Encoding Brain Regions with Sentiment-Relevant Circuits in LLMs

Nursulu Sagimbayeva (nusa00001@uni-saarland.de)
Saarland University
Dota Tianai Dong (tianai.dong@mpi.nl)
Max Planck Institute for Psycholinguistics

Abstract

Large language models (LLMs) generate representations that effectively predict brain responses to natural language, yet the specific circuits within LLMs that drive this alignment remain largely unexplored. We here apply techniques from mechanistic interpretability (MI) to identify LLM circuits (i.e., attention heads) causally relevant to sentiment processing and assess their impact on LLM–brain alignment. Our results show that removing sentiment-related attention heads leads to a greater decrease in alignment with language-processing brain regions compared to random head removal, although this difference does not reach statistical significance. Ongoing work aims to further improve LLM circuit identification in naturalistic settings, enabling more precise mapping of circuits to plausible brain mechanisms and ultimately providing deeper insights into LLM–brain alignment.

Keywords: LLM-brain alignment; mechanistic interpretability

Introduction

Large language models (LLMs) excel at generating rich representations of natural language, achieving state-of-the-art performance across NLP tasks while also showing promising alignment with human brain responses during language processing (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). This alignment—where model activations predict neural activity patterns—suggests that LLMs may capture some computational principles of human language processing. However, the specific circuits within these models that give rise to brain-like representations remain poorly characterized, limiting them as computational proxies for studying neural language processing (Tuckute et al., 2024).

Mechanistic interpretability (MI) offers a promising approach to address this limitation by uncovering the internal workings of neural networks (Cammarata et al., 2020; Elhage et al., 2021). This emerging field examines model weights and activations to identify specific computational components—such as attention heads—that implement particular behaviors (Hanna et al., 2023; Olsson et al., 2022). In this paper, we investigate whether these MI techniques can be adapted to identify circuits (i.e., attention heads) in LLMs that causally contribute to their alignment with brain activity.

We focus on sentiment as an ideal test case , as it is a pervasive feature of natural language processed by both humans and LLMs. Recent work by Tigges et al. has demonstrated that sentiment is represented linearly across various LLMs: a subset of attention heads encodes sentiment along a single direction, with opposing extremes representing positive and negative sentiment. Critically, ablating these heads causes damage to LLMs' ability to perform sentiment-related tasks.

Building on these findings, we first examine whether MI approaches—typically applied to cherry-picked toy tasks—can handle complex naturalistic language stimuli central to neuroscience research. We then identify sentiment-relevant attention heads in a range of language models. Lastly, we test a specific causal hypothesis: if sentiment-processing attention heads in LLMs contribute to brain alignment, selectively patching them with counterfactual activations should reduce LLM-brain alignment more than patching randomly selected heads.

Methods

Datasets and models

To extend the circuit identification approach to a naturalistic setting, we construct a corrupted narrative dataset *HarryPotterSentiment* based on the Harry Potter chapter that is used as stimuli in the fMRI experiments (Wehbe et al., 2014). Our dataset includes 92 examples, such as:

> "Harry had never believed he would meet a boy he *hated* more than Dudley."
> "Harry had never believed he would meet a boy he *loved* more than Dudley."

We label each sentence and its counterpart as "positive" or "negative". Since we are constrained to using the same text as fMRI stimuli, we ended up with an imbalanced dataset: 29 "positive" and 63 "negative" labels. Our experiments use *GPT2* Small, and two task-tuned alternatives: *sentiment-tuned GPT2*[1] trained on tweet sentiment extraction dataset [2], and *summarization-tuned GPT2* trained on CNN news[3]. To evaluate model accuracies, we use the following template:

> "Rate the sentiment of this sentence as positive or negative: [sentence]. The sentiment of this sentence is [model answer]"

If the logits for the correct label are higher than for the wrong label, we count it as a correct answer, and then sum up the number of correct guesses over the whole dataset for each model to arrive at model accuracy.

Identifying attention heads and causal evaluation

We use activation patching (Meng et al., 2022), a causal intervention technique that precisely identifies which subset of attention heads is responsible for encoding some type of information—for example, sentiment information. For our *HarryPotterSentiment* dataset, we perform forward passes with both the original and distorted texts, storing their activations in clean and corrupted caches, respectively. By iteratively patching activations from the distorted text back to the original text for each model component, we observe how this affects the model's predictions (i.e., measuring the accuracy drop on the Harry Potter dataset), thereby revealing which head is most important for sentiment classification. We select heads that

[1] https://huggingface.co/riturajpandey739/gpt2-sentiment-analysis-tweets

[2] https://huggingface.co/datasets/mteb/tweet_sentiment_extraction

[3] https://huggingface.co/gavin124/gpt2-finetuned-cnn-summarization-v2

lead to a more than 5% drop in the original accuracy, and we refer to them as ”circuits”. The 5% threshold is inspired by Tigges et al., where they use the threshold of 5%-or-greater damage to the logit difference. Notably, we patch all sentence tokens, irrespective of whether the token expresses a sentiment.

LLM-brain encoding

We use the standard brain recordings of 8 subjects reading Harry Potter (Wehbe et al., 2014), sampled at a TR of 1.49 seconds per session, capturing activity levels of all voxels (around 28000). We build an encoding model from the last layer to predict brain matrices of participants reading Harry Potter, using ridge-regularized linear functions. After training through 4-fold cross-validation with nested parameter selection, we evaluate using voxel-based mean Pearson correlation between predicted and actual fMRI values.

Results

Finding and evaluating sentiment-relevant circuits

We begin by evaluating the models' baseline performance on *HarryPotterSentiment* dataset. Figure 1 (left) shows that all models perform poorly in a naturalistic setting and achieve low baseline accuracies, with GPT2 Small obtaining the highest accuracy of 64 out of 92. This might be because we have an imbalanced dataset, and GPT2 Small seems to have a bias towards the negative class (87/92 ”negative” predictions), whereas other models seem to be biased towards the positive class (78/92 ”positive” predictions for sentiment-tuned and 79/92 for summarization-tuned models).

However, we assume that the relative drop in accuracy after activation patching is more interpretable than the baseline accuracy, since it allows us to compare model performance before and after intervention. Our results support this: patching individual attention heads in GPT2 Small did not affect task accuracy, whereas in other models, modifying specific heads led to a notable performance drop (e.g., head 7.5 in GPT2 Sentiment and head 9.2 in GPT2 Summarization). We hypothesize that this is because sentiment information is not localized to a single attention head in the pre-trained model, whereas in task-tuned models, it seems to be concentrated in a few heads. Patching all identified circuits in GPT2 Sentiment led to the largest performance drop—approximately a 70% decrease from the original accuracy—indicating that these attention heads encode information relevant to sentiment analysis. Figure 1 (right) illustrates the accuracy drop caused by patching each individual head in the GPT2 Sentiment model. Since our results show that activation patching can most notably disrupt model performance for GPT2 Sentiment, we chose to proceed with this model for brain encoding.

Brain encoding with sentiment-relevant circuits

To test whether sentiment-relevant attention heads in GPT2 Sentiment contribute to brain alignment, we compare alignment between brain activations encoded in three conditions:

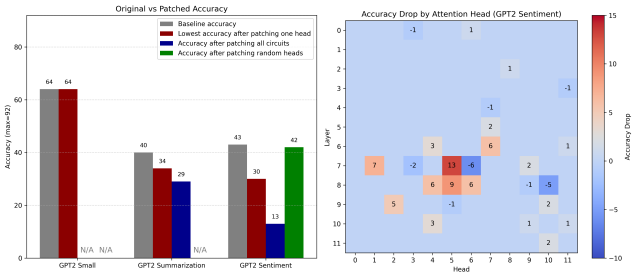

Figure 1: Left: Model performance for three models before and after activation patching. Right: Breakdown of accuracy drop by individual attention head in GPT2 Sentiment.

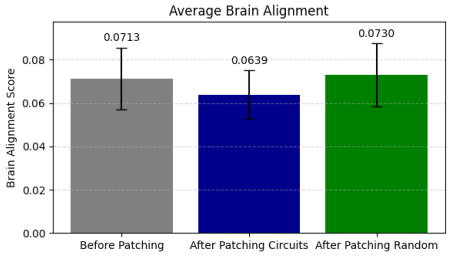

Figure 2: Results for brain encoding averaged over 8 subjects.

using the original model embeddings, and when circuits or random heads are patched with counterfactual dataset activations. In the random patching condition, we select the same number of heads from the same layers as the identified circuits. Figure 1 (left: green bar) shows that patching random attention heads does not result in a meaningful performance drop on the task.

Figure 2 shows the contrast between brain alignment scores in the language network[4] across conditions. As hypothesized, we observe that alignment decreases more after patching the sentiment-relevant heads than after patching random heads. However, these results were not statistically significant, possibly due to incorrect circuits identification or the selection of inappropriate brain regions. Contrary to expectations, patching random heads slightly increased brain alignment, reaching statistical significance. In the future, we aim to explore this phenomenon in greater depth.

Limitations and future work

As the next step, we plan to further improve circuit identification in naturalistic settings by fine-tuning LLMs with narrative data. While we focus the sentiment analysis task in this paper, we believe the approach could be extended to other NLP tasks, and aim to explore them. We also aim to extend our analysis to other naturalistic datasets, such as Narratives (Nastase et al., 2021), expand the dataset, and the range of studied models.

[4]The following regions were used: 'PostTemp', 'AntTemp', 'AngularG', 'IFG', 'MFG', 'IFGorb', 'pCingulate', 'dmpfc

Acknowledgements

We would like to acknowledge Mariya Toneva for the guidance throughout the project. We thank Alexander Koller for supporting the opportunity to present our poster at the conference. We thank Michael Hanna for his feedback.

References

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., ... Lim, S. K. (2020). Thread: Circuits. *Distill*. (https://distill.pub/2020/circuits) doi: 10.23915/distill.00024

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. (https://transformer-circuits.pub/2021/framework/index.html)

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... others (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Hanna, M., Liu, O., & Variengien, A. (2023). How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, *36*, 76033–76060.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 17359–17372). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., ... Hasson, U. (2021). The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), 250. Retrieved from https://doi.org/10.1038/s41597-021-01033-3 doi: 10.1038/s41597-021-01033-3

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... others (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

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

Tigges, C., Hollinsworth, O. J., Geiger, A., & Nanda, N. (n.d.). Linear representations of sentiment in large language models, 2023. *URL https://arxiv. org/abs/2310.15154*.

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

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, *47*(2024), 277–301.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *in press*.