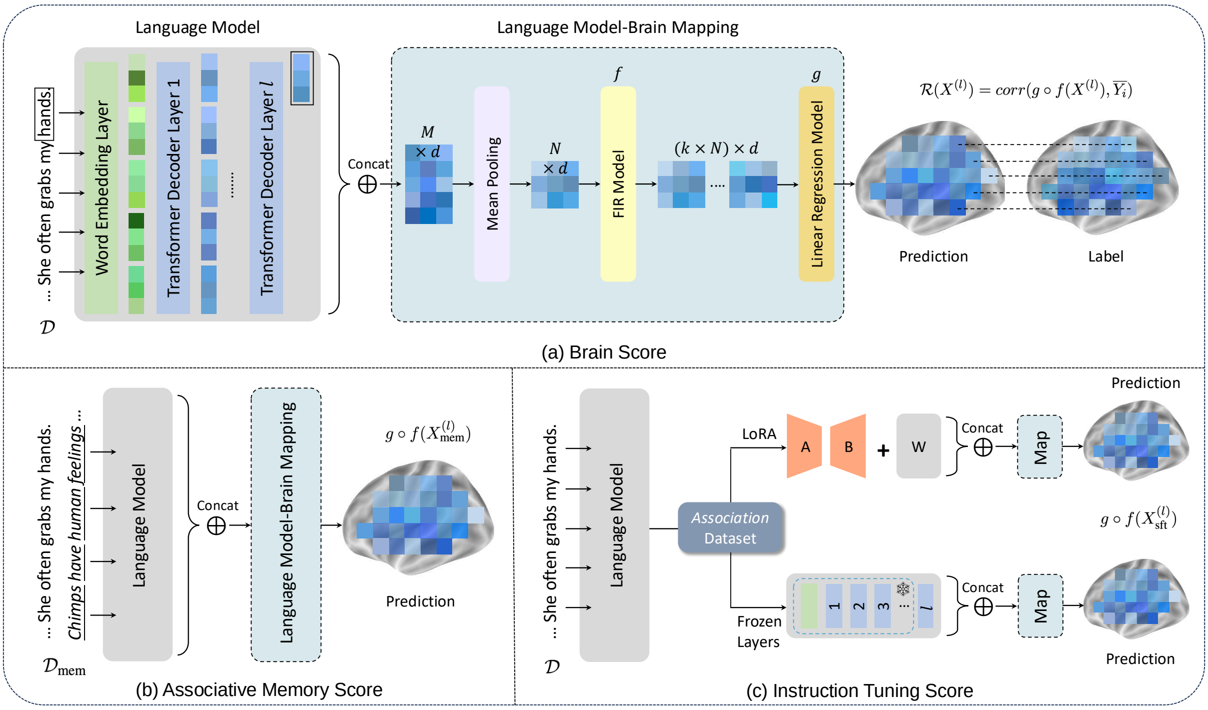

Language Model
Language Model-Brain Mapping
... She often grabs my hands
$\mathcal{D}$
Word Embedding Layer
Transformer Decoder Layer 1
Transformer Decoder Layer l
Concat
$M \times d$
Mean Pooling
$N \times d$
f
FIR Model
$(k \times N) \times d$
g
Linear Regression Model
$\mathcal{R}(X^{(l)}) = corr(g \circ f(X^{(l)}), \overline{Y_i})$
Prediction
Label
(a) Brain Score
... She often grabs my hands.
Chimps have human feelings ...
$\mathcal{D}_{\text{mem}}$
Language Model
Concat
Language Model-Brain Mapping
$g \circ f(X^{(l)}_{\text{mem}})$
Prediction
(b) Associative Memory Score
... She often grabs my hands.
$\mathcal{D}$
Language Model
Association Dataset
LoRA
A
B
+
W
Concat
Map
Prediction
$g \circ f(X^{(l)}_{\text{sft}})$
Frozen Layers
1
2
3
l
Concat
Map
Prediction
(c) Instruction Tuning Score

Figure 2: General framework of calculating brain score, associative memory score, and instruction tuning score.

of associative memories. Subsequent empirical studies (McClelland et al., 1995) demonstrated the importance of synaptic plasticity and the role of long-term potentiation (LTP) in associative learning and memory consolidation. With neuroimaging techniques, some studies have identified key brain regions involved in associative memory, including the medial temporal lobe, the prefrontal cortex, and their interactions (Sperling et al., 2001). Moreover, computational models (Bogacz and Brown, 2003) have been developed to simulate associative memory, providing a deeper understanding of how neurons could support complex associative tasks.

3 Methods

3.1 Overview

We first introduce the method of evaluating language model and brain alignment by mapping language model activations to fMRI signals. The extent of alignment is referred as **brain score**. Then we expand the original dataset with simulated associative memory, and recalculate the brain score to identify newly activated brain regions. Finally, an instruction tuning dataset *Association* is proposed and applied in the supervised fine-tuning (SFT) of large language model (LLM). LLM after SFT is reevaluated on the original dataset to explore whether brain score is improved by instructing large language models to generate associative content.

3.2 Brain Score Calculation

We aim to investigate the alignment between language models and human brain when they process text information. To better map language models to brain, auto-regressive models with left-to-right attention are selected, as brain can't get access to future information like bidirectional Transformers. The activations of language models are mapped to the fMRI recordings with the same text stimuli as input. More precisely, given a sequence $S = (s_1, \ldots, s_i, \ldots, s_M)$ of M words from dataset $\mathcal{D}$, the output embedding x_i of s_i in the l-th layer of language model can be written as

$$x_i^{(l)} = W_l W_{l-1} \cdots W_0(s_{i-c}, \ldots, s_{i-1}, s_i), \quad (1)$$

where W_l indicates the transformation weight matrix of l-th Transformer layer, c is a hyper-parameter deciding the length of history information fused into current word embedding. The representation of word sequence S through language model is denoted as $X^{(l)} = concat(x_1, \ldots, x_M) \in \mathbb{R}^{M \times d}$.

Let $Y = \Psi(S) \in \mathbb{R}^{N \times v}$ be the brain activity elicited by the same word sequence S, which is collected through N continuous fMRI frames. v is the number of voxels in brain. Analysis is conducted for one particular voxel $Y_i \in \mathbb{R}^N$ because it can be easily extended to whole brain. Since fMRI signals are inherently noisy, the average blood-oxygen-level-dependent (BOLD) signal $\overline{Y_i}$ across total T subjects for each voxel is considered.

$$\overline{Y_i} = \frac{1}{T}\sum_{j=1}^{T} Y_{ij} \tag{2}$$

As fMRI is sampled discretely with fixed time intervals (a.k.a. TR) and the sampling frequency is usually much lower than word rate, we take the mean pooling of language model activations to match N fMRI frames, as shown in Figure 2. To mitigate the gap of delayed BOLD responses, we follow previous work (Huth et al., 2016; Affolter et al., 2020) and apply a finite impulse response (FIR) model. For fMRI frame $i \in [1 \dots N]$, the temporal transformation f_i is formally defined as

$$\begin{aligned} f_i : \mathbb{R}^{N\times d} &\to \mathbb{R}^{k\times d} \\ x &\mapsto concat(\widetilde{x}_i, \widetilde{x}_{i-1}, \dots, \widetilde{x}_{i-k+1}) \end{aligned} \tag{3}$$

where

$$\widetilde{x}_i = \frac{1}{m} \sum_{\substack{m\in[\![1\dots M]\!] \\ \mathcal{T}(m)=i}} x_m^{(l)}, \tag{4}$$

$$\begin{aligned} \mathcal{T} : [\![1\dots M]\!] &\to [\![1\dots N]\!] \\ m &\mapsto \min_{k\in[\![1\dots N]\!]} |t_{y_k} - t_{x_m}|, \end{aligned} \tag{5}$$

with $\widetilde{x}$ taking the mean pooling of word embeddings between successive fMRI TRs, k a hyperparameter controlling the delay of FIR feature x, $(t_{x_1}, \dots, t_{x_M})$ the timings of words onsets and $(t_{y_1}, \dots, t_{y_N})$ the timings of N fMRI frames.

After achieving temporal alignment between $X^{(l)}$ and $\overline{Y_i}$ through f, we seek to find a linear model $g \in \mathbb{R}^d$ to map language model activations to brain activity. Ridge regression with ℓ_2-regularization is learned to predict brain activity:

$$\operatorname*{argmin}_{g} \sum_{i\in I_{\text{train}}} \left(\overline{Y_i} - g^T f(X^{(l)})\right)^2 + \lambda\|g\|^2. \tag{6}$$

Finally, similar to previous work (Yamins and DiCarlo, 2016), brain score $\mathcal{R}(X^{(l)})$ is defined as correlation between predicted brain activity and original brain activity. Pearson correlation score $corr(\cdot,\cdot)$ is applied to measure such connection and brain score of each voxel can be written as

$$\mathcal{R}(X^{(l)}) = corr(g \circ f(X^{(l)}), \overline{Y_i}). \tag{7}$$

Moreover, we design a novel brain score ceiling test to explore the limitation of explainable and predictable brain signals. In each iteration, all the subjects hearing the same word sequence are randomly separated into two parts, part A and part B. Instead of using language model activations to predict brain activity, one part of subjects' brain activity Y_A are used to predict the other part of subjects' brain activity Y_B through linear model g. All the brain activity is averaged across corresponding subjects to reduce noise. The brain score ceiling for i-th voxel of brain is calculated as

$$\mathcal{R}_{\text{ceiling}} = corr(g(\overline{Y_{A_i}}), \overline{Y_{B_i}}). \tag{8}$$

3.3 Data Augmentation with Simulated Associative Memory

The concrete mechanism of associative memory in the biological brain is complex, involving the interaction of neurons from multiple brain regions. To investigate whether the alignment between language model and human brain can be improved by associative memory, we don't directly simulate its process in the brain. Instead, we concretize the content of associative memory, namely what people may associate during a passively story hearing test, as natural language input to language models.

The original dataset $\mathcal{D}$ is expanded with simulated associative memory. The dataset after data augmentation is denoted as $\mathcal{D}_{\text{mem}}$. Specifically, sentence-level and word-level associative memory simulation is tried respectively. Sentence-level data augmentation involves grammatically complete sentences, typically focusing on a single aspect of association. In contrast, word-level data augmentation includes words or phrases that capture multiple aspects of associative memory. Both human and GPT-4 annotations are applied. Human annotators are asked to write down what they associate when receiving certain text stimuli that trigger associative memory. GPT-4 is not able to decide when and where to add associative memory content like human, so we give clear instructions and let GPT-4 return associated words or sentences based on the context and its knowledge every four sentences of the original text stimuli. Examples of different data augmentation methods are shown in Appendix B. Considering the latency of fMRI signals, all the expanded content is put at the end of the sentences that trigger associative memory. Onsets are all set to the same as the last word's offset,

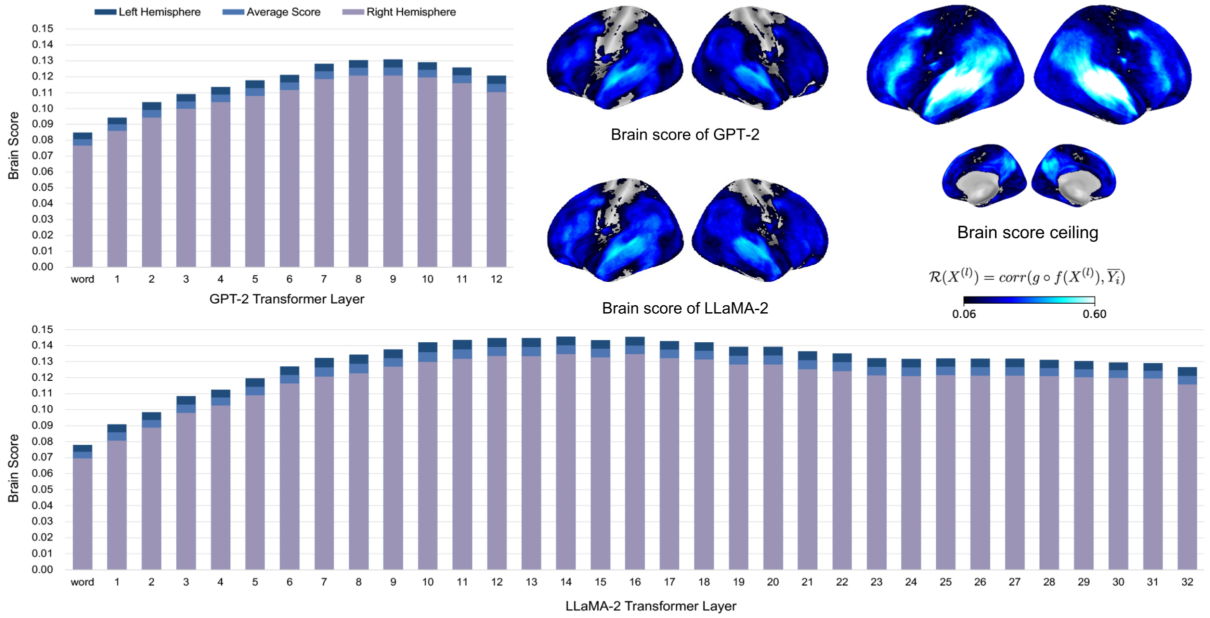


Figure 3: Brain score of different layers with visualization. The colorbar refers to value of brain score.

as if the associative memory forms simultaneously when subject receives specific text stimuli.

Word sequence $S_{\text{mem}} \in \mathcal{D}_{\text{mem}}$ is used to compute the activation of language model $X^{(l)}_{\text{mem}}$. Brain activity maintains $\overline{Y_i}$ as the subjects listen to original dataset D. Following the same process as mentioned before, brain score with simulated associative memory is computed through

$$\mathcal{R}(X^{(l)}_{\text{mem}}) = corr(g \circ f(X^{(l)}_{\text{mem}}), \overline{Y_i}). \tag{9}$$

The associative memory score $\mathcal{F}$ of one specific voxel is defined as the difference between brain score with associative memory and original brain score

$$\mathcal{F}(X^{(l)}) = \mathcal{R}(X^{(l)}_{\text{mem}}) - \mathcal{R}(X^{(l)}). \tag{10}$$

3.4 Instruct LLM to Generate Associative Content

Different from language models with limited parameters like GPT-2, recent large language models with huge number of parameters can be trained to follow instructions through supervised fine-tuning. We build an instruction tuning dataset *Association* containing paragraphs of stories with instructions encouraging associative memory as input, word-level associated content as answers. More details about the dataset are introduced in Appendix A.3 and examples are shown in Appendix B. We build the *Association* dataset to investigate whether the alignment between language models and human brain can be improved by instructing large language model to generate associative content. The improvement is reflected by observing the increment of brain score on certain brain regions.

Two supervised fine-tuning methods are tried: low-rank adaptation (LoRA) (Hu et al., 2021) and frozen layers finetuning. LoRA applies two trainable low-rank matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ during fine-tuning and the original weight $W \in \mathbb{R}^{d\times k}$ of LLM is frozen. When the training finishes, the original weight is replaced as $W + BA$. For frozen layers fine-tuning, layers before the l-th layer are frozen during supervised finetuning. All the parameters in layer l and layers after l-th layer are trainable. The weight matrix of l-th Transformer layer after supervised finetuning is denoted as $W^{(l)}_{\text{sft}}$ and its output is denoted as $X^{(l)}_{\text{sft}}$. Following previous methods, brain score of supervised fine-tuned model is $\mathcal{R}(X^{(l)}_{\text{sft}})$. We define instruction tuning score $\mathcal{M}$ as the growth percentage of supervised finetuned model compared to base model:

$$\mathcal{M}(X^{(l)}) = (\mathcal{R}(X^{(l)}_{\text{sft}}) - \mathcal{R}(X^{(l)}))/\mathcal{R}(X^{(l)}). \tag{11}$$

4 Experimental Setups

4.1 Datasets

We use the publicly accessible "Narratives" dataset (Nastase et al., 2021) which contains fMRI record-

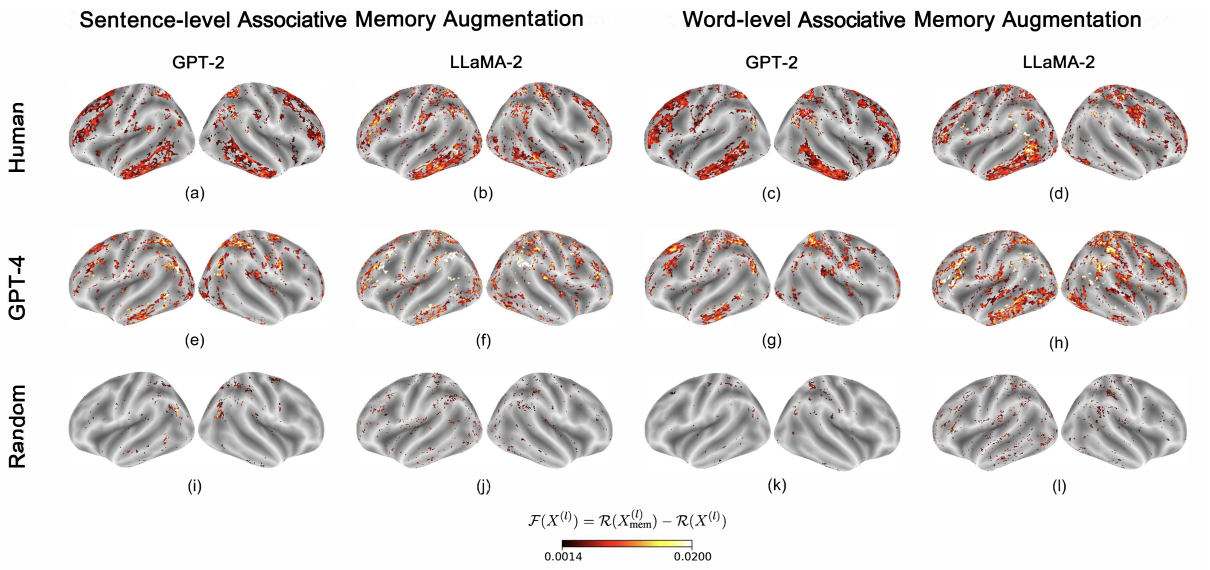

Sentence-level Associative Memory Augmentation
Word-level Associative Memory Augmentation
GPT-2
LLaMA-2
GPT-2
LLaMA-2
Human
GPT-4
Random
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
$\mathcal{F}(X^{(l)}) = \mathcal{R}(X^{(l)}_{\text{mem}}) - \mathcal{R}(X^{(l)})$
0.0014
0.0200

Figure 4: Associative memory score of sentence-level and word-level data augmentation.

ings of 345 individuals listening to 27 spoken English stories. After filtering short articles, 15 stories with corresponding fMRI images are selected for experiments. More details are in Appendix A.3.

4.2 Cortical Parcellation

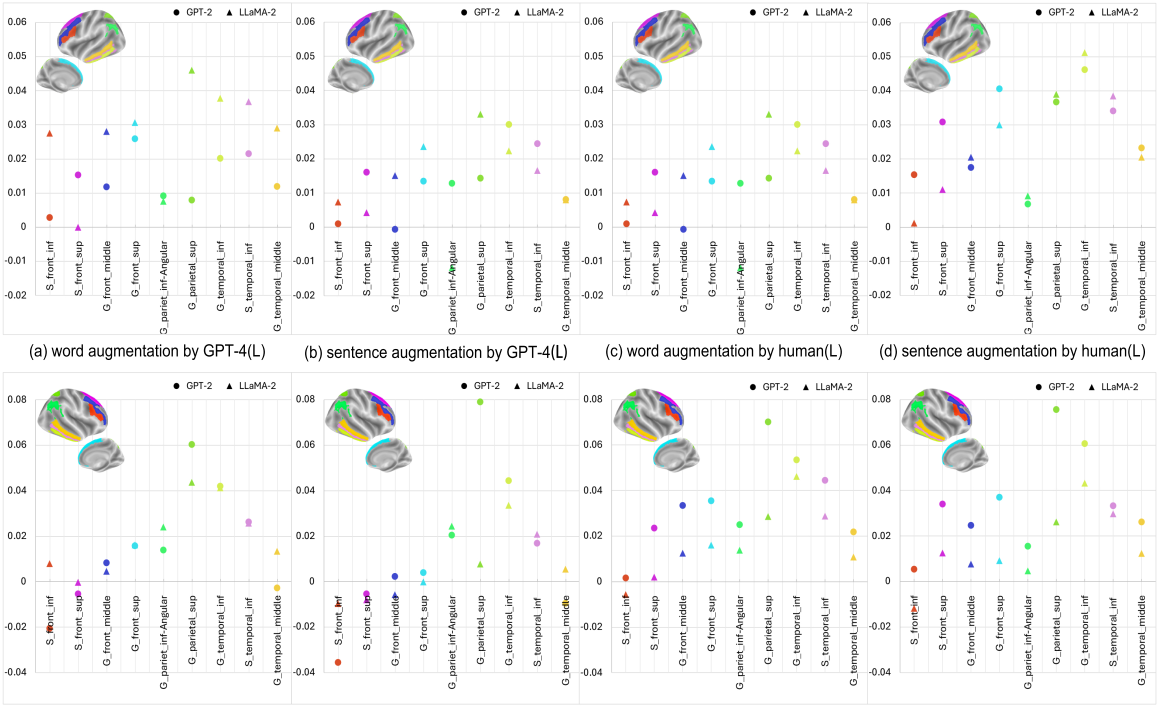
Nine brain regions are selected to analyze brain score changes in different regions of interests (ROIs). Besides inferior temporal gyrus, inferior temporal sulcus and middle temporal gyrus that are known to contribute to associative memory, we also explore regions related to speech processing and working memory, as associative memory involves interaction with working memory in a storying hearing task. These regions include middle and superior frontal gyrus, inferior and superior frontal sulcus, superior parietal lobule, angular gyrus. Details are presented in Appendix A.1.

5 Results and Analysis

First, we analyze the brain score of different layers for GPT-2 and LLaMA-2 models. Second, dataset with associative memory augmentation is applied to explore activated brain regions. Finally, LLaMA-2 fine-tuned in an instruction dataset *Association* is evaluated to verify the improved alignment between language models and human brain.

5.1 Brain Score Comparison

The brain score is calculated by averaging all the fMRI voxel of all the test subjects. Two kinds of language model embeddings are considered: non-contextual word embedding and contextual embedding of each Transformer layer. Results are shown in Figure 3. Overall, LLaMA-2 gets a higher brain score than GPT-2 due to larger number of parameters, more training corpus, and better representation ability. Brain score of left hemisphere is higher than that of right hemisphere for different layers of both models. It's consistent with previous researches (Halpern et al., 2005; Riès et al., 2016) on lateralization of brain function that Broca's and Wernicke's areas related to the production and comprehension of speech are found exclusively on the left hemisphere. Scores calculated through word embedding are significantly lower than other layers for lack of contextual information. Brain scores of both hemispheres peak at the ninth layer for GPT-2 model, achieving an average score of 0.126, which satisfies previous conclusion (Caucheteux and King, 2022) that the activations of $l = n_{\text{layers}} \times 2/3$ layer best fit brain activity. However, we find layer best predicting brain activation becoming shallow for LLaMA-2. The fourteenth out of thirty-two layers reaches the highest brain score of 0.146 and 0.135 for left and right hemispheres, respectively. Relative studies (Durrani et al., 2021; Sajjad et al., 2022; Zhang et al., 2023) on investigating representation inside Transformer-based language models reveal that lower layers are dominated by lexical concepts, whereas middle and higher layers better represent core-linguistic concepts. But why middle front layers best aligned with the brain still remains unexplored. Based on the above findings, we apply the ninth layer and fourteenth layer of GPT-2 and LLaMA-2 separately



(a) word augmentation by GPT-4(L) (b) sentence augmentation by GPT-4(L) (c) word augmentation by human(L) (d) sentence augmentation by human(L)

(e) word augmentation by GPT-4(R) (f) sentence augmentation by GPT-4(R) (g) word augmentation by human(R) (h) sentence augmentation by human(R)

Figure 5: Associative memory score of specific regions of interests (ROIs). (L) and (R) refer to left hemisphere and right hemisphere, respectively. The color of dot corresponds to the color of specific ROI.

for all the following experiments.

We map each voxel's brain score of GPT-2 and LLaMA-2 to brain surface and plot figures for better visualization. Figure 3 also shows the brain maps of the highest possible brain score (i.e. brain score ceiling) under current dataset and linear regression model. Brain scores are witnessed over a distributed and bilateral cortical network, peaking in middle and superior temporal gyrus, middle and superior temporal sulcus, as well as in the supramarginal and the infero-frontal cortex.

5.2 Associative Memory Score

Associative memory score measures the difference between original brain score and brain score with simulated associative memory. We investigate the associative memory score under various settings. Results are shown in Figure 4. Sub-figures (a) to (d) show associative memory score with human annotated associative memory augmentation, while sub-figures (e) to (h) are with GPT-4 augmented associative memory. Besides, we apply random word-level and sentence-level data augmentation as a control group to demonstrate that the improvement of brain score benefits from associative memory. Results are shown in Sub-figures (i) to (l). The random augmentation is conducted in the following manner. For word-level augmentation, we apply GPT-4 to generate 100 unrepeated verbs, nouns, adjectives, and randomly select words among the set of a total of 300 words. For sentence-level augmentation, we directly apply GPT-4 to randomly generate sentences. Sub-figures (a), (b), (e), (f), (i), (j) show sentence-level associative memory augmentation, and sub-figures (c), (d), (g), (h), (k), (l) show word-level augmentation.

Generally speaking, large and continuous regions of the brain, including some areas of frontal gyrus, frontal sulcus and parietal lobule gain increase in brain score ranging from 0.0014 to 0.02. Since these regions get a relatively low brain score without simulated associative memory stimulation, such a gain in brain score is considerable. Moreover, we find random data augmentation leads to none and even negative growth of brain score, which supports the improvement of alignment is caused by introducing associative memory. From Figure 4, it's noticed that word-level augmentation leads to better performance compared to sentence-level augmentation on both models. We think compared to sentence-level augmentation, word-level augmentation probably benefits from multi-aspect association with less introduced noise like proposition and conjunction. Nouns, adjectives, and verbs

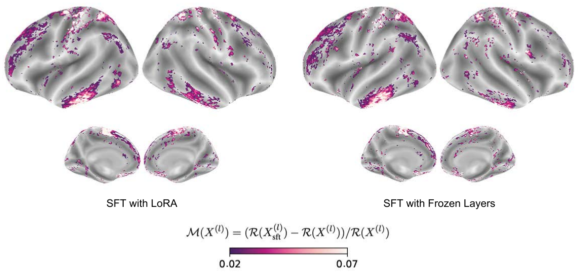
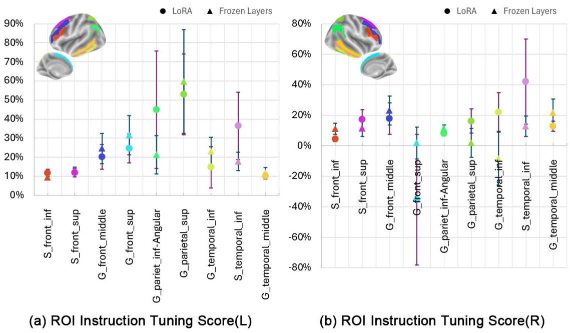

SFT with LoRA
SFT with Frozen Layers
$\mathcal{M}(X^{(l)}) = (\mathcal{R}(X^{(l)}_{\text{sft}}) - \mathcal{R}(X^{(l)}))/\mathcal{R}(X^{(l)})$
0.02
0.07

Figure 6: Instruction tuning score after supervised fine-tuning with two different methods.



(a) ROI Instruction Tuning Score(L)
(b) ROI Instruction Tuning Score(R)

Figure 7: Subject-level instruction tuning score of specific regions of interests (ROIs).

contain more intensive information. This finding is also consistent with previous neuroscience study (Schwering and MacDonald, 2020), which indicates that associative memory is conceptualized by the unit of the word. Human annotation earns higher associative memory score than GPT-4, because GPT-4 can't decide where to generate associative content like human annotators. LLaMA-2 performs better than GPT-2 model under most cases with wider activated brain regions and higher score. Overall, the alignment between two tested language models and human brain gets significantly improved with word-level human-annotated associative memory.

We also compute associative memory score on brain regions of interests (ROIs) and the results are shown in Figure 5. Sub-figure (a) to (d) shows the cases of left hemisphere and sub-figure (e) to (h) shows right hemisphere. Since areas related to associative memory in language comprehension mainly distribute in the left hemisphere (Smith et al., 1998), results on figure (a) to (d) are more confident. Improvements in brain scores were observed across nine regions of interest (ROIs) associated with associative memory, as well as in areas related to speech processing and working memory, ranging from 0 to 0.05. Such trend of improvement is generally consistent with each of the nine ROIs for both models. LLaMA-2 model gets a higher associative memory score than GPT-2 model for most ROIs in the left hemisphere. Superior and middle frontal gyrus, superior parietal lobule related to working memory, inferior and superior frontal sulcus related to speech processing, medial temporal lobe (MTL) area related to associative memory all get improved on both word-level and sentence-level augmentation dataset.

5.3 Instruction Tuning Score

Common instruction tuning will not lead to improvement of brain score (Gao et al., 2023). We explore whether brain score can be improved by instructing language model to simulate associative memory. The *Association* dataset is built, which contains 1000 training samples with story paragraphs and prompts encouraging associative memory as input, associated content as output. We try two different supervised fine-tuning methods, LoRA and frozen layers finetuning, for LLaMA-2 and the results are shown in Figure 6 and 7.

As shown in Figure 6, LLaMA-2 after supervised fine-tuning with both methods shows 2% to 7% gain in regions related to associative memory (i.e. medial temporal lobe (MTL)), which indicates the alignment between language model and brain is improved by associative memory instructed tuning. We also calculate instruction tuning score in specific ROIs and results are shown in Figure 7. Different from Figure 6 where instruction tuning score $\mathcal{M}$ is computed by averaging all the subjects' voxel, which is equivalent to viewing all the subjects' brain activity as one, in Figure 7 we first calculate instruction tuning score for each subject and then average these scores. Results correspond to 95% confidence intervals (CIs) across all test subjects are shown in the figure. Superior parietal lobule related to working memory gets the highest score of 50% to 60%. As to associative memory, instruction tuning score in medial temporal lobe (i.e. G_temporal_inf, S_temporal_inf, G_temporal_middle) also get significantly improved across hundreds of subjects.

6 Conclusion

In this paper, we explore whether the alignment between language model and human brain could be improved by introducing associative memory in a passive story hearing task. By defining brain score, associative memory score, and instruction tuning score in experiments, we answer two research questions: The alignment between language model and human brain can be improved (1) with simulated associative memory (2) by instructing language models to generate associative content.

Limitations

The “Narratives” dataset contains fMRI recordings when subjects listen to English spoken stories. Language and cultural background of the participants and the story should be considered. Therefore, the results could not fully cover all types of languages and cultures. Moreover, annotators involved in associative memory data augmentation may possess different language and cultural backgrounds compared to the subjects in “Narratives” dataset. Even with the same language and cultural background, the fMRI recordings do not perfectly match the associative memory content. This discrepancy will inevitably introduce noise. We hope dataset recording associative memory of subjects is made to better investigate associative memory in the human brain using language models.

Ethics Statement

In this paper, we explore associative memory in the human brain listening to speech through large language models. The proposed *Association* dataset is for non-profit research usage. Experiments are conducted on public accessible cognitive dataset “Narratives” with the authorization from its respective maintainers. The dataset has been de-identified by providers and is used for researches only.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.62476127, No.62106105), the Natural Science Foundation of Jiangsu Province (No.BK20242039), the Basic Research Program of the Bureau of Science and Technology (ILF24001), the Fundamental Research Funds for the Central Universities (No.NJ2023032), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

Francisco Aboitiz and Ricardo Garcıéa V. 1997. The evolutionary origin of the language areas in the human brain. a neuroanatomical perspective. *Brain Research Reviews*, 25(3):381–396.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. 2020. Brain2word: decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*.

John R Anderson and Gordon H Bower. 2014. *Human associative memory*. Psychology press.

Richard Antonello and Alexander Huth. 2024. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79.

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2024. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36.

Alan Baddeley. 2003. Working memory and language: An overview. *Journal of communication disorders*, 36(3):189–208.

Jeffrey R Binder. 2015. The wernicke area: Modern evidence and a reinterpretation. *Neurology*, 85(24):2170–2175.

Jeffrey R Binder and Rutvik H Desai. 2011. The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11):527–536.

Rafal Bogacz and Malcolm W Brown. 2003. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, 13(4):494–524.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pages 1336–1348. PMLR.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Charlotte Caucheteux and Jean-Rémi King. 2020. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, pages 2020–07.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

Edward F. Chang, Erik Edwards, Srikantan S. Nagarajan, Noa Fogelson, Sarang S. Dalal, Ryan T. Canolty, Heidi E. Kirsch, Nicholas M. Barbaro, and Robert T. Knight. 2011. Cortical Spatio-temporal Dynamics Underlying Phonological Target Detection in Humans. *Journal of Cognitive Neuroscience*, 23(6):1437–1446.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15.

Nina F Dronkers, Odile Plaisant, Marie Therese Iba-Zizen, and Emmanuel A Cabanis. 2007. Paul broca's historic cases: high resolution mr imaging of the brains of leborgne and lelong. *Brain*, 130(5):1432–1441.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Howard Eichenbaum. 2017. Time (and space) in the hippocampus. *Current opinion in behavioral sciences*, 17:65–70.

Changjiang Gao, Shujian Huang, Jixing Li, and Jiajun Chen. 2023. Roles of scaling and instruction tuning in language perception: Model vs. human attention.

NORMAN GESCHWIND. 1965. DISCONNEXION SYNDROMES IN ANIMALS AND MAN1. *Brain*, 88(2):237–237.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2020. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *BioRxiv*, pages 2020–12.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Boris Gourévitch, Régine Le Bouquin Jeannès, Gérard Faucon, and Catherine Liégeois-Chauvel. 2008. Temporal envelope processing in the human auditory cortex: Response and interconnections of auditory cortical areas. *Hearing Research*, 237(1):1–18.

Marnie E Halpern, Onur Güntürkün, William D Hopkins, and Lesley J Rogers. 2005. Lateralization of the vertebrate brain: taking the side of model systems. *Journal of Neuroscience*, 25(45):10351–10357.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31.

David Marr, David Willshaw, and Bruce McNaughton. 1991. *Simple memory: a theory for archicortex*. Springer.

Andrew Mayes, Daniela Montaldi, and Ellen Migo. 2007. Associative memory and the medial temporal lobes. *Trends in cognitive sciences*, 11(3):126–135.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

Danielle S McNamara and Joe Magliano. 2009. Toward a comprehensive model of comprehension. *Psychology of learning and motivation*, 51:297–384.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. 2024. Improving semantic understanding in speech language models via brain-tuning. *arXiv preprint arXiv:2410.09230*.

Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):250.

SUBBAREDDY OOTA, Manish Gupta, and Mariya Toneva. 2024. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Amy Poremba, Megan Malloy, Richard C Saunders, Richard E Carson, Peter Herscovitch, and Mortimer Mishkin. 2004. Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature*, 427(6973):448–451.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stéphanie K Riès, Nina F Dronkers, and Robert T Knight. 2016. Choosing words: Left hemisphere, right hemisphere, or both? perspective on the lateralization of word retrieval. *Annals of the New York Academy of Sciences*, 1369(1):111–131.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3082–3101, Seattle, United States. Association for Computational Linguistics.

Roger C Schank. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631.

Steven C Schwering and Maryellen C MacDonald. 2020. Verbal working memory as emergent from language comprehension and production. *Frontiers in human neuroscience*, 14:68.

Edward E Smith, John Jonides, Christy Marshuetz, and Robert A Koeppe. 1998. Components of verbal working memory: evidence from neuroimaging. *Proceedings of the National Academy of Sciences*, 95(3):876–882.

Reisa A Sperling, Julianna F Bates, Andrew J Cocchiarella, Daniel L Schacter, Bruce R Rosen, and Marilyn S Albert. 2001. Encoding novel face-name associations: A functional mri study. *Human brain mapping*, 14(3):129–139.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Endel Tulving et al. 1972. Episodic and semantic memory. *Organization of memory*, 1(381-403):1.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Daniel LK Yamins and James J DiCarlo. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. 2023. Emergent modularity in pre-trained transformers. *arXiv preprint arXiv:2305.18390*.

A Implemention Details

A.1 Cortical Parcellation

The latest version of Destrieux atlas (Destrieux et al., 2010) is applied for cortical parcellation, which leads to 74 regions per hemisphere. To reveal the associative memory in the human brain listening to speech, nine related regions are selected for experiments, including inferior frontal sulcus (S_front_inf), superior frontal sulcus (S_front_sup), middle frontal gyrus (G_front_middle), superior frontal gyrus (G_front_sup), angular gyrus (G_pariet_inf-Angular), superior parietal lobule (G_parietal_sup), inferior temporal gyrus (G_temporal_inf), inferior temporal sulcus (S_temporal_inf), middle temporal gyrus (G_temporal_middle).

A.2 Models and Hyper-parameters

For fMRI data, we apply the AFNI-nosmooth preprocessing step for the Narratives dataset. Analyses are conducted on cortical voxels projected onto the surface and morphed onto an "fsaverage6" template brain. We use 'RidgeClassifierCV' regressor from scikit-learn (Pedregosa et al., 2011) to predict the continuous features and align language models to brain, with 10 possible penalization values log-spaced between 10^{-1} and 10^{8}. The linear model is evaluated on held out data, using 20 cross-validation for averaged score across all subjects and 5 cross-validation for brain score of each subject.

For language models, we choose the small version of GPT-2 and LLaMA-2 with 7B parameters from Huggingface[2]. The supervised finetuning of LLaMA-2 with LoRA or frozen layers is trained for 2 epochs with 10^{-4} learning rate. All experiments are conducted on NVIDIA A100-80G GPUs.

A.3 Associative Memory Data Augmentation and The *Association* Dataset

In the data augmentation process of simulated associative memory, ten annotators are hired to make both word-level and sentence-level annotation. Annotators are asked to write down what they associate when receiving certain text stimuli that trigger associative memory. The hired annotators are Asian undergraduate students with English as their second language. Seven of them are male and three are female. Each annotator is assigned with two or three articles for labeling and is paid about 40 dollars. The annotators are informed that the data will be used for non-profit research.

We also make GPT-4 version of data augmentation with the assistance of `gpt-4-1106-preview` API. The instruction tuning dataset *Association* contains 1000 training samples with encouraging associative memory prompts and word-level association responses. It's composed of sentences from filtered stories of "Narratives" and sentences randomly picked from ROCStories dataset (Mostafazadeh et al., 2016). The *Association* dataset is annotated with the help of GPT-4 through `gpt-4-1106-preview` API, more examples are shown in Appendix B.

B Case Study

In this part, we will take a deeper look into how data augmentation with associative memory is performed, and how the instruction tuning dataset *Association* is made. More examples and cases are given and analyzed.

B.1 Data Augmentation

Table 1 shows four examples of data augmentation. Four different augmentation methods are applied, including word-level augmentation by GPT-4 and human annotators, sentence-level augmentation by GPT-4 and human annotators. Since GPT-4 generates association content every three or four sentences, while human annotators add association stuff according to their ideas, the places of data augmentation are different in most cases. To facilitate a more effective comparison, we present the sentences that have been flagged by both GPT-4 and human annotators below.

B.2 Instruction Tuning Dataset

The *Association* dataset consists of input and output pairs as training samples. As shown in Table 2, input content is made up of instruction and story paragraph. The instruction is prompts encouraging models to generate associative content, and the story paragraph contains sentences extracted from stories in "Narratives" dataset and ROCStories dataset. The output is word-level associated content. Table 2 displays some samples randomly selected from *Association* dataset.

[2] `https://huggingface.co`

Original Sentences	**Method**	**Data Augmentation**
This is Los Angeles. And it's the height of summer. In a small bungalow off of La Cienega, Clara serves homemade chili and chips in red plastic bowls – wine in blue plastic.	word-level, GPT-4	heat, bustling, cozy, spicy, casual, colorful
	word-level, human	hot, comfortable
	sentence-level, GPT-4	The sun blazes down on a cozy home in LA where a casual summer gathering unfolds.
	sentence-level, human	Clara uses plastic bowls of different colors to make thing in a bungalow.
Louis when I first started here. People told him, "Oh no, no she is white man, she's white, she sounds white she's white," and he, convinced, having never met me, that I was black. Well as it turns out, he was right.	word-level, GPT-4	debate, racial identity, assumptions, voice, community perceptions, prejudice, correctness, self-awareness, revelation
	word-level, human	debate, truth, revelation, surprise, race
	sentence-level, GPT-4	In a St. Louis debate about my ethnicity, a stranger's conviction about my race challenged the assumptions tied to my voice,and he was correct.
	sentence-level, human	Debates about the author being black and white were going on long before he came to St Louis Missouri community.
Jane named her Lucy and brought her home on a commercial airline, carried in a bassinet, her face covered with a lacy blanket. We were blissfully unaware of the complexities we were creatin on the day Lucy came home. So the baby was a day or two old.	word-level, GPT-4	Lucy, adoption, chimpanzee, travel, naivety, complexities, infancy, integration, new beginnings
	word-level, human	Lucy, home, expectation, experiments
	sentence-level, GPT-4	Lucy's journey veiled in the innocence of infancy and a lacy blanket, commenced with a flight to an uncharted life, while the Temerlins remained oblivious to the intricate future unfolding from their decision.
	sentence-level, human	Lucy would increase the complexity of the experiments but I think she will make it.
Um, there's a lot of guys in army gear, um, shooting. It's very chaotic, can't really make out any faces, or, really people. Um, and then, so, that, uh, cuts out really quickly and you see this man kind of like start out of the bed.	word-level, GPT-4	battle, soldiers, military, chaos, abrupt anonymous, awakening
	word-level, human	soldier, war, army, grass field, chaotic
	sentence-level, GPT-4	The man wakes up abruptly, haunted by the chaos of battle.
	sentence-level, human	This seems to be the case with the wars in the middle east.

Table 1: Examples of word-level and sentence-level data augmentation with associative memory.

Input		Output
Instruction	**Story Paragraph**	
I'll give you some sentences, you have to perform related association with words.	Sheldon slowly walked into the restaurant, eying the decor suspiciously. His roommate Leonard pushed past him and asked the hostess for a table for two. As they were led to their chairs, Sheldon began to protest yet again.	quirky, cautious, skeptical, friends, dining, impatient
Given some sentences, you are supposed to make related associations and output words.	You know, I think I may have misjudged this restaurant. I won't go out on a limb, but I think we may be looking at my new Tuesday hamburger.	surprise, reconsideration, hamburger, potential favorite
Given a batch of sentences, you need to execute the process of interlinking them based on their relevance with words.	He zipped up Barney's bag and handed it back to him. Quinn followed Barney down the concourse in total confusion. Magic trick? Why wouldn't he tell her what was in the box? She tried to interrogate him as they sat in front of the gate, but he refused to spill the beans.	mystery, secrecy, curiosity, travel, frustration, companionship
You need to engage in divergent thinking based on the sentences I provide, and give me whatever words comes to your mind.	That's fair, that's what we charge in our country. After waiting for their turn to board, they marched down the jetway and onto the plane. George struggled to get into his window seat and fit his bag down by his feet.	equality, travel, patience, boarding, cramped, luggage, discomfort
You will get a set of sentences, and you need to associate some related content with words.	Vinny poked at it with his fork. What's this over here? The cook looked at him in disbelief. You've never heard of grits? Sure, sure, I've heard of grits, I've just never actually… seen a grit before. Go ahead honey, aren't you going to try it? You first, she said with a smile.	curiosity, skepticism, southern cuisine, breakfast, humor
Given some sentences, you are supposed to make related associations and output words.	Anna was filling her bird feeders. But a chunk of suet fell onto the ground. Her dog rushed over and lapped it up! Anna was astonished. She had no idea dogs loved bird food!	surprise, dogs, birds, feeding, accidental, curiosity
Human tend to think relative stuff when receiving text information. Imagine you're human and expand the following sentences with words.	Sam's dog Rex escaped from their yard. Sam was distraught. He went out calling for Rex. Then he saw Rex come running up the street! Sam was so relieved, he almost cried!	worry, search, reunion, joy, pet, relief

Table 2: Training samples randomly picked from the *Association* dataset.