Divergences between Language Models and Human Brains

Yuchen Zhou **Emmy Liu** **Graham Neubig** **Michael J. Tarr** **Leila Wehbe**
Carnegie Mellon University
{zhouyuchen,emmy,gneubig,michaeltarr,lwehbe}@cmu.edu

Abstract

Do machines and humans process language in similar ways? Recent research has hinted at the affirmative, showing that human neural activity can be effectively predicted using the internal representations of language models (LMs). Although such results are thought to reflect shared computational principles between LMs and human brains, there are also clear differences in how LMs and humans represent and use language. In this work, we systematically explore the divergences between human and machine language processing by examining the differences between LM representations and human brain responses to language as measured by Magnetoencephalography (MEG) across two datasets in which subjects read and listened to narrative stories. Using an LLM-based data-driven approach, we identify two domains that LMs do not capture well: **social/emotional intelligence** and **physical commonsense**. We validate these findings with human behavioral experiments and hypothesize that the gap is due to insufficient representations of social/emotional and physical knowledge in LMs. Our results show that fine-tuning LMs on these domains can improve their alignment with human brain responses.[1]

1 Introduction

Language models (LMs) now demonstrate proficiency that may equal or even surpass human-level performance on tasks including generating text [Brown et al., 2020a], answering questions [Lewis et al., 2019], translating languages [Costa-jussà et al., 2022], and even tasks that necessitate reasoning and inference [Dasgupta et al., 2022]. This has inspired researchers to leverage LM representations to investigate and model the human brain's language system, positing that LMs may serve as reliable proxies for human linguistic processes [Abdou, 2022]. Prior studies have found that human neural activity, as measured by neuroimaging techniques such as fMRI [Jain and Huth, 2018, Toneva and Wehbe, 2019], EEG [Hale et al., 2018], MEG [Wehbe et al., 2014a], and ECoG [Goldstein et al., 2022], can effectively be predicted using representations from language models such as BERT [Devlin et al., 2018] or GPT-2 [Radford et al., 2019]. Robust neural prediction is hypothesized to stem from the shared computational objective of both LMs and the human brain: predicting subsequent words based on prior context [Yamins and DiCarlo, 2016, Schrimpf et al., 2021].

Despite the evident behavioral similarities, the extent to which LMs and human brains align functionally for language processing remains an open question. Essentially, the methods that LMs and humans use to acquire language are very different. LMs learn statistical regularities across massive sets of linguistic symbols, whereas humans rely on applying structured linguistic principles across relatively little input. Additionally, LMs that are confined to linguistic data are likely to fail to ground linguistic symbols in real-world contexts [Harnad, 1990, Bender and Koller, 2020, Bisk et al., 2020a]. Furthermore, the learning environments and goals of LMs and humans are markedly different. While humans communicate through active inquiry, expressing needs, directed communication, and

[1] Data and code are available at: `https://github.com/FlamingoZh/divergence_MEG`

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

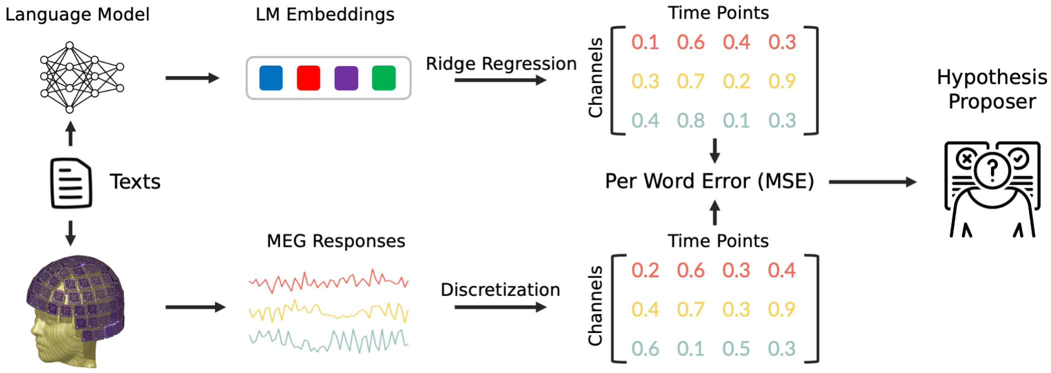


Figure 1: Schematic of our experimental approach. The LM takes as input the current word along with its preceding context to produce the current word's LM embedding. This embedding is then used as input to a ridge regression model to predict the human brain responses associated with the word. The Mean Squared Error (MSE) between the predicted and actual MEG responses is calculated. Finally, an LLM-based hypothesis proposer is employed to formulate natural language hypotheses explaining the divergence between the predicted and actual MEG responses.

scaffolding conversations [Kuhl, 2011], LMs are predominantly trained as passive recipients of raw text data. Consequently, LMs may struggle with comprehending social pragmatics and the nuances of words whose meanings fluctuate across different social contexts [Mahowald et al., 2023].

Previous research exploring the relationship between human and LM language processing has typically focused on the types of linguistic features [Oota et al., 2022a, Sun et al., 2023], neural network architectures [Schrimpf et al., 2021], or training and fine-tuning methods [Sun and Moens, 2023] that may yield better predictions of brain responses. Diverging from this approach, Aw and Toneva [2023] proposed that the divergence between human and LM language processing might stem from LMs' inadequate understanding of texts. They supported this hypothesis by demonstrating that LMs fine-tuned on summarization tasks align more closely with human brain responses. Yet, this hypothesis is only one of many potential explanations. In this work, we adopt a bottom-up, data-driven methodology to systematically investigate the differences between human and machine language processing. Our main contributions are as follows:

1. In contrast to prior studies focusing on the similarities between LMs and human brains, our research emphasizes their differences. We monitor the temporal progression of errors in LM predictions on a word-by-word basis on two datasets with distinct language input modalities (§2).
2. Explaining the prediction errors for every word is challenging due to the vast amount of text. Instead of manually formulating hypotheses, we adopt an LLM-based method that automatically proposes natural language hypotheses to explain the divergent responses between human brains and language models (§3). The top candidate explanations are related to social/emotional intelligence and physical commonsense (§4). We validate these hypotheses with human behavioral experiments.
3. We present evidence that fine-tuning LMs on tasks related to the two identified phenomena can align them more closely with human brain responses. This implies that the observed divergences between LMs and human brains may stem from LMs' inadequate representation of these specific types of knowledge (§5).

2 Predictive MEG Model

2.1 Data Preparation and Preprocessing

While many studies investigating the correlation between brain responses and language models utilize fMRI recordings (e.g., [Caucheteux et al., 2023, Jain et al., 2020]), the comparatively low temporal resolution of fMRI hinders its ability to accurately capture the processing of individual words. To

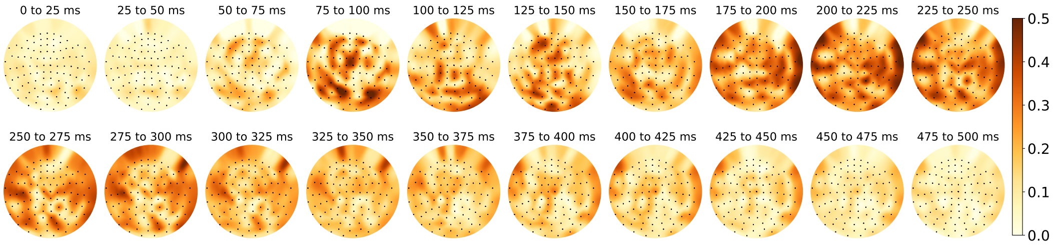


Figure 2: Pearson correlation of actual MEG responses with predicted responses using embeddings from layer 7 of GPT-2 XL on the Harry Potter dataset. The displayed layout is a flattened representation of the helmet-shaped sensor array. Deeper reds indicate more accurate LM predictions. Language regions are well predicted in language processing time windows (refer to §2.4 for more details).

address this limitation, our research employed MEG data. We strategically used two different MEG datasets, each with distinct input modalities, to assess potential variations in the brain's response patterns under these conditions.

The first dataset [Wehbe et al., 2014a] has eight participants reading Chapter 9 of *Harry Potter and the Sorcerer's Stone* (5,176 words) and four participants reading Chapter 10 of the same book (4,475 words) . Each word was exposed on a screen for a fixed duration of 500ms. MEG data were collected on an Elekta NeuroMag MEG with 306 channels at 102 cranial points, and sampled at a rate of 1 kHz. The acquired data underwent preprocessing procedures using the Signal Space Separation (SSS) method [Taulu et al., 2004] and its temporal extension, tSSS [Taulu and Simola, 2006]. The signal was then time-locked with individual words and down-sampled into non-overlapping 25ms time bins. Given the typical low Signal-to-Noise Ratio (SNR) of MEG, we adopted a denoising technique [Ravishankar et al., 2021] that takes advantage of cross-subject correspondences to get an aggregated, denoised version of MEG responses (refer to Appendix A for more details).

To enhance reproducibility and generalizability of our study, we additionally collected MEG data from one participant who listened to six narratives (11,626 words) from The Moth, a platform featuring personal storytelling. These stories were chosen from the stimuli used in a published story listening fMRI dataset [LeBel et al., 2023]. Five of these stories were repeated twice, while one story was repeated five times. The data acquisition was performed using a MEGIN scanner equipped with 306 channels at 102 cranial points. The preprocessing pipeline was similar to that applied to the first dataset. Given that all story repetitions were from the same participant, we averaged the MEG responses for each story's repetitions to enhance SNR without using an alignment method.

2.2 Predicting MEG Responses from LM Embeddings

A substantial number of recent studies exploring the correlation between brain responses and LMs have employed GPT-2 [Pasquiou et al., 2022, Caucheteux et al., 2022, 2023, Toneva et al., 2022]. To ensure consistency and comparability with these studies, we utilized the pre-trained GPT-2 XL model with 1.5B parameters, sourced from HuggingFace's *transformers* library [Wolf et al., 2020a], as the backbone language model. Following previous work [Toneva and Wehbe, 2019], for every word w, we provided the model with a context consisting of the preceding 19 words. We used the output of hidden layers of the LM, subsequently referred to as LM embeddings, to predict the MEG responses associated with each word (Figure 1). For comparison, we also replicated some analyses on Llama-2 7B [Touvron et al., 2023a] (refer to Appendix D for more details).

Building upon established research that demonstrates the capability of LM embeddings to linearly predict MEG responses [Wehbe et al., 2014a, Jain and Huth, 2018, Caucheteux and King, 2022a], we utilized a linear ridge regression model as the encoding model. Considering the time-correlated nature of MEG data, it was essential to maintain the temporal structure when partitioning the data for training and testing purposes [Yang et al., 2019]. Therefore, we implemented a 10-fold cross-validation procedure that splits the MEG data into 10 continuous chunks. We denote the actual MEG responses as M and LM embeddings as L. For split i, we set aside one fold as the test set $(M^{i,test}, L^{i,test})$ and fitted a ridge regression model with weight matrix W^i and bias b^i using the remaining folds, denoted as $(M^{i,train}, L^{i,train})$. The regularization parameters were chosen via

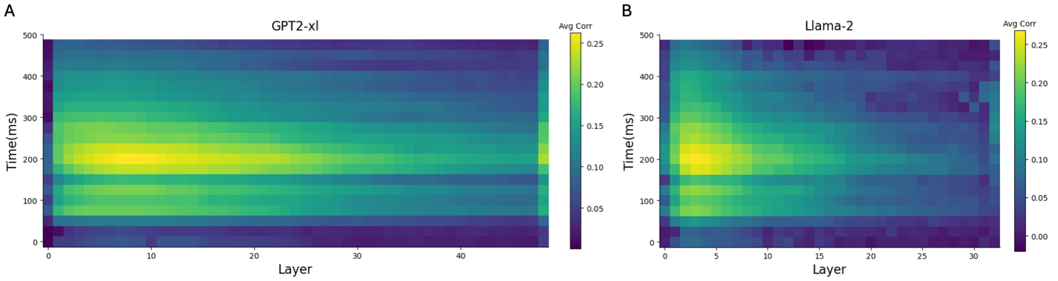


Figure 3: Pearson correlation between actual MEG responses and predicted responses from (A) GPT-2 XL and (B) Llama-2 across LM layers and time after word onset on the Harry Potter dataset. Both models exhibit high correlations in early and intermediate layers at around 200ms. Correlation is computed across words and averaged across MEG channels.

nested cross-validation. Following model training, we applied the trained weight matrix and bias to predict the brain responses from the LM outputs for the test set:

$$\hat{M}^{i,test} = L^{i,test}\hat{W}^{i} + \hat{b}^{i}$$

Finally, the test predictions from all folds were concatenated to form the comprehensive prediction of MEG responses from the LM:

$$\hat{M} = concat[\hat{M}^{i,test}]$$

This process is performed for each of the different time windows relative to word onset.

2.3 Best Language Model Layer for Predicting MEG Responses

Prior research has shown that intermediate layers of language models often best predict human brain responses [Toneva and Wehbe, 2019, Jain and Huth, 2018, Oota et al., 2022b]. Therefore, we selected the layer that best predicts brain responses. Figure 3 illustrates the Pearson correlation between actual MEG responses and those predicted by LM embeddings across layers and time points relative to word onset. We used the average correlation across all words and time windows as the metric to select the best layer. Echoing previous findings, we confirmed that intermediate layers exhibit higher correlations, with layer 7 being the best at predicting brain responses in GPT-2 XL. Similarly, for Llama-2, layer 3 was identified as the most predictive.

2.4 Spatio-temporal Pattern of Predictions

As a sanity check, we examined if the predictive model can effectively predict the brain areas and time course of language processing. These areas include the inferior frontal gyrus, superior temporal gyrus, certain sections of the middle temporal gyrus, and angular gyrus [Blank et al., 2016, Rogalsky et al., 2015, Sahin et al., 2009, Brennan and Pylkkänen, 2012, Friederici, 2002, Visser et al., 2010, Rogalsky and Hickok, 2009].

As shown in Figure 2, we observe a temporal progression of accurately predicted areas after word onset. The prediction performance peaks first in the occipital lobe between 75-100ms. Given that LM embeddings encode information (e.g., word frequency) correlated to the number of letters in a word and MEG is sensitive to abrupt changes in visual inputs, we attribute this early peak to the initial visual perception of a word. This is followed by heightened prediction performance in the bilateral temporal lobe between 175-250ms, when we expect semantic processing to start. This observation aligns with previous research indicating that most language experiments with naturalistic stimuli reveal bilateral language representations [Wehbe et al., 2014b, Huth et al., 2016, Deniz et al., 2019, Toneva et al., 2022]. Finally, between 250-375ms, the anterior temporal lobe and frontal lobe show increased prediction performance, which is likely related to further semantic processing. This sequential pattern of prediction performance replicates the spatio-temporal dynamics of language processing found in previous literature [Wehbe et al., 2014a, Toneva et al., 2022].

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

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

Hypothesis	**Validity**	***p*-value**
contain elements of fiction or exaggeration	0.212	0.012
feature emotional or dramatic language	0.150	0.090
refer to cultural or societal norms	0.138	0.107
include sensory details or imagery	0.137	0.107
have strong emotional or dramatic content	0.100	0.173
show a lack of coherence or logical flow	0.100	0.111
contain elements of surprise and unpredictability	0.094	0.201
contain emotional, personal narratives	0.088	0.201
use idiomatic expressions or figurative language	0.088	0.178
refer to specific events or incidents	0.087	0.237

3 Identifying Phenomena of Interest

Our objective is to investigate the elements of MEG responses that cannot be well explained by the LM. We work with an average of cleaned MEG responses from a group of subjects and multiple trials, which illustrate the common elements of language processing across individuals. Therefore, for words where MEG responses are not well predicted, it is likely that this marks a genuine divergence between human brains and the LM. It is important to clarify that our approach trains an encoding model to predict human brain responses based on language model outputs, rather than the reverse. This means our methodology identifies information that is captured by MEG but is not present in the language model, rather than information captured by the language model but is not present in MEG responses.

Leveraging the high temporal resolution of MEG, we computed the Mean Squared Errors (MSEs) between actual and predicted MEG responses for each individual word on channels that demonstrated statistically significant correlations (one-sided, p=0.001). For word w,

$$MSE(w) = \frac{1}{|S|} \cdot \sum_{i \in S} (\hat{M}(w)_i - M(w)_i)^2 \tag{1}$$

where S is the set of significant channels.

3.1 Automatically Discovering Differences between Brain Responses and LM Predictions

Given the vast amount of text, manual pattern discovery becomes challenging. Figure 7 presents sample sentences color-coded based on prediction error, illustrating the challenges in formulating hypotheses from observations.

To discover subtle differences between MEG responses and LM predictions, we used a method that automatically describes differences between text corpora using proposer and verifier LMs [Zhong et al., 2023]. This system consists of first prompting an LLM (GPT-3; Brown et al. [2020b]) with a

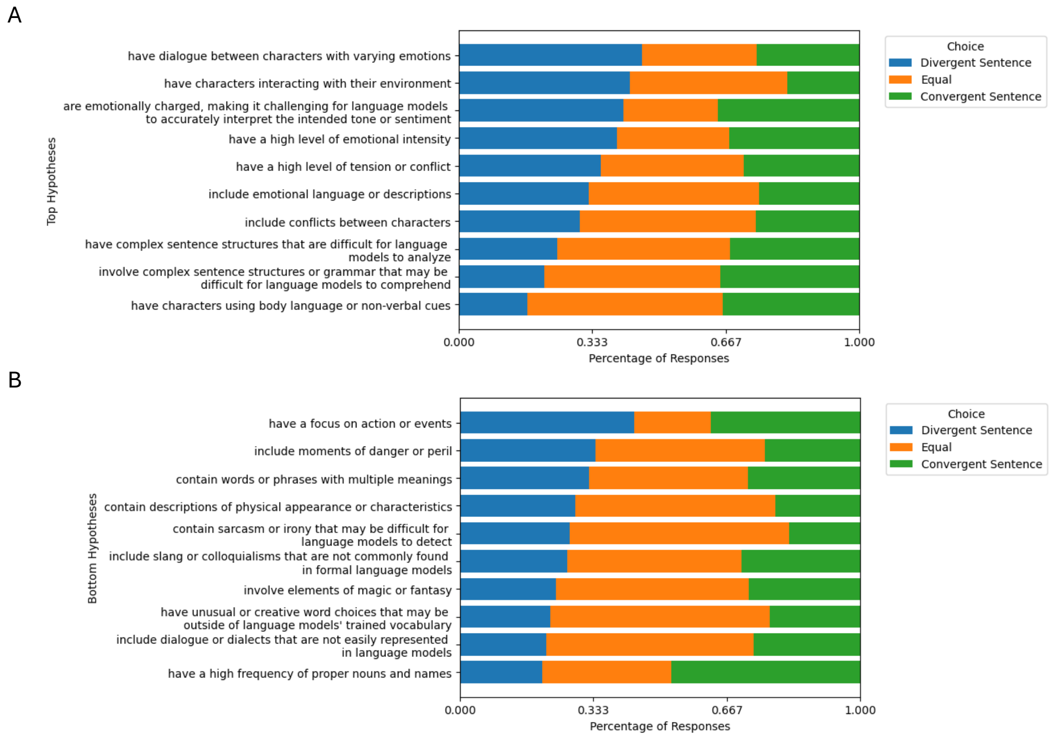


Figure 4: Distribution of human responses for (A) the top 10 and (B) the bottom 10 hypotheses, ranked by the percentage of 'Divergent Sentence' responses.

number of samples from two corpora (D_0, D_1) to generate many hypotheses on how the first corpus differs from the second, and then using a fine-tuned validator model (FLAN-T5-XXL; Chung et al. [2022]) to validate how often each proposed hypothesis is true based on pairs from each corpus sampled from a held-out set. Specifically, the verifier is presented with a prompt containing two sentences from D_0 and D_1, and asked whether or not the hypothesis is true, and this is repeated across the development set for each hypothesis. For the exact prompts used in the proposer and verifier, please refer to Appendix E. Sentences were then ranked based on their mean MSE. The top 100 least-predicted sentences constituted the $D0$ set, while the 100 well-predicted sentences constituted the $D1$ set. This process of hypothesis proposal and verification was repeated across 3 cross-validation folds.

3.2 Proposed Hypotheses

The top ten hypotheses from the Harry Potter dataset ranked by validity[2] are listed in Table 1. We identified two primary differences between the language model and the human brain: firstly, the processing of social and emotional information, and secondly, the capacity for interaction with the surrounding environment. These are henceforth referred to as **social/emotional intelligence** and **physical commonsense**, respectively. Importantly, these hypotheses resonate with conclusions drawn in prior research, as detailed in §4. Similarly, we ran the hypothesis proposer on the Moth dataset. This replication produced slightly varied but fundamentally similar topics to those discovered in the Harry Potter dataset (Table 2). This congruence across datasets with different input modalities aligns with previous research showing that after initial sensory processing, the brain's language processing is consistent across reading and listening [Deniz et al., 2019, Chen et al., 2023].

[2] Validity measures the difference in certainty that the hypothesis is true between the two corpora, see Zhong et al. [2023] for more details.

3.3 Manual Hypothesis Verification

We conducted an experiment involving human participants for additional validation of our hypotheses. We gathered data from 10 participants using Qualtrics, resulting in a collection of 1,400 trials. In each trial, participants were presented with a hypothesis selected either from the top 10 or bottom 10 hypotheses generated from the Harry Potter dataset, along with a pair of sentences — one from $D0$ and the other from $D1$ — in a randomized order. The task for participants was to determine which sentence aligned more closely with the given hypothesis, choosing between "More True for Sentence A", "More True for Sentence B", or "Equally true".

The response distribution for each hypothesis is shown in Figure 4. Note that a given hypothesis is not expected to apply to all divergent sentences (e.g., it might suffice for a sentence to be emotionally intense or grammatically complex to be divergent). If a hypothesis does not align with a sentence from the divergent set, participants should show no preference between the two sentences presented. Chi-square analysis revealed statistically significant differences in the distribution of responses between the top and bottom hypotheses ($p = 0.024$). A preference towards divergent sentences was observed in the top hypotheses condition while a preference towards "Equally True" was observed in the bottom hypothesis condition. This pattern can be attributed to the role of the proposer, which was instructed to generate hypotheses that effectively distinguish D_0 (comprising divergent sentences) from D_1. As a result, the bottom hypotheses tend to be those that fail to differentiate between D_0 and D_1, rather than those that are more explanatory of D_1 compared to D_0 (refer to Appendix F for more details).

4 Selected Phenomena

Comprehending social/emotional and physical commonsense requires humans use a broad spectrum of contextual knowledge. We briefly discuss the insights and challenges highlighted in the existing neuropsychological and NLP literature regarding these domains.

Human social and emotional intelligence requires both introspection and predicting the feelings of others [Salovey and Mayer, 1990]. Neuropsychological research on social cognition has identified a network of brain regions that support understanding other people's intentions, actions, and emotions [Saxe et al., 2006]. Crucially, emotions are intrinsic to the human experience and pervasively interact with other mental facilities, including language [Satpute and Lindquist, 2021]. As such, creating agents with social and emotional intelligence has been a longstanding goal of NLP [Gunning, 2018, Paiva et al., 2021]. However, LLMs still fall short of human abilities for inferring the mental states and emotions of others ("theory-of-mind" tasks) [Sap et al., 2022].

Physical commonsense refers to knowledge about the physical properties of everyday objects and physical phenomena [Forbes et al., 2019, Bisk et al., 2020b]. From a neuropsychological perspective, language is not the primary means through which humans acquire commonsense physical knowledge. Instead, humans rely on sensory inputs and interactions with their environment [Baillargeon, 1994]. Notably, the category of a physical object affects which brain regions are recruited when interacting with that object. For example, interacting with people activates the theory of mind areas [Saxe et al., 2006], the visual face areas [Sergent et al., 1992, Kanwisher et al., 1997], and body areas [Downing et al., 2001], interacting with corridors while navigating recruits the visual place [Epstein and Kanwisher, 1998] and spatial navigation areas, and interacting with tools recruits the dorsal object-processing stream [Almeida et al., 2010]. Interestingly, reading about these domains has also been found to recruit these same visual regions [Wehbe et al., 2014b, Huth et al., 2016]. Given how physical commonsense knowledge is acquired, it is not surprising that, within NLP, this domain poses a challenge for language models. While these models can potentially learn representations capturing specific physical properties of the world, such as an object's color or a game board's state [Abdou et al., 2021, Li et al., 2023], it remains unclear whether purely text-based representations can capture the richness and complexity of physical commonsense as exhibited by humans [Forbes et al., 2019, Bisk et al., 2020b].

5 Improving Brain Alignment via Fine-tuning

We hypothesize that the inability of language models to accurately predict associated brain responses stems from their inadequate representations of social/emotional understanding and physical world

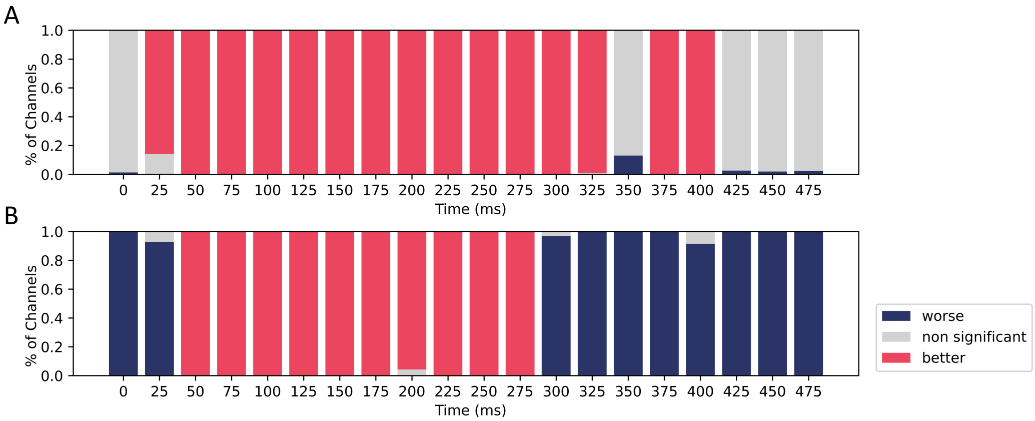


Figure 5: Performance comparison of the base model with models fine-tuned on (A) social and (B) physical datasets. Each panel's y-axis shows the percentage of channels in the fine-tuned model with better, worse, or non-significantly different performance (measured by Pearson correlation) compared to the base model. Fine-tuned models outperform the base model during language processing time windows. Refer to Appendix K for a detailed view of each MEG channel plotted.

knowledge. Drawing inspiration from Aw and Toneva [2023], we fine-tuned the GPT-2 XL model on datasets specific to the two phenomena to examine if targeted fine-tuning could enhance the model's alignment with brain activity.

Furthermore, we examined whether domain-specific fine-tuning would specifically bolster the model's capability in predicting MEG responses associated with words from that domain, as compared to words outside that domain. To this end, we recruited three raters to annotate Chapter 9 of *Harry Potter* across the two domains. We release these annotations as a resource for the dataset to facilitate further analysis. Details on the annotation process can be found in Appendix H. Examples of each phenomenon within the *Harry Potter* text can be found in Appendix I.

5.1 Datasets

Social/Emotional Intelligence We study social and emotional intelligence using the Social IQa dataset [Sap et al., 2019]. This dataset contains questions about people's feelings and their social implications.

Physical Commonsense We study physical commonsense using the PiQA dataset [Bisk et al., 2020b]. This dataset contains goal-driven questions based on everyday situations. These questions were taken from the website instructables.com, where people share DIY project instructions.

We also provide examples from each dataset in Table 3.

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**

5.2 Fine-tuning Setup

In order to keep the architecture of fine-tuned models consistent with the base model, we format the multiple choice task as N language modeling tasks, where N is the number of options. Specifically, for the combined context and question x, we directly concatenate each possible multiple-choice answer $\{y_1, ..., y_N\}$ to x to form N different sentences. After passing the concatenated sequences through the model, we sum the logits of all tokens corresponding to each multiple-choice option to obtain a score proportional to its log-likelihood. These scores are then gathered into a size $(1, N)$ tensor, and cross-entropy loss relative to the correct multiple choice answer is used to train the model. Further details on the fine-tuning setup can be found in Appendix G.

5.3 Comparing Fine-tuned Models with the Base Model

We evaluated the fine-tuned GPT-2 XL model on the Harry Potter dataset. To identify channels with statistically significant differences between the base and fine-tuned model, we calculated empirical *p*-values by comparing the true correlation value with 10,000 simulated ones obtained by permuting the brain data. Details of the algorithm can be found in Appendix J.

Fine-tuned models are better aligned with the brain on both tasks. As illustrated in Figure 5A, the model fine-tuned on the social dataset exceeds the base model in performance across the majority of channels within the 50ms to 300ms time interval post word onset. Notably, this interval corresponds to the language processing time windows, as identified in §2.4. In a similar vein, the fine-tuned physical model exceeds the base model's performance in almost all channels during the 50-275ms interval post word onset (Figure 5B). However, interestingly, almost all channels are worse than the base model outside this time window. This time selectivity may indicate that the improvements of the fine-tuned model are tailored towards linguistic comprehension rather than broader brain functionalities.

Fine-tuning improves alignment more for words annotated with that category. We compared the reduction in prediction error for words annotated within each category and words outside each category by computing the difference in MSE between the model fine-tuned on the corresponding task and the base model. As demonstrated in Figure 6A, prediction errors for social words exhibit a significant reduction compared to non-social words 200-275ms post word onset. Additionally, there is a significant improvement in MSE for physical words over non-physical words 150-225ms post word onset (Figure 6B). We also ran additional control experiments to check if MSE improvement is specific to words that match the category of the dataset on which the model was fine-tuned. Specifically, we evaluated the prediction improvement of physical words on the model fine-tuned on the social dataset, and vice versa (Appendix L).

Improvements are not related to increased language-modeling ability. Prior work has found that LMs with lower perplexity can better predict brain activity [Schrimpf et al., 2021]. Therefore, additional fine-tuning may have improved the language model's ability to perform the LM task in general, leading to improved alignment. To rule out this possibility, we performed 3-fold cross-validation on *Harry Potter and the Sorcerer's Stone*, excluding Chapters 9 and 10, which were used as data in this study. We trained the base model, as well as the fine-tuned models, on the train set in each fold with the language modeling objective, and found that the final average losses on the test sets were similar (See Appendix M for details).

6 Related Work

Numerous studies have found that LM hidden states can linearly map onto human brain responses to speech and text measured by MEG, EEG, and fMRI [Wehbe et al., 2014a, Hale et al., 2018, Jain and Huth, 2018, Abnar et al., 2019, Jat et al., 2019, Gauthier and Levy, 2019, Toneva and Wehbe, 2019, Caucheteux and King, 2022a, Jain et al., 2020, Toneva et al., 2022, Aw and Toneva, 2022].

At a more foundational level, studies have identified shared computational principles between LMs and human brains. Evidence suggests that both human brains and LMs are perpetually engaged in predicting the subsequent word [Schrimpf et al., 2021]. LM surprisal is found to be positively correlated with brain activation, reaching its peak approximately 400 ms post word onset [Goldstein et al., 2022]. This aligns well with N400, which denotes a decline in brain activation upon encountering unexpected words around 400 ms after word onset [Lau et al., 2009, Parviz et al., 2011, Halgren et al.,

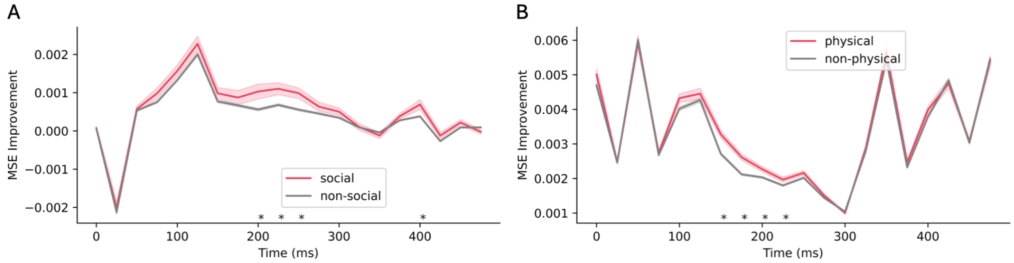


Figure 6: Comparison of improved MSE between (A) social and (B) physical words and those outside each category evaluated on models fine-tuned on corresponding datasets. Positive values denote lower MSEs in the fine-tuned model. Shaded region indicates standard error. Asterisks denote time points with significant differences between the two groups (Student's t-test with FDR correction, *p*=0.05).

2002]. Moreover, LM representations can predict the hierarchy of brain responses [Caucheteux and King, 2022b, Caucheteux et al., 2023]. Despite this, Antonello and Huth [2022] have pointed out that a high correlation between brain activity and LMs does not necessarily imply that they operate under similar computational principles.

We not only observe this LM-brain alignment but can also actively intervene in it. Research has demonstrated that the alignment between LMs and human brains can be improved by task-specific fine-tuning. A notable instance is the study by Schwartz et al. [2019], where the fine-tuning of BERT using both fMRI and MEG signals enhanced its ability to predict fMRI responses. Importantly, this improvement was not participant-specific and could be transferred to hold-out individuals. Another study [Aw and Toneva, 2023] showed that task-oriented fine-tuning, particularly for narrative summarization, also facilitated better alignment with brain activity. Furthermore, altering the architecture of BERT such that it aligns better with the brain improves its performance on downstream NLP tasks [Toneva and Wehbe, 2019]. These findings suggest a potentially symbiotic relationship between enhancing task performance in LMs and boosting their alignment with brain responses.

7 Conclusions, Limitations, and Future Work

We explore a critical question connecting language models with human neural activity: How do LMs differ from human brains in processing language? We employed an LLM-based approach to automatically propose hypotheses explaining the elements of human brain responses that cannot be well explained by language models. Social/emotional intelligence and physical commonsense emerged as the two dominant themes. After fine-tuning GPT-2 XL on datasets related to these themes, we observed an improved alignment between LM predictions and human brain responses.

Limited by the availability of datasets with aligned brain data, our study was conducted on a relatively narrow range of texts. While we observed consistent patterns across two language modalities, it is important to note that both datasets utilized were exclusively narrative stories. This limited scope raises the possibility that additional, undetected divergences exist, potentially obscured by the quantity of text and the methodology employed for hypothesis generation from the sentences. By developing more robust tools for pattern discovery and incorporating a wider array of textual materials, our approach can be adapted to more comprehensively address the question in future studies.

Acknowledgments and Disclosure of Funding

This work is supported by the James S. McDonnell Foundation. We also acknowledge the support of the National Sciences and Engineering Research Council of Canada (NSERC) through a postgraduate fellowship given to EL (PGSD). This work was also partially supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number R01DC020088. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020a.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Mostafa Abdou. Connecting neural response measurements & computational models of language: a non-comprehensive guide. *arXiv preprint arXiv:2203.05300*, 2022.

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

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

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, 2018.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014a.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

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

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198, 2020.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language, 2020a.

Patricia K Kuhl. Early language learning and literacy: Neuroscience implications for education. *Mind, brain, and education*, 5(3):128–142, 2011.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: a cognitive perspective, 2023.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? *arXiv preprint arXiv:2205.01404*, 2022a.

Jingyuan Sun, Xiaohan Zhang, and Marie-Francine Moens. Tuning in to neural encoding: Linking human brain and artificial supervised representations of language. *arXiv preprint arXiv:2310.04460*, 2023.

Jingyuan Sun and Marie-Francine Moens. Fine-tuned vs. prompt-tuned supervised representations: Which better account for brain language representations? *arXiv preprint arXiv:2310.01854*, 2023.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, pages 1–12, 2023.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749, 2020.

Samu Taulu, Matti Kajola, and Juha Simola. Suppression of interference and artifacts by the signal space separation method. *Brain topography*, 16(4):269–275, 2004.

Samu Taulu and Juha Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in medicine and biology*, 51(7):1759, 2006.

Srinivas Ravishankar, Mariya Toneva, and Leila Wehbe. Single-trial meg data can be denoised through cross-subject predictive modeling. *Frontiers in Computational Neuroscience*, 15:737324, 2021.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. Neural language models are not born equal to fit brain data, but training helps. *arXiv preprint arXiv:2207.03380*, 2022.

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

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023a.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):1–10, 2022a.

Yuxiao Yang, Omid G Sani, Edward F Chang, and Maryam M Shanechi. Dynamic network modeling and dimensionality reduction for human ecog activity. *Journal of neural engineering*, 16(5): 056014, 2019.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *arXiv preprint arXiv:2212.08094*, 2022b.

Idan Blank, Zuzanna Balewski, Kyle Mahowald, and Evelina Fedorenko. Syntactic processing is distributed across the language system. *Neuroimage*, 127:307–323, 2016.

Corianne Rogalsky, Diogo Almeida, Jon Sprouse, and Gregory Hickok. Sentence processing selectivity in broca's area: evident for structure but not syntactic movement. *Language, cognition and neuroscience*, 30(10):1326–1338, 2015.

Ned T Sahin, Steven Pinker, Sydney S Cash, Donald Schomer, and Eric Halgren. Sequential processing of lexical, grammatical, and phonological information within broca's area. *Science*, 326(5951):445–449, 2009.

Jonathan Brennan and Liina Pylkkänen. The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage*, 60(2):1139–1148, 2012.

Angela D Friederici. Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2):78–84, 2002.

Maya Visser, Elizabeth Jefferies, and MA Lambon Ralph. Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of cognitive neuroscience*, 22(6):1083–1094, 2010.

Corianne Rogalsky and Gregory Hickok. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, 19(4): 786–796, 2009.

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

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, Jack L Gallant, Wendy a De Heer, Thomas L Griffiths, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. doi: 10.1038/nature17637. Natural.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

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.

Catherine Chen, Tom Dupré la Tour, Jack Gallant, Dan Klein, and Fatma Deniz. The cortical representation of language timescales is shared between reading and listening. *bioRxiv*, pages 2023–01, 2023.

Peter Salovey and John D. Mayer. Emotional intelligence. *Imagination, Cognition and Personality*, 9 (3):185–211, 1990.

R. Saxe et al. Uniquely human social cognition. *Current opinion in neurobiology*, 16(2):235–239, 2006.

Ajay B. Satpute and Kristen A. Lindquist. At the Neural Intersection Between Language and Emotion. *Affective Science*, 2(2):207–220, 2021.

David Gunning. Machine common sense concept paper, 2018.

Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patrícia Arriaga. *Empathy and Prosociality in Social Agents*, page 385–432. Association for Computing Machinery, New York, NY, USA, 1 edition, 2021.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense?, 2019.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020b.

Renée Baillargeon. How do infants learn about the physical world? *Current Directions in Psychological Science*, 3(5):133–140, 1994.

J Sergent, S Ohta, and B MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992.

N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

R Epstein and N Kanwisher. A cortical representation of the local visual environment. *Nature*, 392 (6676):598–601, 1998.

Jorge Almeida, Bradford Z Mahon, and Alfonso Caramazza. The role of the dorsal visual processing stream in tool identification. *Psychological science*, 21(6):772–778, 2010.

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color, 2021.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2023.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. *arXiv preprint arXiv:1906.01539*, 2019.

Sharmistha Jat, Hao Tang, Partha Talukdar, and Tom Mitchell. Relating simple sentence representations in deep neural networks and the brain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5137–5154, 2019.

Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China, November 2019. Association for Computational Linguistics.

Khai Loong Aw and Mariya Toneva. Training language models for deeper understanding improves brain alignment. *arXiv preprint arXiv:2212.10898*, 2022.

Ellen Lau, Diogo Almeida, Paul C Hines, and David Poeppel. A lexical basis for n400 context effects: Evidence from meg. *Brain and language*, 111(3):161–172, 2009.

Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the australasian language technology association workshop 2011*, pages 38–46, 2011.

Eric Halgren, Rupali P Dhond, Natalie Christensen, Cyma Van Petten, Ksenija Marinkovic, Jeffrey D Lewine, and Anders M Dale. N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage*, 17(3):1101–1116, 2002.

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

Richard Antonello and Alexander Huth. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, pages 1–16, 2022.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems*, pages 14123–14133, 2019.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. Flute: Figurative language understanding through textual explanations, 2022.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States, July 2022. Association for Computational Linguistics.

Verna Dankers, Christopher G Lucas, and Ivan Titov. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. Technical report.

Emmy Liu and Graham Neubig. Are representations built from the ground up? an empirical examination of local composition in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9053–9073, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020b.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

A MEG Denoising

Because of the typical low Signal-to-Noise Ratio (SNR) associated with MEG, we adopted a denoising technique [Ravishankar et al., 2021] that takes advantage of cross-subject correspondences to get an aggregated, denoised version of MEG responses. Specifically, this process involves modeling the MEG responses M_t of subject t as a linear function of the MEG responses M_s from a source subject s:

$$\hat{M}_{t\leftarrow s} = \hat{W}_{t\leftarrow s} M_s + \hat{b}_{t\leftarrow s}$$

We estimated the target subject's MEG responses from all other subjects:

$$\hat{M}_t = \frac{1}{N-1} \sum_{s\in S, s\neq t} \hat{M}_{t\leftarrow s}$$

where S is the set of subjects and N is the number of subjects. These individual estimates are then aggregated to generate a denoised version of MEG responses:

$$\hat{M} = \frac{1}{N} \sum_{s\in S} \hat{M}_t$$

B Sample Sentences

Given the vast amount of text in the datasets, manually discovering patterns becomes challenging. Figure 7 provides an illustrative example by presenting a set of sample sentences that are color-coded based on the magnitude of their prediction error. This visualization demonstrates the complexity involved in formulating hypotheses from observations.

1. He had been looking forward to learning to fly more than anything else.
2. "Of course he has," said Ron, wheeling around.
3. But Neville, nervous and jumpy and frightened of being left on the ground, pushed off hard before the whistle had touched Madam Hooch's lips.

most divergent 🟥 🟫 🟪 ◼ ⬛ least divergent

Figure 7: Sample sentences from the Harry Potter dataset, with colors indicating prediction error levels. Each of the five colors corresponds to a 20-percentile range of words from the entire dataset.

C Additional Results on GPT-2 XL

C.1 Results on Last Layer

In addition to the best layer, we also performed analyses on the last layer of the language model.

C.1.1 Spatial-Temporal Patterns of Predictions

The spatial-temporal pattern of predictions observed in the last layer (Figure 8) is similar to that of the best layer, However, there is a notable difference in the magnitude of the values. Specifically, the maximum correlation in the last layer is lower, decreasing from 0.53 in the best layer to 0.43.

C.1.2 Proposed Hypotheses

We also generate hypotheses from the predictions of the last layer (Table 4). These hypotheses exhibit similarities with those derived from the best-performing layer, notably in their inclusion of emotions

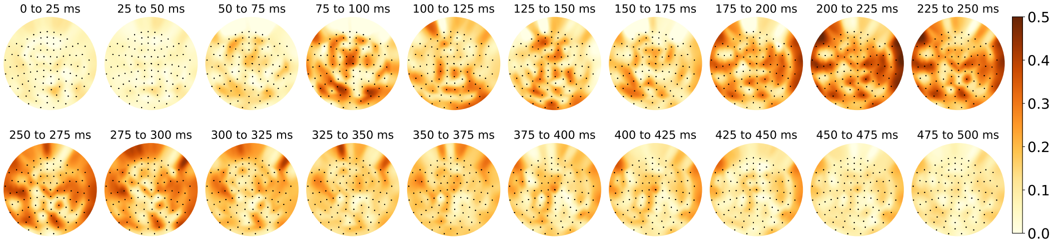


Figure 8: Pearson correlation of actual MEG responses with those predicted by LM embedding from the last layer of GPT-2 XL (evaluated on the test set). The displayed layout is a flattened representation of the helmet-shaped sensor array. Deeper reds indicate more accurate LM predictions. Language regions are effectively predicted in language processing time windows (refer to §2.4 for more details).

and social interactions. However, a distinctive aspect of these hypotheses is their association with supernatural and magical elements. Additionally, we observe the emergence of figurative language, aligning with previous research that indicates language models underperform humans in both the interpretation and generation of figurative language [Chakrabarty et al., 2022, Liu et al., 2022] and the correct representation of idiomatic phrases [Dankers et al., Liu and Neubig, 2022].

Table 4: Top 10 hypotheses generated from the last layer of GPT-2 XL on the Harry Potter dataset

Hypothesis	**Validity**	***p*-value**
contain descriptions of unusual settings or creatures	0.1750	0.0754
has a lot of dialogue, with characters speaking to each other	0.1719	0.0855
contain figurative language	0.1367	0.1409
contain rhetorical questions or exclamations	0.1125	0.1685
contains references to obscure facts or trivia, such as the longest game of Quidditch	0.1094	0.0627
mentions the unknown or unexpected, such as an unknown creature or a surprise announcement	0.1094	0.1856
contain references to the emotions of characters	0.1062	0.1996
contain references to the supernatural	0.0875	0.1827
mentions dangerous creatures and events, such as trolls and duels	0.0813	0.2383
contain references to magic	0.0688	0.2550

D Replication on Llama-2 7B

Although GPT-2 is a widely used language model in brain research, it's not the latest model in the field. Models with more parameters and advanced training methods could show different results. Therefore, we replicated some analyses on Llama-2 [Touvron et al., 2023b]. We used the implementation in the HuggingFace library [Wolf et al., 2020b] with 7B parameters. As Figure 3B shows, early layers exhibit high correlations, with layer 3 identified as having the highest correlation.

D.1 Spatio-Temporal Patterns of Predictions

The spatial-temporal pattern of predictions in layer 3 of Llama-2 (Figure 9) closely resembles those found in GPT-2 XL. This similarity implies that both language models effectively capture the representations of words.

D.2 Proposed Hypotheses

Hypotheses from the predictions of layer 3 of Llama-2 7B can be found in Table 5. Interestingly, the focus of these hypotheses is primarily on physical objects and events. In comparison to the hypotheses produced by GPT-2 XL, there is a notable absence of social and emotional aspects, suggesting that Llama-2 7B could have a more advanced comprehension of social and emotional contexts.

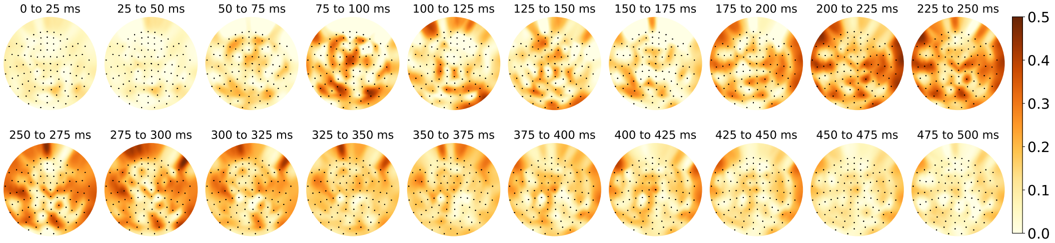


Figure 9: Pearson correlation of actual MEG responses with those predicted by LM embedding from the best layer (layer 3) of Llama-2 (evaluated on the test set). The displayed layout is a flattened representation of the helmet-shaped sensor array. Deeper reds indicate more accurate LM predictions. Language regions are effectively predicted in language processing time windows (refer to §2.4 for more details).

Table 5: Top 10 hypotheses generated from the best layer of Llama-2 on the Harry Potter dataset

Hypothesis	**Validity**	***p*-value**
involve action or movement, such as running or tiptoeing	0.300	0.005
refer to specific events or actions, such as a flying lesson or a spell not working	0.237	0.029
refer to specific objects or locations, such as the front steps or the trophy room	0.237	0.013
describe physical actions or movements	0.175	0.081
discuss or describe dangerous or frightening situations	0.150	0.056
include actions or physical movements	0.150	0.117
contain words or phrases that are specific to the wizarding world	0.140	0.130
have a sense of chaos or disorder	0.137	0.040
have a high level of tension or suspense	0.137	0.128
include words or phrases that are specific to the wizarding world	0.127	0.152

E Proposer and Verifier Prompts

The prompt for the proposer is:

> {A_block}
>
> {B_block}
>
> The dataset includes two chapters from "Harry Potter and the Sorcerer's Stone". The two groups are generated based on the difference between language model and human responses to these sentences. The Group A snippets sentences where language models and humans show divergent responses, while the Group B snippets sentences where language models and humans show similar responses.
>
> I am a literary analyst investigating the characteristics of words. My goal is to figure out which sentences induce different responses for language models and human responses.
>
> Please write a list of hypotheses about the datapoints from Group A (listed by bullet points "-"). Each hypothesis should be formatted as a sentence fragment. Here are three examples.
>
> - "{example_hypothesis_1}"
> - "{example_hypothesis_2}"
> - "{example_hypothesis_3}"
>
> Based on the two sentence groups (A and B) from the above, more sentences in Group A ...

The prompt for the validator is:

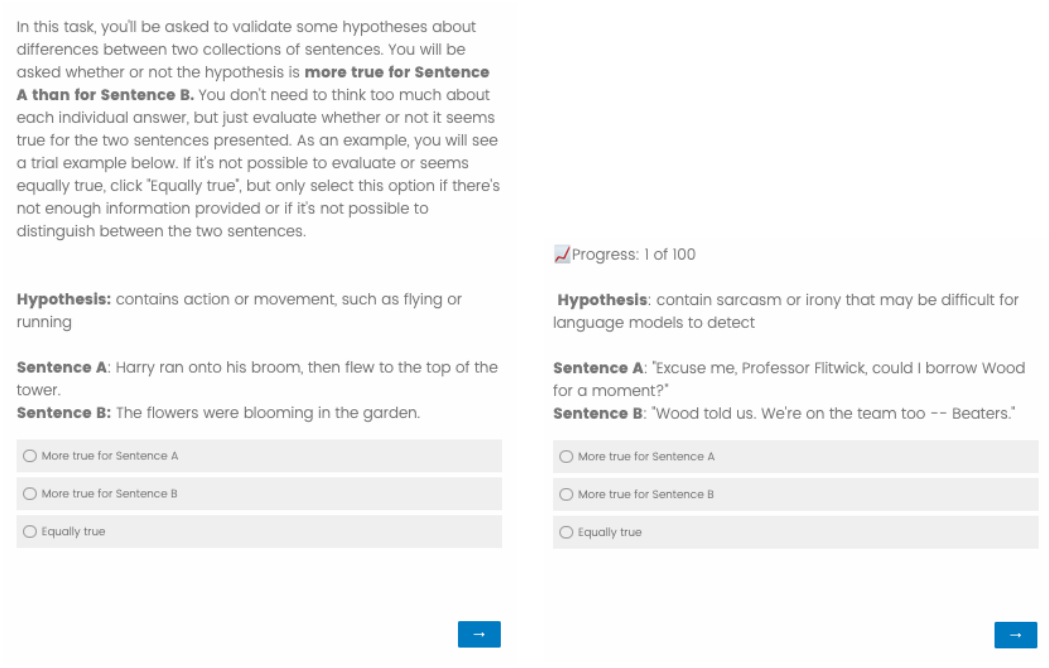
> Check whether the TEXT satisfies a PROPERTY. Respond with Yes or No. When uncertain, output No.
>
> Now complete the following example -
>
> input: PROPERTY: {hypothesis}

TEXT: {text}
output:

F Manual Hypothesis Verification

F.1 Experiment Setup

We recruited 10 participants through Qualtrics. Of these, 9 participants completed 100 trials each, while one participant completed 500 trials. In each trial, participants were presented with a hypothesis selected either from the top 10 or bottom 10 hypotheses generated from the Harry Potter dataset, along with a pair of sentences — one from $D0$ (the divergent sentence set) and the other from $D1$ (the convergent sentence set) — in a randomized order. The task for participants was to determine which sentence aligned more closely with the given hypothesis, choosing between "More True for Sentence A", "More True for Sentence B", or "Equally true". Note that a given hypothesis is not expected to apply to all divergent sentences (e.g., it might suffice for a sentence to be emotionally intense or grammatically complex to be divergent) so it is expected that some of the responses will be "Equally true". A screenshot of the experiment can be found in Figure 10.



In this task, you'll be asked to validate some hypotheses about differences between two collections of sentences. You will be asked whether or not the hypothesis is **more true for Sentence A than for Sentence B.** You don't need to think too much about each individual answer, but just evaluate whether or not it seems true for the two sentences presented. As an example, you will see a trial example below. If it's not possible to evaluate or seems equally true, click "Equally true", but only select this option if there's not enough information provided or if it's not possible to distinguish between the two sentences.

Hypothesis: contains action or movement, such as flying or running

Sentence A: Harry ran onto his broom, then flew to the top of the tower.
Sentence B: The flowers were blooming in the garden.

○ More true for Sentence A

○ More true for Sentence B

○ Equally true

→

📈Progress: 1 of 100

Hypothesis: contain sarcasm or irony that may be difficult for language models to detect

Sentence A: "Excuse me, Professor Flitwick, could I borrow Wood for a moment?"
Sentence B: "Wood told us. We're on the team too -- Beaters."

○ More true for Sentence A

○ More true for Sentence B

○ Equally true

→

Figure 10: Screenshots of the experiment.

F.2 Results

We constructed a contingency table with dimensions (Top Hypothesis, Bottom Hypothesis) by (Prefer Divergent, Equal, Prefer Convergent) (Table 6). A binomial test conducted on the contingency table (looking only at the "Prefer Divergent" and "Prefer Convergent" responses) showed that divergent sentences were more likely to be chosen over convergent sentences on average ($p < 10^{-10}$). A Chi-square test revealed statistically significant differences in the distribution of responses between the top and bottom hypotheses ($p = 0.024$). Additionally, we utilized the Chi-square test to compare the frequency of "Prefer Divergent," "Equal," and "Prefer Convergent" responses in two conditions. Notably, a preference towards divergent sentences was observed in the top hypotheses condition compared to the bottom hypotheses ($p = 0.093$). In contrast, in the bottom hypothesis condition, there was a marked preference for "equally true" ($p = 0.008$). No significant difference was observed in the preference for convergent sentences between the conditions ($p = 0.676$).

Table 6: Contingency Table for Human Responses

	Divergent	**Equal**	**Convergent**
Top	377	122	209
Bottom	336	159	196

G Fine-tuning details

G.1 Computational Details

GPT-2 XL was trained separately on each of the two datasets in subsection 5.1 on 4 A6000 GPUs with 16-bit quantization and a batch size of 1 per GPU. Deepspeed with ZeRo stage 2 optimization was used in order to parallelize training [Rasley et al., 2020]. The Adam optimizer was used with a learning rate of 1e-5, betas of $(0.9, 0.999)$, epsilon of 1e-8, and no weight decay. Models were trained with early stopping with a patience of 3 [Kingma and Ba, 2017].

G.2 Multiple-choice training

Let x_i represent the concatenation of the context, if applicable, and the question. Then for each answer choice y_i, we concatenate it with the question and context, and feed it to the model to obtain a sequence of logits.

$$\ell_i = \text{Model}(x_i \oplus y_i) \tag{2}$$

Then we sum the logits corresponding to the sequence, where $t \in [1, T]$ represents the total length of $x_i \oplus y_i$.

$$\text{score}_i = \sum_{t=1}^{T} \ell_{i,t} \tag{3}$$

Finally, we take the cross-entropy loss of these values relative to a one-hot encoding of the correct option, where $t_i = 1$ if option i is correct, or else 0.

$$P_i = \frac{\exp(\text{logit}_i)}{\sum_{j=1}^{N} \exp(\text{logit}_j)}$$
$$L = -\sum_{i=1}^{N} t_i \log(P_i)$$

G.2.1 Performance on Multiple-Choice Datasets

We note that the performance of the final model may not approach that of GPT-2 XL fine-tuned with an output size of N denoting each option, as we keep the output dimension the same as the size of the vocabulary. However, we report the final accuracy achieved by each model on the original datasets here.

Table 7: Summary of model performance

Dataset	Best epoch	Accuracy (%)	Baseline (random) accuracy
Social IQa	4	54.86%	33.33%
PiQA	1	73.88%	50.00%

H Annotations

To decide which category a word belongs to, we employed three raters who used binary coding to indicate if a word belonged to the target category. The consistency among raters was evaluated using Krippendorff's alpha. Their consistency was 0.54 for social/emotion and 0.87 for physical. Finally, if at least two out of the three people annotated a word as fitting a category, we counted it as belonging to that category.

H.1 Annotation Guidelines

H.1.1 Social/Emotional Intelligence

- Include words that depict the emotions of characters and/or social interactions.
- Exclude words that suggest emotions or social interactions indirectly. For instance, "slam the door" shouldn't be annotated.

H.1.2 Physical commonsense

- Annotate words referring to tangible entities, such as characters (people) and physical objects.
- Do not annotate words that represent concrete ideas but lack physical substance, like "laughter".
- Pronouns should also be excluded.

I Examples of phenomena in Harry Potter

We give some examples of the two phenomena in the dataset according to the annotations. Words of that category are marked in bold.

I.1 Social/Emotional

- Harry had never believed he would meet a boy he **hated** more than Dudley.
- Hermione Granger was almost as **nervous** about flying as Neville was.
- But Neville, **nervous** and **jumpy** and **frightened** of being left on the ground, pushed off hard before the whistle had touched Madam Hooch's lips.

I.2 Physical Commonsense

- Up the **front steps**, up the **marble staircase** inside, and still **Professor McGonagall** didn't say a word to him.
- **Ron** had a piece of **steak** and **kidney pie** halfway to his **mouth**, but he'd forgotten all about it.
- They pulled on their **bathrobes**, picked up their **wands**, and crept across the **tower room**, down the **spiral staircase**, and into the **Gryffindor common room**.

J Algorithm for Permutation Test

To identify channels on which the performance of the fine-tuned model and the base model has statistically significant differences, we calculated empirical p-values by comparing the true correlation value with 10,000 simulated ones obtained by permuting the brain data as shown in Algorithm 1. Given that we are assessing multiple hypotheses simultaneously, we also used the Benjamini-Hochberg False Discovery Rate (FDR) [Benjamini and Hochberg, 1995] to correct for multiple comparisons, at level $\alpha = 0.05$.

Algorithm 1 Permutation test (for one channel, one time window)

```
**Input:** Brain data $D$, Prediction from base model $P_1$, Prediction from fine-tuned model $P_2$
$D$, $P_1$, and $P_2$ are all of size $(1, N)$, where $N$ is the number of words in the dataset.
**Output:** $pvalue$
$X = \text{corr}(D, P1) - \text{corr}(D, P2)$
$Counter = 0$
**for** $i = 1$ **to** $10,000$ **do**
    $D_i = \text{permute}(D)$
    $X_i = \text{corr}(Di, P1) - \text{corr}(Di, P2)$
    **if** $X_i > X$ **then**
        $Counter = Counter + 1$
    **end if**
**end for**
$pvalue = \frac{Counter+1}{10{,}000+1}$
```

K Comparison between Fine-tuned models and the Base Model

We provide a detailed view of the comparison between the base language model and the models fine-tuned on social (Figure 11) and physical (Figure 12) datasets with each channel plotted.

L Additional Control Experiments on MSE Improvement

We conducted additional control experiments to evaluate whether MSE improvement is specific to words that match the category of the dataset on which the model was fine-tuned. Specifically, we evaluated the improvement of physical words on the model fine-tuned on the social dataset, and vice versa.

This analysis reveals that the performance of the model fine-tuned on the social dataset does not significantly differ when assessed with physical and non-physical words (Figure 13B). This finding implies that the enhancements observed are specifically tied to social words. On the other hand, in the model fine-tuned on the physical dataset, we observed a marginal, though not statistically significant, boost in performance with social words (Figure 13C). We propose that this marginal improvement could be attributed to the presence of social and emotional knowledge embedded within the physical dataset. To substantiate this hypothesis, we conducted a thorough review of the physical dataset and identified items that indeed pertain to social or emotional scenarios.

Examples of such items include:

- how do you give a surprise party?
- To help your child feel less afraid when they're going to sleep
- How can you get a child to smile in a photo?
- To help a friend feel better when they are sad
- how to avoid danger?
- To determine if someone has romantic feelings for you

These findings led us to conclude that the marginal improvement in processing social words by the model fine-tuned on the physical dataset may result from exposure to social and emotional content.

M Cross-validation on language modelling task

We perform 3-fold cross-validation on the remaining chapters of the *Harry Potter* book (excluding chapters 9 and 10), where we randomly shuffle paragraphs and assign to train:validation:test sets respectively 77%, 16.5%, and 16.5% of the data. Paragraphs that exceeded the context length were excluded. Both the base GPT-2 XL model as well as each model fine-tuned on the three domains were trained to predict the next word for 3 epochs, with the same hyperparameters used in Appendix G.

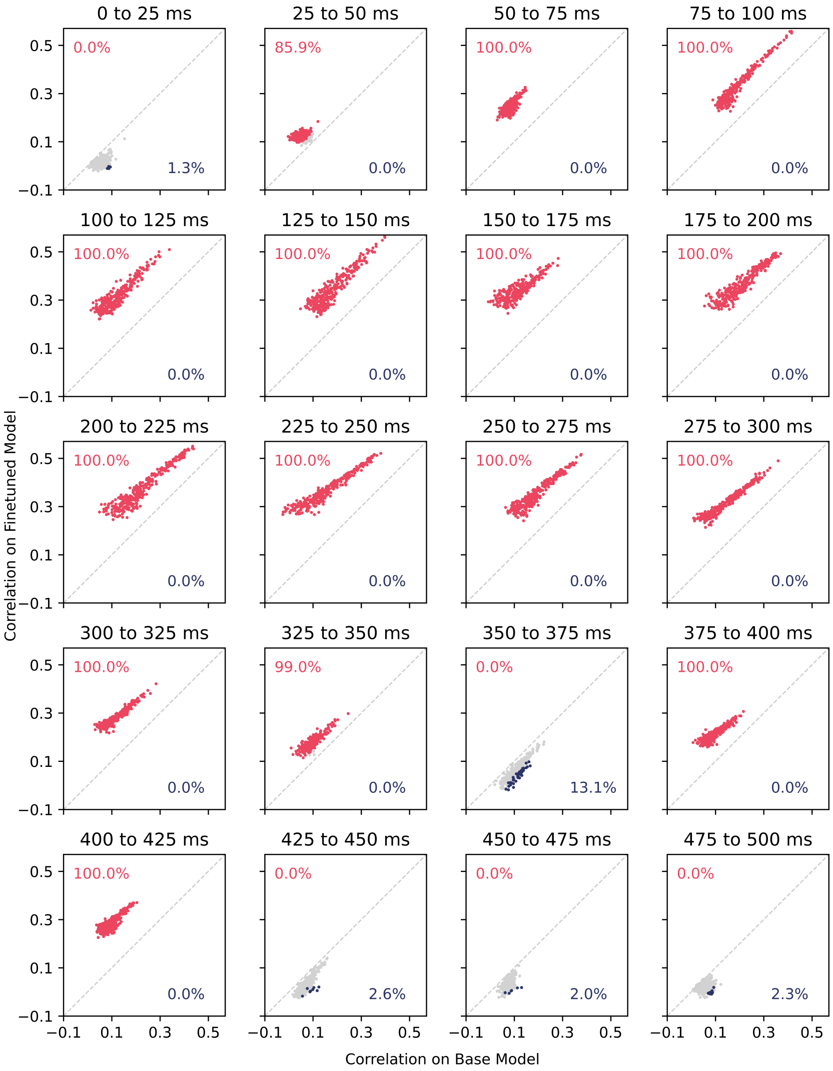


Figure 11: Performance evaluation of the model fine-tuned on the Social IQa (social and emotional) dataset versus the base model using Pearson correlation. Each dot represents a MEG channel. Red channels indicate better predictions by the fine-tuned model, blue channels indicate better predictions by the base model, and gray dots denote non-significant differences. The fine-tuned model outperforms the base model in predicting most channels during language processing time windows.

Results on the test set for each fold are listed below. The average negative-log-likelihood loss per token at the end of training is reported in Table 8.

Table 8: Summary of language-modeling loss across cross-validation folds for models on the remaining chapters of *Harry Potter*.

Model	Avg. Loss (%) ± St.dev	Fold 1 Loss	Fold 2 Loss	Fold 3 Loss
Base	0.08795 ± 0.01707	0.09794	0.06391	0.1020
Social	0.1148 ± 0.00286	0.1119	0.1187	0.1138
Physical	0.1001 ± 0.00184	0.1019	0.1009	0.0976

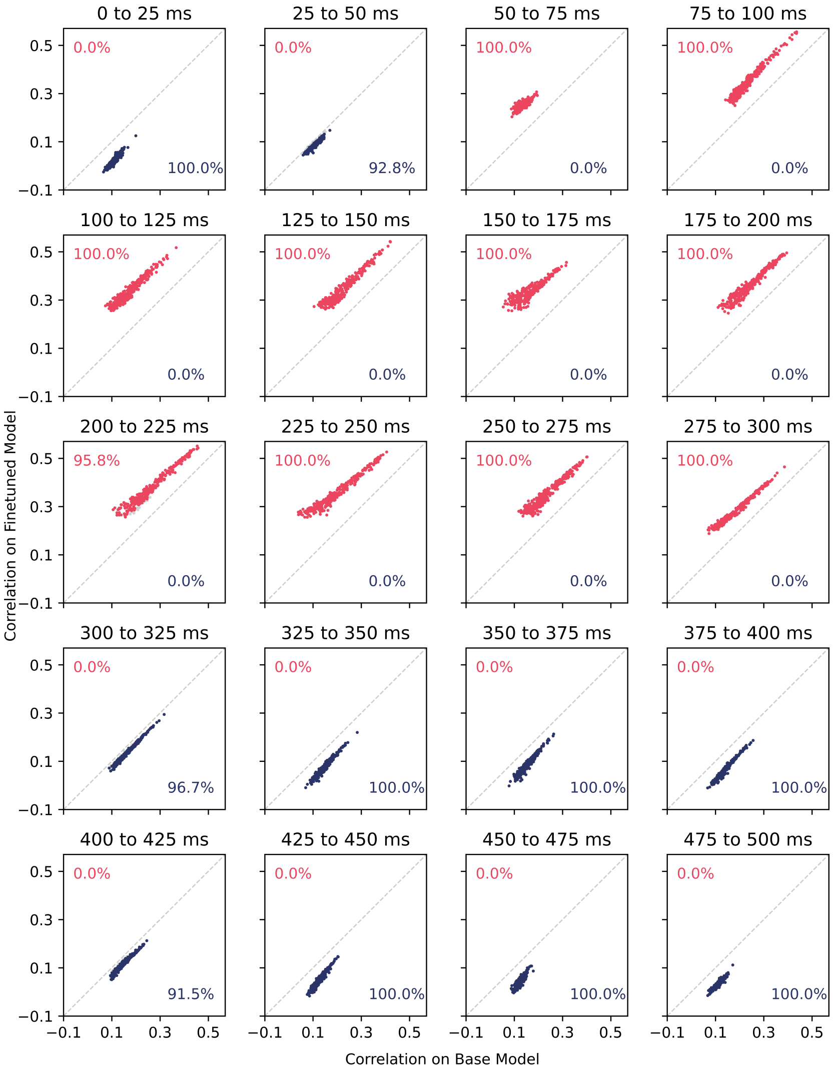


Figure 12: Performance evaluation of the model fine-tuned on the PiQA (physical) dataset versus the base model using Pearson correlation. Each dot represents a MEG channel. Red channels indicate better predictions by the fine-tuned model, blue channels indicate better predictions by the base model, and gray dots denote non-significant differences. The fine-tuned model outperforms the base model in predicting most channels during language processing time windows.

N Societal Impacts

This study represents a significant intersection between Neuroscience and Machine Learning, striving to push the boundaries of machine learning models while deepening our understanding of how the human brain functions. In a broader context, this research lays the groundwork for future breakthroughs in the field of neuroscience and for making human-computer interfaces more efficient and intuitive.

However, the development of human-computer interfaces may cause problems in privacy and security, as these interfaces often require the collection and processing of personal data, increasing the risk of data breaches and unauthorized access. There are also ethical concerns regarding the potential for surveillance and the impact on employment, as advanced interfaces might automate tasks currently performed by humans, potentially displacing workers.

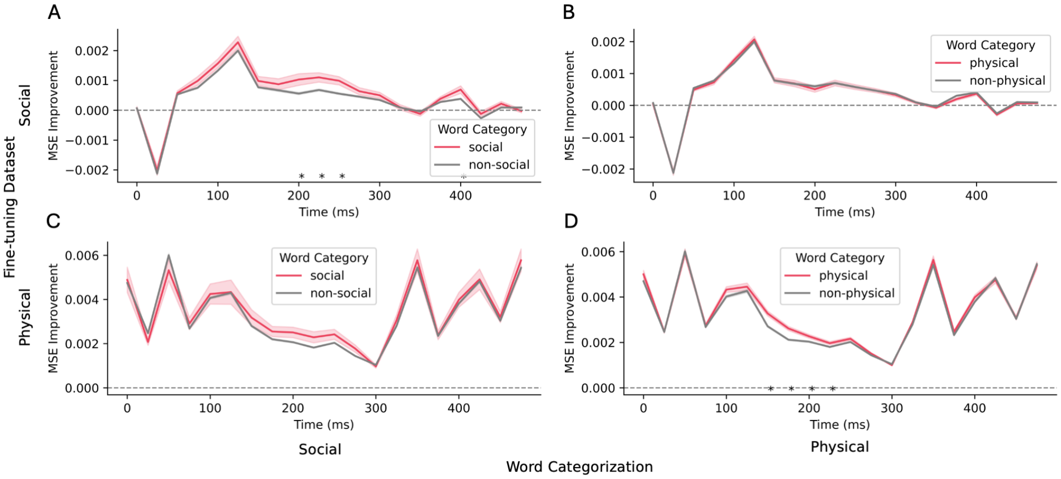


Figure 13: Comparison of improved MSE for A) social vs. non-social words on the social model, B) physical vs. non-physical words on the social model, C) social vs. non-social words on the physical model, and D) physical vs. non-physical words on the physical model. Positive values denote lower MSEs in the fine-tuned model. Shaded region indicates standard error. Asterisks denote time points with significant differences between the two groups (Student's t-test with FDR correction, *p*=0.05).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The abstract provides a concise summary of the key contributions, while the introduction elaborates on the assumptions and contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the work are discussed in section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides a comprehensive disclosure of all the necessary information required to reproduce the main experimental results. The methodologies are described in sufficient detail, allowing for accurate replication of the experiments. Additionally, the code implementation is accessible via an online repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code implementation for all experiments is accessible via an online repository. The Harry Potter dataset is open-sourced. The Moth dataset, which is a component of a larger research initiative, will be publicly released once additional data collection is complete. Instructions on command and environment to reproduce experimental results can be found in the online repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details necessary to understand the results. The training of the encoding model is detailed in subsection 2.2, and the fine-tuning details of the language models are specified in Appendix G. Additionally, the code for training these models is available in the online repository provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper appropriately reports error bars and p-values, providing clear indications of the statistical significance of the experiments. Additionally, p-values are corrected for multiple comparisons.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides detailed information on the computational resources required to reproduce the experiments, including the type and number of GPUs used, as well as the specific hyperparameters employed for fine-tuning the language model.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research fully conforms to the NeurIPS Code of Ethics. It ensures fair compensation and adherence to protocols for human research participants. Data privacy and consent are prioritized, with fair use of datasets. The paper transparently addresses potential societal impacts, including safety, security, discrimination, and environmental concerns, providing mitigation strategies. It avoids facilitating illegal activities and ensures fairness and human rights protection. Documentation of data and models is thorough, with necessary licenses and privacy protocols. Responsible release of models and accessibility of research artifacts are ensured, with all essential elements for reproducibility disclosed.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The societal impacts of the work are discussed in detail in Appendix N, covering both the potential positive and negative effects on society.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applicable. The paper does not involve data or models that carry a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The assets used in the paper are properly cited, with the creators or original owners credited. The versions we used are explicitly stated, and the license and terms of use are respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, the new assets introduced in the paper are well documented. Detailed instructions on how to use the code and data are provided in the online repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes the full text of instructions given to participants and a screenshot, which can be found in Appendix F.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The experiments conducted in the study posed no risk to participants. Additionally, all experiments were thoroughly reviewed and approved by the Institutional Review Board (IRB).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.