Carnegie Mellon University

Divergences between Language Models and Human Brains

Yuchen Zhou Emmy Liu Graham Neubig Michael Tarr Leila Wehbe

Divergences between LMs and Human Brains

- Brain activity data collected using MEG while participants listened to or read narratives
- Language model embeddings generated from GPT-2 XL (1.5B) and Llama-2 7B
- Ridge regression with cross-validation used to compute prediction error per word

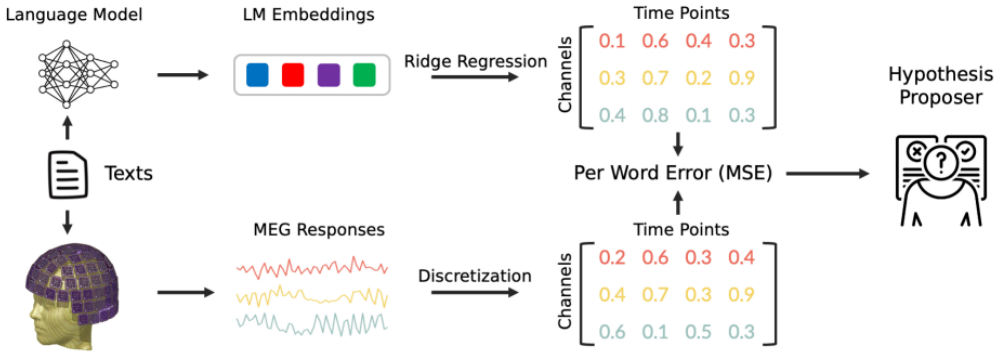


Image Source: https://www.mrn.org/collaborate/elekta-neuromag-meg

Carnegie Mellon University

Automatic Hypothesis Proposer

- Natural language hypotheses that explain the differences between two text corpora (D0, D1) are generated using a proposer-verifier system (Zhong et al., 2023)
- GPT-3 (Brown et al. ,2020) serves as the proposer, generating hypotheses on how corpus D0 differs from D1
- FLAN-T5-XXL (Chung et al., 2022) acts as the verifier, evaluating and ranking hypotheses

Hypotheses

Two topics: **Social/Emotional Intelligence** and **Physical Commonsense**

Table 1: Top 10 hypotheses generated from the best layer of GPT-2 XL on the Harry Potter dataset

Hypothesis	Validity	*p*-value
have a high level of emotional intensity	0.250	0.010
involve complex sentence structures or grammar	0.250	0.015
include emotional language or descriptions	0.238	0.008
have a high level of tension or conflict	0.237	0.023
have characters using body language or non-verbal cues	0.225	0.032
are emotionally charged, making it challenging for language models to accurately interpret the intended tone or sentiment	0.213	0.020
include conflicts between characters	0.200	0.035
have characters interacting with their environment	0.188	0.059
have complex sentence structures	0.175	0.081
have dialogue between characters with varying emotions	0.175	0.022

Validity measures the difference in certainty that the hypothesis is true between the two corpora, see (Zhong et al., 2023) for details

Can targeted fine-tuning improve LM-brain alignment?

Hypothesis

LMs fail to predict brain responses due to limited representations of social/emotional understanding and physical world knowledge

Prediction

Fine-tuning on domain-specific data will enhance LMs' alignment with human brain activity

Fine-tuning

Each multiple-choice option is concatenated with the question to format it as a language modeling task

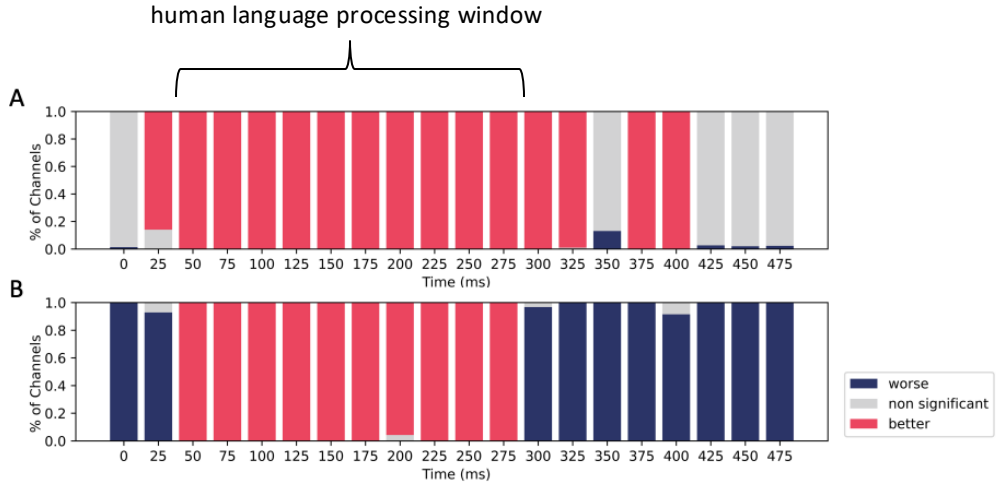
Table 3: Datasets for Fine-Tuning with Sample Questions and Answers (Correct Answer in Bold)

Dataset	Type	Num train	Options	Sample question	Sample answers
Social IQa	Social/Emotion	33.4k	3	Sydney had so much pent up emotion, they burst into tears at work. How would Sydney feel afterwards?	1. affected 2. **like they released their tension** 3. worse
PiQA	Physical	16.1k	2	When boiling butter, when it's ready, you can	1. Pour it onto a plate 2. **Pour it into a jar**

Performance comparison of the base model with fine-tuned models

Social/Emotional Intelligence

Physical Commonsense



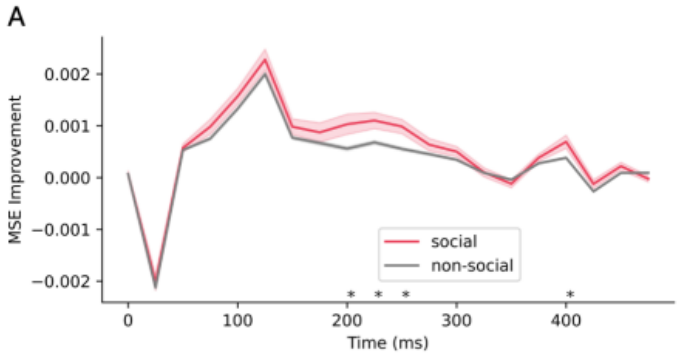
y-axis shows the percentage of MEG channels in the fine-tuned model with better, worse, or non-significantly different performance (measured by Pearson correlation) compared to the base model

x-axis is the time relative to word onset

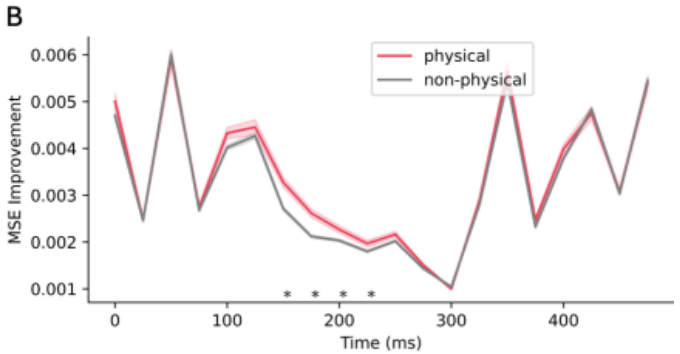
Carnegie Mellon University

Fine-tuning improves alignment more for words annotated with that category

Social/Emotional Intelligence



Physical Commonsense



Comparison of improved MSE between (A) social and (B) physical words and those outside each category evaluated on models fine-tuned on corresponding datasets. Positive values denote lower MSEs in the fine-tuned model.

Carnegie Mellon University

Takeaways

- LMs differ from human language processing in social/emotional intelligence and physical commonsense
- The observed divergences between LMs and human brain activity may stem from LMs' inadequate representation of these specific types of knowledge
- Fine-tuning LMs on tasks related to the two identified topics can align them better with human brain responses

References

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.