# INVESTIGATING NATURAL SCENE REPRESENTATIONS SHARED BETWEEN LARGE LANGUAGE MODELS AND THE HUMAN VISUAL SYSTEM

**Yota Kawashima**
Bernstein Center for Computational Neuroscience Berlin
yotakawashima@gmail.com

**Supervisor: Adrien Doerig**
Freie Universität Berlin

March 27, 2025

**ABSTRACT**

What kind of information does the brain extract from natural scenes? Large Language Models (LLMs) provide a useful representational format to study natural scene representations in the brain. Specifically, when LLMs receive captions of natural scene images, their internal representations can predict brain responses to the corresponding images. However, it remains unclear which aspects of LLM's representations are important for this alignment. To address this question, we will investigate which abilities of LLM correlate with predictive power. Using publicly available LLMs and a functional magnetic resonance imaging (fMRI) dataset, we will analyse the predictive power in relation to LLM benchmarks for formal linguistic competence (e.g. grammar) and functional linguistic competence (e.g. world knowledge and reasoning). We hypothesize that functional linguistic competence will correlate with predictive power more strongly than formal linguistic competence.

## 1 Introduction

What kind of information does the brain extract from natural scenes? Natural scenes contain complex contextual associations among objects, such as object co-occurrence statistics [Oliva and Torralba, 2007]. Several studies have suggested different brain regions responsible for different aspects of contextual associations, offering insights into how the brain represents natural scenes. However, it remains unclear how the human visual system collectively encodes complex contextual associations.

Large Language Models (LLMs) provide a useful framework for studying natural scene representations in the brain (Figure 1). Given the LLM's ability to process context information, Doerig et al. [2022] explored whether natural scene representations in LLMs aligned with those in the human visual system. The study used LLM representations of natural scene image captions to predict brain responses to the corresponding natural scene images (brain alignment scores). Their findings suggest a similarity between LLM and brain representations.

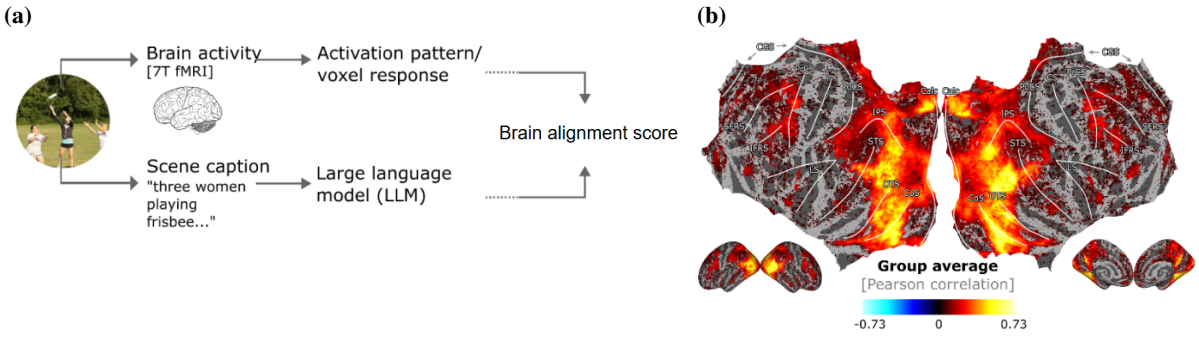


Figure 1: Large Language Models (LLMs) share natural scene representations with the human visual system. **(a)** Analysis pipepline. LLM representations of natural scene image captions can predict functional magnetic resonance imaging (fMRI) responses to the corresponding natural scene images. **(b)** Correlation between predicted and actual responses. Modified from Doerig et al. [2022]

However, brain alignment scores alone do not reveal which aspects of LLM representations are important for this alignment to the brain. Doerig et al. [2022] tried to address this issue by shuffling image captions and assessing whether LLM representations could still predict brain responses. The study showed that the predictive power remained intact, suggesting that language grammar was not a key factor in the alignment. But what specific aspects of LLM representations contribute to this alignment?

LLM benchmarks offers a way to interpret the alignment more directly. For example, AlKhamissi et al. [2025] examined the human language system by tracking brain alignment scores and LLM's performance on different linguistic tasks. The study found a stronger correlation between brain alignment scores and LLM's performance on English grammar questions (formal linguistic competence) than on world knowledge and reasoning questions (functional linguistic competence). This suggests that the aligned representations are more relevant to formal linguistic competence. Brain alignment scores, together with LLM benchmarks, explain which aspects of LLM representations align with brain's representations.

In this project, we aim to investigate natural scene representations shared between LLMs and the human visual system through LLM benchmarks. Specifically, we will analyze which type of LLM's linguistic competence–formal or functional–correlates more strongly with brain alignment scores. We will use publicly available LLM models and benchmarks from AlKhamissi et al. [2025] and will analyse LLM representations to predict functional magnetic resonance imaging (fMRI) data in the Natural Scenes Dataset (NSD) [Allen et al., 2022]. Based on the result in Doerig et al. [2022], we hypothesize that brain alignment scores will correlate with functional linguistic competence more strongly than formal linguistic competence.

## 2 Method

We will apply the analysis reported in AlKhamissi et al. [2025] to the Natural Scenes Dataset (NSD), a fMRI dataset recorded during a visual recognition task [Allen et al., 2022]. We will follow AlKhamissi et al. [2025], including the selection and evaluation of LLMs. Here, we will provide the information sufficient to understand this study. See the original papers for details.

We will analyze Pythia, a publicly available set of autoregressive language models from different training checkpoints [Biderman et al., 2023]. We will evaluate 34 training checkpoints from each of the five models with a different model size (410M, 1B, 1.4B, 2.8B, and 6.9B).

We will use eight benchmarks to evaluate the formal and functional linguistic competence of LLMs. We will use two benchmarks for formal linguistic competence: BLiMP [Warstadt et al., 2019] and SyntaxGym [Gauthier et al., 2020]. These benchmarks measure the model's ability to choose a grammatically sound sentence from a given pair of sentences (acceptability judgment). We will use six benchmarks for different aspects of functional linguistic competence: ARC [Clark et al., 2018] for world knowledge, SocialIQA [Sap et al., 2019] for social reasoning, PIQA [Bisk et al., 2020] for physical reasoning, and WinoGrande [Sakaguchi et al., 2020] and HellaSwang [Zellers et al., 2019] for common sense reasoning. These benchmarks measure the model's ability to answer multiple choice questions. We will use the benchmark scores reported in AlKhamissi et al. [2025].

We will analyze functional magnetic resonance imaging (fMRI) data from the Natural Scenes Dataset (NSD) [Allen et al., 2022]. The NSD provides 7T fMRI responses from 8 participants who performed a visual recognition task. The task involved natural scenes images from Microsoft Common Objects in Context (COCO) dataset [Lin et al., 2014]. We will use the preprocessed fMRI data from Doerig et al. [2022].

We will measure brain alignment scores through a voxel-wise linear encoding model. We will train a ridge regression model to predict voxel-wise fMRI responses from LLM representations for each participant's data. We will measure the Pearson correlation between the predicted and actual brain responses on a hold-out test dataset. The test dataset includes shared 515 images that all participants saw three times. We will use the remaining images to train the regression model.

Finally, we will evaluate correlations between brain alignment scores and formal/functional linguistic competence. We will average the brain alignment scores across all voxels and participants to get one brain alignment score for each model size and training checkpoint. We will compute the Pearson correlation between the averaged brain alignment scores and each LLM benchmark scores for each model size.

## 3 Result

We will update this section after analyzing the data.

## 4 Discussion

We will update this section after analyzing the data.

## 5 Code Availability

We will upload code on the following Github repository: `https://github.com/yotaKawashima/LLM-encoding-model-FU`.

## References

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How LLMs outgrow the human language network. *arXiv [cs.CL]*, March 2025.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126, January 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *arXiv [cs.CL]*, April 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proc. Conf. AAAI Artif. Intell.*, 34(05):7432–7439, April 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv [cs.AI]*, March 2018.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Visual representations in the human brain are aligned with large language models. *arXiv [cs.CV]*, September 2022.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv [cs.CV]*, May 2014.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends Cogn. Sci.*, 11(12):520–527, December 2007.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Proc. Conf. AAAI Artif. Intell.*, 34(05):8732–8740, April 2020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. SocialIQA: Commonsense reasoning about social interactions. *arXiv [cs.CL]*, April 2019.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. *arXiv [cs.CL]*, December 2019.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.