Decoding Semantic Representations in the Brain Under Language Stimuli with Large Language Models

Anna Sato[1], **Ichiro Kobayashi**[1]

[1]Ochanomizu University, Tokyo, Japan
{g1920519,koba}@is.ocha.ac.jp

Abstract

Brain decoding technology is paving the way for breakthroughs in the interpretation of neural activity to recreate thoughts, emotions, and movements. Tang et al. (2023) introduced a novel approach that uses language models as generative models for brain decoding based on functional magnetic resonance imaging (fMRI) data. Building on their work, this study explored the use of three additional language models along with the GPT model used in previous research to improve decoding accuracy. Furthermore, we added an evaluation metric using an embedding model, providing higher-level semantic similarity than the BERTScore. By comparing the decoding performance and identifying the factors contributing to good performance, we found that high decoding accuracy does not solely depend on the ability to accurately predict brain activity. Instead, the type of text (e.g., web text, blogs, news articles, and books) that the model tends to generate plays a more significant role in achieving more precise sentence reconstruction.

1 Introduction

Brain decoding technology has recently gained considerable attention for its potential. This technology, which analyzes brain activity in real time to decode thoughts, emotions, and movements, is expected to bring major breakthroughs in areas such as medicine, rehabilitation, communication support, scientific research, and beyond. Many brain-machine interfaces (BMIs) designed for practical use rely on invasive methods like electrocorticography (ECoG), which require brain surgery (Willett et al., 2023; Metzger et al., 2022). Although these methods provide clearer data, allowing for accurate analysis of brain activity even in complex tasks, they come with surgical risks and practical limitations, making them unsuitable for large-scale deployment.

In contrast, non-invasive BMIs using functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) are safer and more cost-effective alternatives. However, these methods face challenges, including noisy data and lower temporal or spatial resolution, which restrict their applications to simpler tasks such as recognizing a limited set of words or basic motion commands (Lopez-Bernal et al., 2022). Non-invasive BMI technologies remain far from being practically deployed, with several challenges yet to be addressed.

Tang et al. (2023) took a novel approach by not directly decoding stimuli from non-invasive data, but instead utilizing neural data to support the reconstruction process. Their method involved using a language model to generate several possible next words, then selecting the one that most closely aligns with the brain's current state. Although this method is based on off-line brain decoding using data acquired through fMRI, its innovative approach has sparked widespread interest from researchers.

In this study, we extend the work of Tang et al. (2023) by using three additional language models, along with the Fine-tuned GPT model (Radford et al., 2018a) they employed for language generation, in order to reconstruct sentences with higher similarity scores to the actual stimulus sentences, and compare the accuracy of the decoders. We investigate whether higher accuracy of the encoding model that predicts brain state leads to more precise decoding, as well as the factors that contribute to decoding accuracy.

2 Related Work

Tang et al. (2023) proposed a decoder that reconstructs continuous natural language from fMRI data acquired non-invasively, corresponding to any stimuli that participants are listening to or imagin-

Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025), pages 53–67
January 20, 2025.

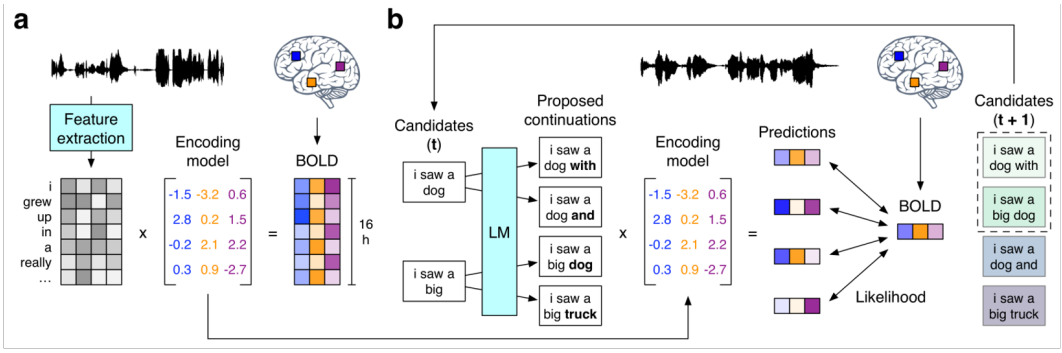

Figure 1: Reconstruction of sentences from brain data under language stimuli (adapted Tang et al., 2023). (a) An encoding model was constructed to predict BOLD responses obtained during an fMRI experiment from word sequences presented to participants. A total of 16 hours of data was used for training. (b) The language model generated candidate word sequences that could follow the given input. Using the trained encoding model, brain responses that can be evoked by these candidate sequences were predicted. The top k candidates, whose predicted responses were closest to the observed brain responses, were retained for the next time step.

ing. The overview is shown in Figure 1 (adapted from Tang et al., 2023). This decoder uses a language model to generate a set of candidate words and an encoding model trained to estimate the brain activity evoked by each candidate. The most likely word sequence, which best aligns with the actual brain state, is selected from these candidates. This approach mitigates the limitations of fMRI, which has low temporal resolution, enabling the reconstruction of sentences that participants are listening to.

Encoding models generally estimate brain states from vectors that represent stimuli, typically extracted from deep learning models. Since the introduction of word2vec (Mikolov et al., 2013), which represents the meaning of words in natural language as vectors, it has become possible to extract features from language stimuli presented to the human brain. More recently, intermediate representations from language models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2018b), and Llama (Touvron et al., 2023) have been increasingly used as vectors that capture sentence features for brain state estimation (Schrimpf et al., 2021; Caucheteux et al., 2021; Nakagi et al., 2024; Antonello et al., 2024). The performance of encoding models depends on the language model used. Antonello et al. (2024) reported that there is a scaling law between the number of parameters in the language model used for feature extraction and the accuracy of the resulting encoding model. As the number of parameters in the model increases, the accuracy of the encoding model improves in a logarithmic-linear fashion.

In this study, we introduce the Pre-trained GPT, the original model before fine-tuning in the research of Tang et al. (2023), and investigate how fine-tuning affects decoder accuracy. Additionally, while Tang et al. (2023) and other studies using encoding models have commonly employed GPT or GPT-2 for feature extraction, we use the powerful language models Llama3 and OPT to build a more accurate encoding model. Furthermore, we introduce a new evaluation approach that provides further insights into their performance to evaluate the effectiveness of Tang et al.'s decoding methods.

3 Method

3.1 Semantic reconstruction of language

The decoders developed in this study are based on the framework introduced by Tang et al. (2023). (Figure 1). Neural activity data were collected using fMRI while participants were exposed to auditory stimuli consisting of multiple stories narrated by a single speaker. To model the brain's response to natural language stimuli, an encoding model is first constructed to predict Blood-oxygen-level-dependent imaging (BOLD) responses under language stimuli using features extracted by a language model (Figure 1a). Theoretically, it is possible to identify the stimulus being perceived or imagined by the participant by comparing the measured neural response with the predicted responses for all possible word sequences. However, the number of potential word sequences is prohibitively large, and many of these sequences are unlikely to adhere to typical grammatical rules or resemble natural language. To address this, Tang

et al. (2023) used a language model trained on large text datasets to constrain the candidates to grammatically coherent word sequences. The decoder employs beam search to retain the top k candidates that produce neural responses most similar to the measured brain activity at each time step (Figure 1b).

3.2 MRI Data and Experimental Tasks

In this study, we use the same dataset (LeBel et al., 2024) as the previous research, which is openly available through the neuroimaging database OpenNeuro[1]. The MRI data were acquired at the Biomedical Imaging Center of the University of Texas at Austin using a Siemens 3T MRI scanner. The dataset includes data from three healthy participants (one female) aged 23 to 36.

The fMRI parameters were as follows: repetition time (TR) = 2.00 s, echo time (TE) = 30.8 ms, flip angle = 71° multi-band factor (simultaneous multi-slice) = 2, and voxel size = 2.6 mm × 2.6 mm × 2.6 mm (slice thickness = 2.6 mm).

The stimulus dataset consists of 82 stories, each with a duration ranging from 5 to 15 minutes, extracted from *The Moth Radio Hour* and *Modern Love*. In each story, a single speaker narrates an autobiographical story as an audio stimulus. In this study, we use fMRI data that has been preprocessed by LeBel et al. (2023). The test data was collected while the participants listened to the story "Where There's Smoke" (10 minutes) from *The Moth Radio Hour*, under the same conditions as the training data. To enhance the signal-to-noise ratio, the experiment was repeated five times in separate MRI sessions, and the BOLD responses were averaged across these trials for each participant.

3.3 Language Model

We use the Fine-tuned GPT model, which was employed in the previous research, as the baseline. To assess decoder performance with different language models, we also utilize the Pre-trained GPT, Llama3-8B, and OPT-6.7B models (Table 1). The baseline Fine-tuned GPT was trained on a corpus consisting of over 20 billion words from Reddit comments and 240 autobiographical stories (over 400,000 words) extracted from *The Moth Radio Hour* and *Modern Love*, which were not used in the fMRI experiments. The GPT was pre-trained on a story-like dataset, while the Llama3 and OPT models were pre-trained on corpora from books, news, websites, etc. All the Pre-trained models were obtained from Hugging Face Hub (details in Table A4) and were not trained by the authors.

The same language model was used for both feature extraction in the encoding model and for generating candidate words in the decoder.

3.4 Encoding Model

The encoding model explains information about stimuli or tasks represented in the activity of single voxels by predicting BOLD signals using linear regression based on features extracted from the stimuli (Naselaris et al., 2011). Language features used in the encoding model are extracted from the hidden states of the target token by feeding a sequence of the previous five tokens and the target token into a language model. The token features are downsampled to match the MRI repetition time (TR) using a Lanczos filter. To account for the temporal delay in the BOLD response, features from 1 to 4 TRs[2] before the stimulus are combined and included in the regression.

Ridge regression, commonly used in encoding models, is employed in this study. The regularization parameter α is selected from 10 values within the range of 10^1 to 10^3 for each voxel, based on a 50-fold cross-validation.

3.5 Token Rate Model

For each participant, we estimate a model to predict the number of tokens at specific time points, corresponding to when a new word was perceived or imagined. BOLD signals from voxels in the auditory cortex are used to train a linear regression model that predicts the number of tokens presented between time $t-1$ and t. The auditory cortex of each participant was defined using an auditory localizer task where participants listened to a one-minute stimulus, repeated 10 times, consisting of 20 seconds of music (Arcade Fire), speech (Ira Glass, This American Life), and natural sounds (such as a babbling brook).

Similar to the encoding model, we account for the temporal delay in the BOLD signal response to the stimulus by combining features from 1 to 4 TRs after the stimulus and performing regression. Next, we divide the predicted number of tokens by 1 TR to estimate the token input times. Although

[1] https://openneuro.org/

[2] 1 TR = 2.0 seconds

Model	Dim.	Layers	Params	Vocab	Training Data
FT GPT	768	12	120M	17378	Reddit comments and autobiographical stories
PT GPT	768	12	120M	40478	Unpublished books across various genres
PT Llama3	4096	32	8B	128000	Large public text datasets
PT OPT	4096	32	6.7B	50272	Books, story-like data, news, Reddit posts, web text

Table 1: Language models used in this study. "FT" represents Fine-tuned, and "PT" represents Pre-trained. Fine-tuned GPT, as employed in previous research, as the baseline, with additional models including Pre-trained GPT, Llama3-8B, and OPT-6.7B, which differ in training datasets and model sizes. All Pre-trained models used in this study were on Hugging Face.

this model is referred to as the word rate model in previous study, this study extends the word rate model to a token rate model since not all language models treat words as tokens.

3.6 Beam Search Decoder

Evaluating all possible word sequences is computationally impractical, so the decoders use a beam search algorithm to approximate the most likely sequence.

When a new token is detected by the token rate model, the language model generates candidate continuation words for each beam. The encoding model is then used to estimate the predicted brain state for all candidates. The likelihood of a candidate word sequence given the observed brain response is calculated using a multivariate Gaussian distribution, and the most likely word sequence is kept in the beam.

3.7 Evaluation Method

To evaluate how well the decoders reconstruct sentences from brain activity, we measure the similarity between the decoder-generated sentences and the actual stimuli the participants heard. Previous study used metrics such as word error rate (WER), BLEU, METEOR, and BERTScore (Zhang et al., 2020) for evaluation. However, considering that the language model used in previous study was fine-tuned on the same corpus used for testing and had vocabularies closely matching the actual stimuli, it is more challenging for the three new models, which were trained on entirely different corpora, to perfectly match the decoded words with the actual stimuli. As WER, BLEU, and METEOR are low-level metrics based on word matching, they proved less meaningful for the three new models (see Figure A5). Therefore, we focus on BERTScore, a higher-level metric that evaluates the semantic similarity between the generated and reference texts. We calculate BERTScore in the same manner as described in previous study, using inverse document frequency (IDF) weights derived from the training dataset and computed the recall score. In order to provide a more accurate evaluation, this study adopt the 750M DeBERTa (He et al., 2021) xlarge model which has been reported by the BERTScore authors to achieve the best performance, while previous study used the 355M RoBERTa (Liu et al., 2019) large model to calculate BERTScore.

In addition to BERTScore, this study incorporates sentence similarity evaluation using an embedding model. Although we have not directly compared accuracy with the model used for BERTScore, LLM-based embedding models have become widely used in tasks such as clustering, search, and retrieval-augmented generation (RAG) (Lewis et al., 2021) in recent years (Lee et al., 2024). We use OpenAI's embedding model[3] to extract embeddings for each sentence, and the similarity between the actual stimulus and the decoded sentence is assessed by calculating the Pearson correlation coefficient between their embeddings.

Sentence similarity is evaluated in terms of both window similarity and story similarity. Following previous research, window similarity is calculated based on word sequences within a 20-second window, while story similarity is calculated by averaging the window similarities.

4 Experiments

4.1 Performance of Encoding Model

Figure 2 shows the performance of encoding models built for three participants using different language models, evaluated with Pearson correlation

[3]text-embedding-3-small

on the test dataset. For each participant, the average correlation between the predicted and observed test brain data was calculated across cortical voxels that met the false discovery rate (FDR) threshold ($q < 0.05$). The gray bars represent the average values across all participants ($n = 3$). Encoding models constructed with Llama3 and OPT outperformed those built with GPT models in their highest-performing layers. This result aligned with previous studies showing that larger language models tend to achieve better accuracy in predicting BOLD signals (Antonello et al., 2024). Additionally, GPT and OPT models were reported to peak in deeper layers, while Llama family model showed peak performance in shallower layers, consistent with prior findings (Antonello et al., 2024; Wang et al., 2024).

Figure 3 presents a cortical flat map showing the accuracy of the encoding model for participant S02 using the Fine-tuned GPT($q(FDR) < 0.05$). Results for other participants and language models can be found in Figure A6. As observed in prior work with the same dataset (LeBel et al., 2023), regions like the parietal cortex, temporal cortex, and prefrontal cortex showed high accuracy.

The encoding models used in the decoders were chosen based on the layers that exhibited the highest prediction accuracy in an initial analysis without test data. For Fine-tuned GPT, Layer 9 was used; for Pre-trained GPT, Layer 10; for Llama3, Layer 13; and for OPT, Layer 22.

4.2 Performance of Token Rate Model

The accuracies of the token rate model on the test data, measured by Pearson correlation, are shown in Table 2 ($n = 3$).

Model	Pearson correlation
FT GPT	0.740 ± 0.012
PT GPT	0.708 ± 0.011
Llama3	0.722 ± 0.009
OPT	0.729 ± 0.008

Table 2: The Pearson correlation coefficients for the token rate models of each language model.

4.3 Decoder Setting

In this study, we used top-p sampling as the candidate word generation strategy for the generative model. Specifically, we used the probability mass parameter P_{mass}, which was set to 0.9, to represent the cumulative probability of the candidate words, and the relative probability threshold parameter P_{ratio}, set to 0.1, to evaluate whether a candidate word retains sufficient probability compared to the most probable word. This approach prioritized high-probability vocabulary while minimizing the loss of generation diversity.

Large language models typically include a special token to indicate the beginning of a sentence. However, to align with the settings of previous studies, the sentences generated by the decoders were set to begin with one of the following pronouns: 'He,' 'I,' 'It,' 'She,' or 'They,' and decoding was performed using beam search with $k = 5$.

The top 10,000 voxels with the highest accuracy in cross-validation were used for each participant to calculate the likelihood $P(S|R)$ of each candidate word sequence S given the observed brain state R.

4.4 Statistical Testing

We evaluated 300 sentences generated by the same language models used for the decoders without using brain activity, in order to assess whether the decoder-generated sentences scored significantly higher. Null distributions were established by calculating the similarity between each of the 300 generated sentence and the actual sentences. We then conducted a hypothesis test under the null hypothesis that the decoder cannot reconstruct sentences reflecting brain activity. The p-value was calculated as the proportion of the 300 sentences that had a score equal to or higher than those generated by the decoders, with multiple comparisons corrected using FDR.

4.5 Decoding Results

Figure 4a illustrates the results for story similarity, demonstrating whether the entire decoded sentence is significantly similar to the actual stimulus sentence. The null distribution, depicted as Chance, is composed of sentences generated by each language model without brain data and thus varies across models. For all language models and participants, the reconstructed sentences were significantly more similar to the actual stimuli than chance level ($q(FDR) < 0.05$). Figure 4b illustrates the results for window similarity, demonstrating whether the decoded sentence at each time point is significantly similar to the actual stimulus sentence (results for other participants are pro-

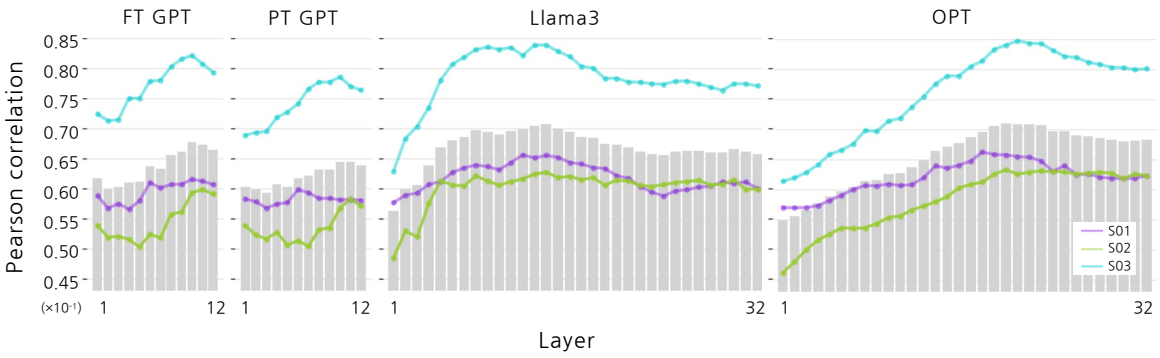

Figure 2: Encoding model accuracy for each language model (q(FDR) < 0.05). The gray bars represent the average scores across all participants (n $= 3$).

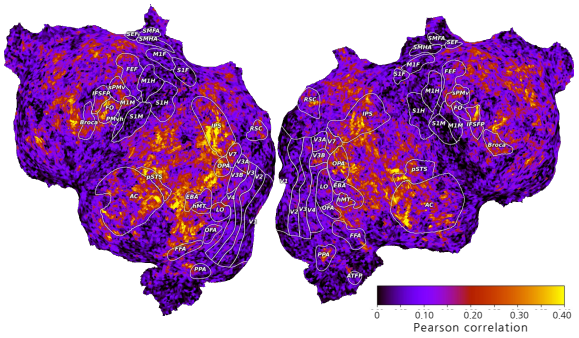

Figure 3: The encoding model accuracy mapped onto the cortical surface for a single participant (q(FDR) $<$ 0.05).

vided in Figure A7). BERTScore analysis of window similarity revealed that Fine-tuned GPT exhibited significant similarity at most time points (94%), whereas the other three models showed significant similarity at only 28-44% of the time points. Evaluations of window similarity using the embedding model showed significant scores at most time points for all language models (58-82%).

The actual sentences heard by the participants and the corresponding parts generated by each decoder are shown in Table 3 (see more in Table A5-A8). Decoders based on larger models, like Llama3 and OPT, produced more "rich" sentences, with distinctions between uppercase and lowercase letters and the inclusion of symbols. However, for evaluation, the text was standardized to match the dataset's notation, with all text converted to lowercase and punctuation (except apostrophes) removed. For all language models, we observed that the highlighted portions of the reconstructed sentences contained word sequences that closely resembled the meaning of the actual stimuli. For instance, in Example 1, the word *light* was matched with terms such as *candle* and *screen was brighter*, and a scene involving multiple people conversing was also reconstructed. In Example 2, for a stimulus sentence containing words like *car* and *road*, the decoders reconstructed sentences with terms such as *car*, *road* and *drive* which also suggests the concept of a vehicle.

5 Discussion

5.1 BERTScore vs. Embedding score

When examining the BERTScore for both story similarity and window similarity, we observed that the decoder using the Fine-tuned GPT yielded significantly higher scores than the decoder scores based on the other three language models (Figure 4a, b). The null distribution generated without using brain activity for Fine-tuned GPT, also yielded higher scores than the scores for the other decoder (Figure 4a), suggesting that the sentences generated by Fine-tuned GPT tended to be more similar to the actual stimuli compared to those generated by the other language models. We hypothesize that this is attributable to two factors: (1) the inclusion of a dataset in the training of Fine-tuned GPT that closely resembles the actual stimuli (though not used in the fMRI experiment), and (2) the relatively limited vocabulary size of Fine-tuned GPT compared to the other language models, which facilitates the frequent appearance of words and phrases from the actural stimuli in the generated sentences. These result in higher scores for both the decoded sentences and the null distribution in the Fine-tuned GPT.

In the evaluation using the embedding model, while there was no change in the rankings, the differences across the language models are smaller

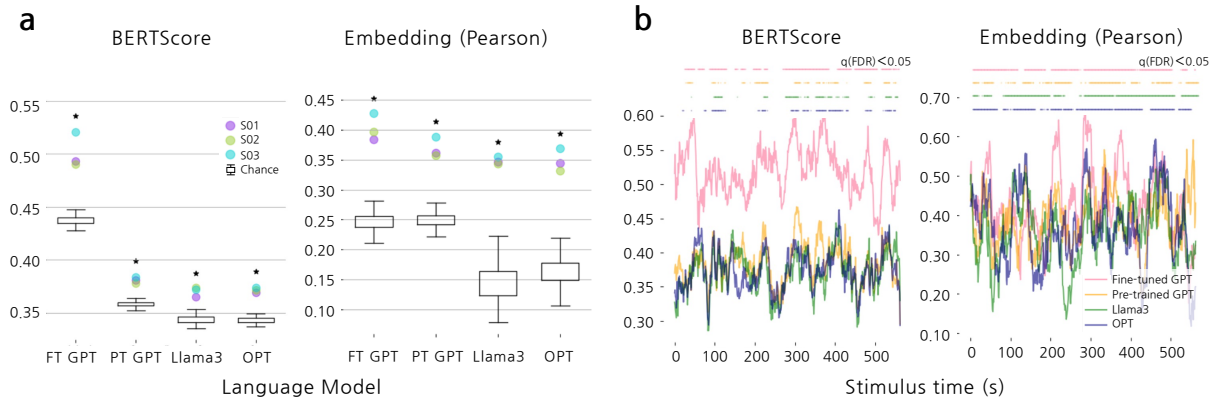

Figure 4: Score of sentence reconstruction by the decoders. (a) Story similarity, representing the overall similarity of reconstructed sentences. Box plots indicate the null distribution, and stars denote significantly higher scores (q(FDR) < 0.05). (b) Window similarity, representing the similarity within a 20-second window for an single participant. Lines above the graph indicate time points where each language model achieved significantly higher scores (q(FDR) < 0.05).

	Example 1	Example 2
Actual	**in that little crack of light** and i hear the man and he says where were you and she says never mind i'm back and **he says you alright**	**the roads** are getting wider and wider and there's more **cars** and i see um lots of stores you know laundromats and
FT GPT	the windshield a minute later and **the guy said to me are you okay** and i replied well i'm fine and he says ok	little trail and then **the main road** and the trees and there are houses and some kind of town hall and a gas station
PT GPT	**candle** in the foyer burning bright is it time to leave yet no i'll be back soon	i'll rent a **car** and **drive** my first step is to find a **car** rental agency a small town a bank and
Llama3	**my phone's screen was brighter** than the sun it's time to sleep i'll see you soon okay i love you	as we **drive** i explain what we'll do when we arrive the warehouse is an old military surplus store now a gun shop
OPT	dozen different calls how long are you here i have to go i'm sorry i'll see	i **drove** i **drove** to the only place i knew of a diner a greasy spoon a diner in a strip

Table 3: The actual stimulus sentences and the sentences reconstructed by the decoders of each language model at two different time points for a single participant. Parts with similar expressions are highlighted in bold.

than those observed with BERTScore for both story similarity and window similarity. Even with BERTScore, a method that compares the hidden states of models and measures the semantic similarity between tokens in two sentences, we believe that the high scores are likely observed due to the presence of identical words, especially considering that DeBERTa XL, the model used to calculate the scores, is not a "large" model. On the other hand, the evaluation using the embedding model is considered to measure similarities in higher-level semantic representations rather than at the word/token level. In this evaluation, all language models demonstrated accuracy surpassing the null distribution of Fine-tuned GPT. It can be concluded that all language models were able to reconstruct sentences that were significantly similar to those the participants might have heard or imagined.

5.2 Factors Underlying Variations in Scores

When comparing Pre-trained GPT with two larger models (Llama3 and OPT), despite the higher accuracy of the encoding models in the larger models (Figure 2), indicating better predictions of brain states, the decoder based on Pre-trained GPT achieves slightly higher accuracy (Figure 4a). We hypothesize that this discrepancy is attributable to differences in the training datasets used for each model. Larger models typically require vast amounts of training data, which often includes datasets that differ significantly from the autobiographical stories used as actual stimuli. In contrast, Pre-trained GPT was trained on story-like data, making it more likely to generate sentences similar to the actual stimuli. The null distribution of Pre-trained GPT being positioned higher than that of the larger language models further supports this assumption.

It is important to note that while a larger language model may improve the accuracy of brain state estimation, it does not necessarily guarantee to more precise reconstructions of the brain's representations. In scenarios like the this study, where the stimulus dataset applied to the decoder is already well-defined, using a language model capable of generating outputs similar to the stimulus dataset allows for more precise reconstructions. On the other hand, when the stimulus dataset is not clearly defined in the fMRI experience, employing a language model with a larger vocabulary or one trained on diverse datasets may be crucial, as it allows for the generation of a wider array of possible outputs.

6 Conclusion

In this study, we examined and expanded upon Tang et al.'s research, which proposed the use of language models for brain decoding. Specifically, in addition to the Fine-tuned GPT model used in previous study, we constructed decoders using three additional language models, clarified the accuracy of the encoding models and the token rate models used in the decoders, and compared their decoding performance.

Regardless of the language model used, we confirmed that the decoders could significantly reconstruct sentences similar to the actual stimuli presented to participants. Although larger models like Llama3-8B and OPT-6.7B demonstrated superior performance in predicting brain activity, we found that the GPT (120M) models achieved higher decoding scores. We hypothesize that this result is attributable, at least, to the training dataset of the GPT models being more similar to the actual stimulus sentences.

Moreover, this study added a similarity evaluation metric using an embedding model by computing higher-level semantic similarities between sentences, demonstrating that all language models successfully reconstructed sentences with significantly high scores at most time points.

While this study focused solely on evaluating the similarity between the actual stimulus sentences and the decoded sentences, such similarity does not necessarily guarantee an accurate reflection of brain status. Unlike this study, when the stimulus dataset in the fMRI data is not explicitly known, using language models trained on more diverse datasets could potentially result in a better reconstruction of brain states.

7 Limitation

In this study, we examined whether decoders reported in previous research function similarly across different language models and compared the decoding accuracy between them. Although this decoder's main objective is to reconstruct sentences that participants are likely hearing or imagining, the sentences participants are hearing are clearly defined in the experiment while the sentences they may be imagining remain unknown. We confirmed the decoder's accuracy by assessing the similarity to the sentences the participants are hearing, but if participants are imagining sentences that differ from the given stimuli (e.g., based on personal experiences or different contexts), a decoder closely matching the stimulus sentences may not necessarily be ideal. To evaluate the similarity with the sentences participants are imagining, relying solely on similarity measures between the actual and decoded sentences would be insufficient, and additional evaluations, such as comparing the similarity between predicted brain responses from the decoded sentences and actual brain responses, would likely be required.

This study also supported the differences in decoder performance due to variations in the training dataset. However, identifying the differences in performance based on model size remains a challenge for future work.

Finally, while we confirmed the decoder's effectiveness by applying it to data from the same participants used for training, the performance of the decoder across different participants remains unverified.

Acknowledgments

This study was supported by JSPS KAKENHI (23K18489).

References

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. 2024. Scaling laws for language encoding models in fmri. *Preprint*, arXiv:2305.11863.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2021. Gpt-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

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

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. 2024. "an fmri dataset during a passive natural language listening task".

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. 2022. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in Human Neuroscience*, 16:867281.

Sean L. Metzger, Jessie R. Liu, David A. Moses, Maximilian E. Dougherty, Margaret P. Seaton, Kaylo T. Littlejohn, Josh Chartier, Gopala K. Anumanchipalli, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. 2022. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13:6510.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Preprint*, arXiv:1310.4546.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. 2024. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20313–20338, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. 2011. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners.

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

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yuejiao Wang, Xianmin Gong, Lingwei Meng, Xixin Wu, and Helen Meng. 2024. Large language model-based fmri encoding of language functions for subjects with neurocognitive disorder. *Preprint*, arXiv:2407.10376.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Tianyi Zhang, Varsha Kishore*, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Language Models On Hugging Face

The Hugging Face model IDs used are listed in Table 4.

Model	ID
GPT	openai-community/openai-gpt
Llama3	meta-llama/Meta-Llama-3-8B
OPT	facebook/opt-6.7b

Table 4: The IDs of the Hugging Face models used.

A.2 Other Similarity Evaluation Metrics

As discussed in Section 3.7, the previous study has evaluated performance using metrics such as WER, BLEU-1, and METEOR. In our experimental setting, summarized in Figure 5, only the Fine-tuned GPT decoder, optimized for generating sentences closely resembling the actual stimuli, achieved statistically significant scores across all metrics. It consistently outperformed the other three language models, showing a much higher degree of word-level similarity. The lower scores observed for the other models suggest that generating identical words poses a greater challenge for them.

A.3 Performance of Encoding Model

Figure 6 presents the results of the encoding models constructed for each subject and each language model. Across all language models, higher accuracies were consistently observed in regions such as the parietal cortex, temporal cortex, and prefrontal cortex, with no discernible differences between the language models.

A.4 Window Similarity Between Actual and Reconstructed Sentences

The window similarity between the stimulus sentences heard by the participants and those reconstructed by the decoder was computed using the procedure outlined in Section 3.7. Figure 7 presents the results for participants not included in the main text. As detailed in Section 5.1, the Fine-tuned GPT exhibited significantly higher scores in the BERTScore evaluation. On the other hand, the differences in performance were not as pronounced in the evaluation using the embedding model.

A.5 Decoder predictions for a perceived story

The reconstructed sentences produced by each decoder are presented in Table 5-8. Line breaks were removed during preprocessing to improve readability.

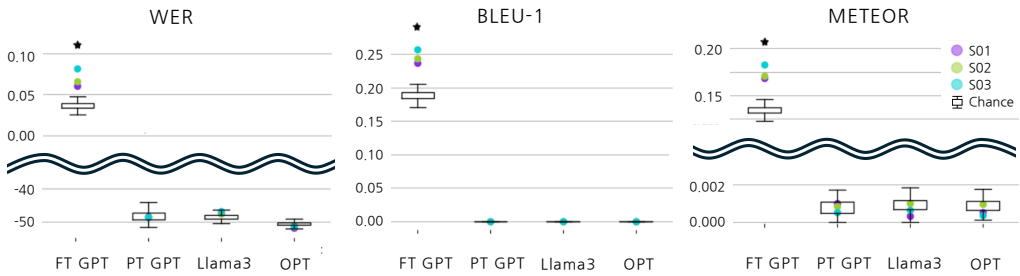

Figure 5: Story similarity based on word-level evaluation metrics. Box plots indicate the null distribution, and stars denote significantly higher scores (q(FDR) < 0.05).

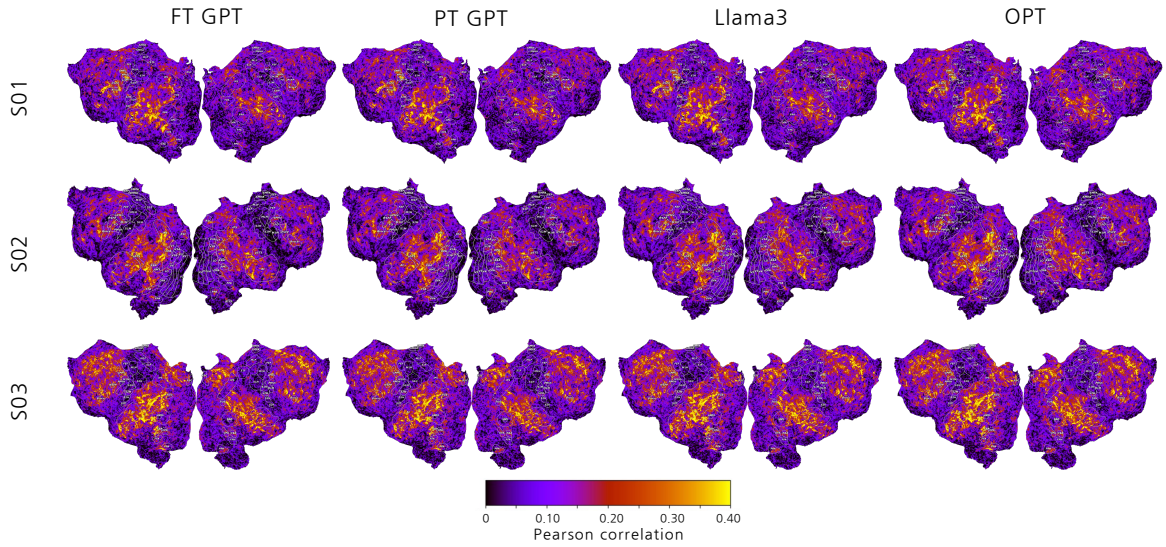

Figure 6: Cortical maps showing the encoding model accuracy for each participant and each language model used.

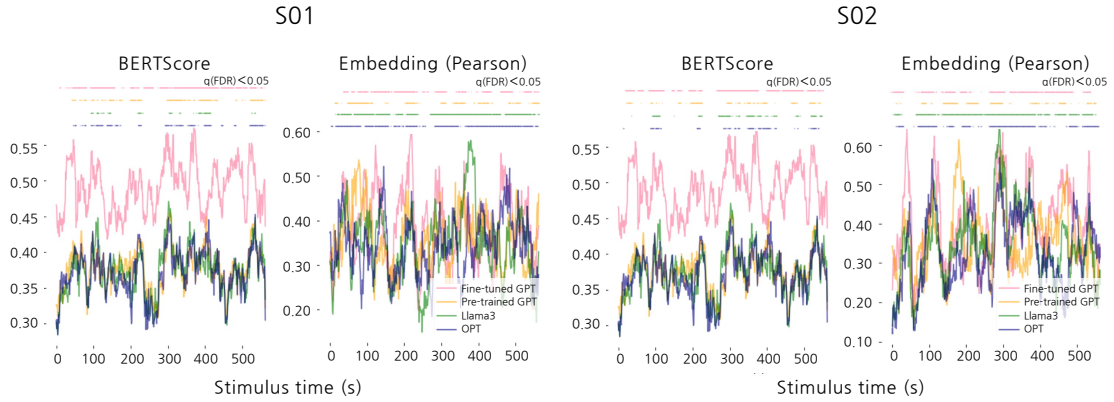

Figure 7: Window similarity, representing the similarity within a 20-second window. Lines above the graph indicate time points where each language model achieved significantly higher scores (q(FDR) < 0.05)

Actual	S01	S02	S03
she digs back in the front again deep deep and she pulls out a pack of matches that had been laundered at least once ukgh we open	got down to my underwear and pulled it out of his pants to find a huge pile of cash that was probably on the floor he	had to go back to the apartment or even look for anything i was homeless in a really nice area so i had some	pulled the top of the bag aside and found a large amount of weed that was probably half a pack the guy had
it up and there is one match inside ok oh my god this takes on it's like nasa now we got to like oh how are we gonna do it ok and we we hunker down	took it all and said i'm so sorry about this i don't think anyone can help you now it's all done now so it's really good to be	money saved up and had enough for a few drinks to take the edge off so i decided to just sit in the car with my feet on the	to get some and i was like ok we need some you know how you want to go with the flow so we did this thing where you put your
we crouch on the ground and where's the wind coming from we're stopping i take out my cigarettes let's get the cigarettes ready oh my brand she says not surprising and	able to see the light on the way out my mom says ok let me go grab the rest of the food i am pretty sure this	seat and the engine running i took my hand out and said you can help me with the gas my dad was right there at that point	feet up and then you lean over to get your balance and the guy says can you grab your seat belt i got you my friend and he does this i take
we both have our cigarettes at the ready she strikes once nothing she strikes again yes fire puff inhale mm sweet kiss of that cigarette	is my mother so i do i start eating and it is delicious it tastes like heaven i feel so relaxed and happy	so we put it in and it blew up with a little pop and a puff of smoke in it and the woman	it and we start to roll he pulls it tight and the ball explodes with a loud explosion of blood and
and we sit there and we're loving the nicotine and we both need this right now i can tell the night's been tough immediately we start to reminisce	as we sit and drink we have the perfect time to be together as a family i remember this when i was about we had been married for	got on her hands and knees and tried to get my friends to do it too because it was such a huge problem we all stood around for	gore and i feel this overwhelming need to cry for my family that i am in such pain over i think the last months of my life
about our thirty second relationship i didn't think that was gonna happen me neither oh man that was close oh i'm so lucky i saw you yeah then she	about six months and the day i asked him why he said you know what you did right and i didn't and then i	about minutes thinking how stupid this was we didn't see a damn thing i said hey guys get over here you two and i heard	was spent being afraid to ask what happened to me to make you hate me and what have you my friend and i were
surprises me by saying what was the fight about and i say wha what are they all about and she said i know what you mean she said was it a bad one and and i said	got an email from her asking me if i knew the guy that did this i replied no i did not know him i don't think you	a few of us say to each other are you sure i said something like you don't know i think he meant you did you see his	talking about when he told me what was happening i said what did you mean by that he replied oh nothing really i don't remember he was

Table 5: The reconstructed sentences by the **Fine-tuned GPT** decoder.

Actual	S01	S02	S03
she digs back in the front again deep deep and she pulls out a pack of matches that had been laundered at least once ukgh we open	i put it in a small envelope and sealed it with a plastic wrapper, hoping the little bit of gold was still in there.	last time i went back to the museum, it was full of creepy old people and weird stuff, but i got my	he reached back and found the nearest box. it was full of empty bottles, which meant the bottle would have to
it up and there is one match inside ok oh my god this takes on it's like nasa now we got to like oh how are we gonna do it ok and we we hunker down	he closed it and went on. "i don't know how it is that you can walk so slowly, but it's amazing. i could never	own place, so i've had a little extra to go around, so i'm just trying to keep my balance, "he replied, sitting up and	be filled. i grabbed two. "okay, i'll drink my fill of the whole thing, but i'm going to take my chances." i leaned
we crouch on the ground and where's the wind coming from we're stopping i take out my cigarettes let's get the cigarettes ready oh my brand she says not surprising and	walk so slowly, i have to carry my bag and the water bottle with me, but i will take the first water bottle, as i will	adjusting the sleeping bag, then reaching for the fire starter. "you could have brought your own lighter," she said to the room at large. she needed	over the bar to get a shot, and she held my hand for me to get her another. i got the shot and the bartender asked if she was a
we both have our cigarettes at the ready she strikes once nothing she strikes again yes fire puff inhale mm sweet kiss of that cigarette	be taking the other, so i give it a shake and it pops open, water splashing out in a cloud around me.	the flame to go out, and it did. she extinguished it with a few puffs of smoke from her pipe and held the	fan of my song, so she took the shot and shot it in the mouth and let the alcohol slide down the throat,
and we sit there and we're loving the nicotine and we both need this right now i can tell the night's been tough immediately we start to reminisce	i sit on the ground and wait, waiting to be able to see what i'm up against. the world seems to have gone on for a	pipe in her mouth as she smoked, waiting for me to join her. i was glad to have a cigarette after the long day. i found myself	letting it settle in and relax me, relax me until the alcohol took over and i was feeling everything i'd been feeling since i walked into his office the morning
about our thirty second relationship i didn't think that was gonna happen me neither oh man that was close oh i'm so lucky i saw you yeah then she	hundred years without me being there to tell you that you have to stop blaming yourself, you know that, right ?" i turned	wondering how much longer i could last out here. it was like a war zone here. "where are you, son ?"	after our little "break-in". i couldn't have been happier for him and his friends to see me. he told me to come to
urprises me by saying what was the fight about and i say wha what are they all about and she said i know what you mean she said was it a bad one and and i said	to see if she was talking about her friend. "who ?" "i didn't ask you who it was, i asked you if you	a man asks me. "don't you think the answer is obvious ?" "what is obvious is that he didn't come home	his house when he got home. i asked him what happened, he said it wasn't you who did it, he said it wasn't you and

Table 6: The reconstructed sentences by the **Pre-tuned GPT** decoder.

Actual	S01	S02	S03
she digs back in the front again deep deep and she pulls out a pack of matches that had been laundered at least once ukgh we open	went through the bag and found a plastic bag inside it with a few packets of instant noodles and biscuits. They were probably thrown away,	back to see if the tree grew up and turned into a fairy tale castle or something like that, but the house just	it wide and reached in, finding a plastic bag that was almost certainly filled with cocaine. He'd have to throw
it up and there is one match inside ok oh my god this takes on it's like nasa now we got to like oh how are we gonna do it ok and we we hunker down	and I'm glad. I hate when people try to be clever. This is a game. Let's all play. It's not about getting	got a few coats of paint, so I'm going to take my time with this and let it dry. So I got my 6š ladder and	it away. Right. Okay. We can't let that happen. Let's do some brainstorming on the fly and try to get our ducks in a row
we crouch on the ground and where's the wind coming from we're stopping i take out my cigarettes let's get the cigarettes ready oh my brand she says not surprising and	to the top of the ladder and then pushing the ladder away and telling the kids you'll find another ladder. (Yes, I was the parent	climbed up. I placed the nozzle of the hose just right and was ready to spray. My wife and daughter had no idea I had this set up	so that we can get the next three out the door. We need to get a hold of our ducks. He wasn't a fan of the new marketing
we both have our cigarettes at the ready she strikes once nothing she strikes again yes fire puff inhale mm sweet kiss of that cigarette	with the ladder.) So we all climbed up and down and up and down until the tree was stripped bare. My mother and	and when I hit the button, it sprayed out with a satisfying hiss and a puff of steam. I grabbed the towel	term. He'd tried it a few times, and the word made his tongue feel fat and heavy, and he swallowed
and we sit there and we're loving the nicotine and we both need this right now i can tell the night's been tough immediately we start to reminisce	I would stand and watch in silence, holding hands. I knew we were in a good place when my mother would go into a trance of sorts and	and held it out for her to dry her face. I wanted her to do this for me. The rain had stopped, and we had a little	it down and let it stay there, warm and sweet and making him feel better. The whole day seemed to change for him when he took a break from the pressure
about our thirty second relationship i didn't think that was gonna happen me neither oh man that was close oh i'm so lucky i saw you yeah then she	begin speaking to my father, saying things like you will do this and you will do that. And then I got	more daylight left. I knew I'd never see this place again. It was a shame. Hey, look at that! I	of the exam to talk about the weather with us. We all felt good for him and happy that he'd come to talk to us.
urprises me by saying what was the fight about and i say wha what are they all about and she said i know what you mean she said was it a bad one and and i said	a text from someone who knew that I was a Christian and they said, "I'm not sure if you know this but God doesn't hate you	heard a woman exclaim. What do you mean, not all'? Of course he was a woman. I saw his face	Where are you going to school now? Why do you ask? Did I do something wrong? Did I? You look different

Table 7: The reconstructed sentences by the **Llama3** decoder.

Actual	S01	S02	S03
she digs back in the front again deep deep and she pulls out a pack of matches that had been laundered at least once ukgh we open	I pulled down the zipper and found a pocket of tissue paper and a couple of loose pieces of gum. The gum was already	like and to imagine them living and working in this beautiful, richly-wooded country. I had a little money,	it over and found a little flap that was easily ripped open with a sharp knife. The plastic was already broken,
it up and there is one match inside ok oh my god this takes on it's like nasa now we got to like oh how are we gonna do it ok and we we hunker down	gone. "Oh, man, this is crazy. It's not funny, but it's not not funny. It's like when	so I bought a few bottles of wine. It was fun to mix them and then we all had a drink. I just sat on the floor with	so I used a little glue and it worked! I will definitely use this idea! I was so proud of my creation, I took it outside and held it
we crouch on the ground and where's the wind coming from we're stopping i take out my cigarettes let's get the cigarettes ready oh my brand she says not surprising and	a clown is running through the audience and he accidentally slips on the banana peel. I'll take the banana peel. Oh, great.	my legs dangling and had my friend hold my drink for me. You can always use your own cup! I don't get this thing that you have	in the air to get the best picture. I then got my son to help me with the light meter. He was not impressed. My daughter's
we both have our cigarettes at the ready she strikes once nothing she strikes again yes fire puff inhale mm sweet kiss of that cigarette	So I'll just slip and slide and then slide and slip and slide until I'm a quivering mass of ice.	to wait for the bar to get empty. Just pour a shot into the cup, put a straw in, and put	light meter is much better. She took one shot, set it to 100 and let it sit on my face for
and we sit there and we're loving the nicotine and we both need this right now i can tell the night's been tough immediately we start to reminisce	I sit down and watch the boys play, my hands trembling. I know what I saw. A long time ago I wrote a series	the cup in front of her, and let her drink. "This is all I could spare. I had no more than two coins left after	a bit. The warmth and moisture help me wake up and get my day going. I find that when I am in the office, the morning routine is often interrupted
about our thirty second relationship i didn't think that was gonna happen me neither oh man that was close oh i'm so lucky i saw you yeah then she	of poems, beginning with the line, I will never know what you do not know. But this	I got home." "You can't leave now! There are so many people here!" I shouted back at him,	by people wanting to chat about the previous night. I love it when people are happy to see you, and it's just nice to see you.
urprises me by saying what was the fight about and i say wha what are they all about and she said i know what you mean she said was it a bad one and and i said	time, when the woman asked me if I was gay, I said "I don't have to answer that, but no I'm not gay".	and I heard someone say, "Are you serious? Why is that a question? She's obviously a lesbian. Why else	"When I heard you were in town, I said, 'Why, she's not the one, is she?' 'No, she

Table 8: The reconstructed sentences by the **OPT** decoder.