The Brain Tells a Story: Unveiling Distinct Representations of Semantic Content in Speech, Objects, and Stories in the Human Brain with Large Language Models

Yuko Nakagi [1 2 *] **Takuya Matsuyama** [1 2 *]
Naoko Koide-Majima [1 2] **Hiroto Yamaguchi** [1 2] **Rieko Kubo** [1 2]
Shinji Nishimoto [1 2 † ‡] **Yu Takagi** [1 2 † ‡]

Abstract

In recent studies, researchers have utilized Large Language Models (LLMs) to investigate semantic representation within the brain. However, in many of these studies, the researchers often examined various semantic information contents separately, such as speech content, objects in scenes, and background stories. To quantitatively evaluate the contribution of various semantic contents in the brain, we recorded brain activity using functional magnetic resonance imaging (fMRI) while participants watched a total of 8.3 hours of videos of dramas or movies. Importantly, we densely annotated these videos at multiple semantic levels related to video contents, which allowed us to extract latent representations of LLMs for a range of semantic contents. We show that LLMs explain human brain activity more accurately than traditional language models, particularly for the high-level background story. Additionally, we show that distinct brain regions correspond to different semantic contents, thereby underscoring the importance of simultaneously modeling various levels of semantic contents. We will make our fMRI dataset publicly available for future research as a biological metric of the alignment between LLMs and humans.

1. Indroduction

Language models, which learn the statistical structure of languages from large corpora to enable machines to understand semantics, have made significant advances (Devlin et al., 2018; Radford et al., 2019; Zhang et al., 2022; Touvron et al., 2023). Because semantic comprehension is also fundamental to human intelligence, the correspondence between human brain activity during semantic comprehension and the latent representations of language models has been an intriguing subject of research. To examine the correspondence between human and language models, in recent studies, researchers used **brain encoding models** (Naselaris et al., 2011; Nishimoto et al., 2011; Huth et al., 2012) to predict brain activity based on the high-dimensional latent representations of language models (Jain & Huth, 2018; Jat et al., 2019; Toneva & Wehbe, 2019; Caucheteux et al., 2021; Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux et al., 2022; Antonello et al., 2023). The quantitative assessment of the correspondence between the two can serve as a unique benchmark for modern large language models (LLMs) because it potentially provides a biological benchmark for the alignment between LLMs and humans.

In previous studies, researchers typically focused on a single aspect of semantic comprehension (e.g., speech content); however, realistic scenarios are inherently multifaceted. For instance, a scene in which two people are speaking can be depicted through multiple semantic levels in language: their speech content, their identities and visual appearance of their outfits, the location and time of the scene, and the broader context of the conversation. In traditional studies, researchers often addressed these levels of semantics separately, which has led to a lack of clarity on how multiple semantic contents are attributed to brain activity and how the latent representations of language models might align with the human brain processing of multiple contents. Understanding the world involves these multiple semantic levels, which makes it important to examine how the brain and language models might represent them in a comprehensive manner.

In this study, we addressed these issues by recording human brain activity using functional magnetic resonance imaging (fMRI) while participants viewed 8.3 hours of videos of dramas or movies. Importantly, these videos were densely

* Equal first author. † Equal last author.
[1]Osaka University, Japan [2]NICT, Japan.
‡ Correspondence to: Shinji Nishimoto <nishimoto.shinji.fbs@osaka-u.ac.jp>, Yu Takagi <takagi.yuu.fbs@osaka-u.ac.jp>.

Preprint. Under review.

annotated across multiple levels of semantic contents related to the drama, including speech, visual objects, and background story. We extracted latent representations of these annotations from LLMs, then built encoding models that predicted brain activity from these latent representations to quantitatively compare how each type of information was represented in different brain regions. Furthermore, we quantitatively assessed how each brain region uniquely captured different aspects of semantic contents, in addition to the relationships with the cortical representations of visual and auditory modalities.

Our contributions are as follows:

1. We show that latent representations of LLMs explain human brain activity more accurately than traditional language models, particularly for high-level story contents.
2. Additionally, we show that distinct brain regions uniquely correspond to different semantic contents. We also demonstrate that these semantic contents possess unique variance that is explained when compared with vision and audio modalities. These findings underscore the importance of simultaneously modeling various levels of semantic contents.
3. We collected and will publish densely annotated fMRI datasets acquired while the participants watched 8.3 hours of videos. In addition to existing fMRI data, which focus on images or short movie clips, our dataset can be another benchmark for the biological metric of the alignment between LLMs and humans.

2. Related work

In many studies, researchers have focused on the relationship between the latent representations of language models and brain activity during speech comprehension. In these studies, researchers mainly investigated the correspondence between the latent representations of language models obtained from speech transcription and brain activity (Jain & Huth, 2018; Jat et al., 2019; Toneva & Wehbe, 2019; Schmitt et al., 2021; Caucheteux et al., 2021; 2022), and recently indicated that larger language models explain brain activity well compared with traditional language models (Schrimpf et al., 2021; Goldstein et al., 2022; Antonello et al., 2023).

Regarding the semantic contents of the visual object in the scene, the correspondence between the latent representations of deep learning models related to the displayed objects and human brain activity has also been studied (Güçlü & van Gerven, 2015; Horikawa & Kamitani, 2017; Groen et al., 2018; Wen et al., 2018; Khosla et al., 2021; Allen et al., 2022; Chen et al., 2023; Takagi & Nishimoto, 2023). In most of these studies, researchers used neuroimaging data while participants watched static images or short video clips.

Although several neuroscience studies have been conducted in which researchers explored high-level story contents in the brain using a naturalistic movie watching experiment (Hasson et al., 2008; Baldassano et al., 2017; Chang et al., 2021; Nastase et al., 2021), these researchers typically did not explicitly model high-level story content using language models. Furthermore, they have not examined the unique explanatory power of story-specific semantic content on brain activity when compared to other types of semantic contents.

While there is a growing consensus that LLMs mirror human brain activity more accurately than traditional language models during semantic comprehension, individual studies addressed discrete aspects of semantic comprehension independently, and in most studies, researchers did not explicitly model the background story. This is problematic because humans process different types of semantic contents simultaneously in naturalistic scenarios, and these may be represented differently in LLMs and the human brain. For this study, we collected a neuroimaging dataset while participants watched 8.3 hours of naturalistic videos with dense annotations for different levels of semantic contents. We quantitatively evaluated how much each type of semantic content uniquely explained the brain activity.

3. Methods

3.1. fMRI experiments

We collected brain activity data from six healthy participants with normal and corrected-normal vision (three females; age 22–40, mean = 28.7) while they freely watched 8.3 hours of videos of movies or drama series. All participants were Japanese and provided written informed consent for the study and the release of their data. The ethics and safety committees of the National Institute of Information and Communications Technology approved the experimental protocol.

MRI data were acquired using a 3T MAGNETOM Vida scanner (Siemens, Germany) with a standard Siemens 64-channel volume coil. Blood oxygenation level-dependent (BOLD) images were acquired using a multiband gradient echo-planar imaging sequence (Moeller et al., 2010) (TR = 1,000 ms, TE = 30 ms, flip angle = $60°$, voxel size = $2 \times 2 \times 2$ mm^3, matrix size = 96×96, 72 slices with a thickness of 2 mm, slice gap 0 mm, FOV = 192×192 mm^2, bandwidth 1736Hz/pixel, partial Fourier 6/8, multiband acceleration factor 6). Anatomical data were collected using the same 3T scanner using T1-weighted MPRAGE (TR = 2530 ms, TE = 3.26 ms, flip angle = $9°$, voxel size = $1 \times 1 \times 1$ mm^3, FOV = 256×256 mm^2). The preprocessing of the

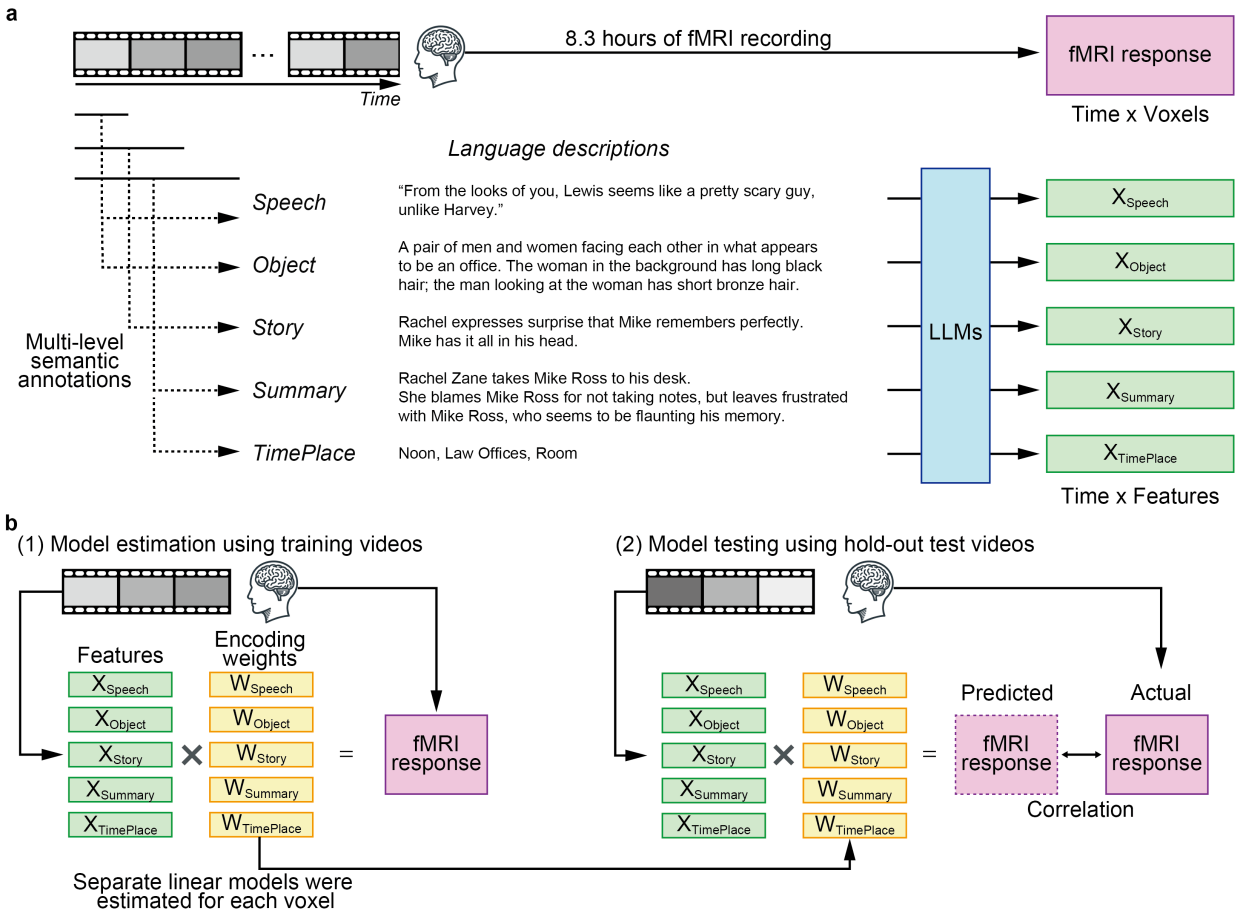

Figure 1. **Overview of the experiment and brain encoding models. a** Participants watched 8.3 hours of videos of dramas or movies while their brain activity was measured inside an fMRI. The videos were densely annotated with five types of distinct semantic contents, which were used for extracting latent representations from multiple language models. The annotations provided examples from specific scenes in the series '*Suits*'. **b** For each semantic feature obtained by the language models, linear weights for predicting brain activity across the cerebral cortex from the feature were estimated using ridge regression. The estimated weights were subsequently applied to the features of the test data to estimate brain activity. Prediction performance was calculated using Pearson's correlation coefficient between the predicted and actual fMRI responses.

functional data included motion correction, coregistration, and detrending. See Section A.1 for details of the acquisition and preprocessing procedures.

3.2. Stimuli

We used nine videos of movies or drama series as stimuli (10 episodes in total), with dense annotations related to those videos. The videos encompassed various genres. Eight were international videos and one was a Japanese animation. The average playback time of the 10 episodes was 49.98 minutes (ranging from 21 minutes to 125 minutes). Each episode was divided into two to nine parts, each approximately 10 minutes long, for use as stimulus videos during the fMRI experiment. All the international videos were played in Japanese dubbed versions, and the subjects understood them in Japanese. We will make our stimuli and fMRI dataset publicly available for future research upon acceptance. See Section A.1.1 for details of data collection.

The annotations included five levels of semantic contents from the videos: transcriptions of spoken content (*Speech*), objects in the scene (*Object*), background story of the scene (*Story*), summary of the story (*Summary*), and information about time and location (*TimePlace*) (Figure 1a). The semantic contents consisted of annotations that described the stimulus videos in natural language, but they differed in the nature of the description content and the timespan of the annotations. Specifically, *Speech* corresponded to the intervals of speaking, whereas *Object* was annotated every second,

Model	Layers	Width	Parameters
Word2Vec (Mikolov et al., 2013)	1	300	900M[1]
BERT (Devlin et al., 2018)	12	768	110M
GPT2 (Radford et al., 2019)	36	1280	812M
OPT (Zhang et al., 2022)	32	4096	6.7B
Llama 2 (Touvron et al., 2023)	32	4096	6.74B

Table 1. **Summary of the language models' architecture.**

Story every 5 seconds, *Summary* every 1 to 2 minutes, and *TimePlace* every time the screen changed. Each semantic content, except for *Speech*, was independently labeled by multiple annotators: five for *Object*, three each for *Story* and *TimePlace*, and two for *Summary*. See Section A.2 for details and more examples of annotations.

We divided the data into training and test datasets, and calculated all the prediction performance results presented in the paper using the test dataset. Specifically, we used the fMRI scanning sessions corresponding to the last split of each movie or drama series, 7,779 seconds in total, as test data. We used the remaining sessions, 22,214 seconds in total, as training data.

3.3. Feature extraction

To extract the latent representations of each semantic content, we obtained latent representations from five language models (Figure 1a). Information about the language models used is presented in Table 1. We used Word2Vec (Mikolov et al., 2013) as a traditional language model. For recently developed language models, we used BERT (Devlin et al., 2018), GPT2 (Radford et al., 2019), OPT (Zhang et al., 2022), and Llama 2 (Touvron et al., 2023). We used GoogleNews-vectors-negative300, bert-base-uncased, gpt2-large, facebook/opt-6.7bm, and meta-llama/Llama-2-7b-hf models available on Huggingface for Word2Vec, BERT, GPT2, OPT, and Llama 2, respectively.

Except for Word2Vec, where we obtained word embeddings, we extracted the latent representations of language models for the annotations from each of the hidden layers. For each latent representation from each hidden layer, we built brain encoding model (see Section 3.4). Because there were multiple annotations for each second except for *Speech* annotations, we extracted the latent representations of each annotation from each language model and then averaged the latent representations among all annotations. We also averaged the latent representations of LLMs across tokens, and averaged the word embeddings of Word2Vec across words.

We also explored how uniquely different semantic contents explained brain activity compared with audio and visual features. For visual features, we extracted latent representations from DeiT (Touvron et al., 2021). We input the first and middle frames of each second of the video into the model and extracted the output from each hidden layer of DeiT. For audio features, we extracted the latent representations of audio data in the video using AST (Gong et al., 2021). We input audio data into the model in 1-second intervals. We extracted the output from each hidden layer in response to the input. For both the vision and audio modalities, we reduced the dimensions of the flattened stimulus features using principal component analysis (PCA) and set the number of dimensions to 1280. We calculated the PCA loadings based on the training data and applied these loadings to the test data. We confirmed that different architectures, ResNet (He et al., 2016) and MMS (Pratap et al., 2023), provided similar results for the visual and audio modalities, respectively (Figure A.5). We used the facebook/deit-base-distilled-patch16-224, MIT/ast-finetuned-audioset-10-10-0.4593, microsoft/resnet-50, and facebook/mms-1b-all models available on Huggingface for DeiT, AST, ResNet, and MMS, respectively.

3.4. Brain encoding models

To investigate how different levels of semantic contents are represented differently in the human brain, we first built brain encoding models to predict brain activity from the latent representations of language models for each semantic content independently (see Figure 1b). Because LLMs have multiple layers, separate encoding models were constructed independently for each layer.

We modeled the mapping between the stimulus features and brain responses using a linear model $Y = XW$, where Y denotes the brain activities of voxels from fMRI data, X denotes the corresponding stimulus features, and W denotes the linear weights on the features for each voxel. We estimated the model weights from training data using L2-regularized linear regression, which we subsequently applied to test data. We explored regularization parameters during training for each voxel using cross-validation procedure. Because the dataset contained nine different dramas and movies, for tuning the regularization parameters during training, we selected sessions from two to three dramas or movies as validation data and used the remaining videos as training data. We repeated this procedure across all dramas and movies. For the evaluation, we use Pearson's correlation coefficients between the predicted and measured fMRI sig-

[1]The number of parameters for Word2Vec is calculated by multiplying the embedding size (300) by the vocabulary size (3M).

nals. We computed the statistical significance using blockwise permutation testing. Specifically, to generate a null distribution, we shuffled the voxel's actual response time course before calculating Pearson's correlation between the predicted response time course and the permuted response time course. During this process, we shuffled the actual response time course in blocks of 10TRs to preserve the temporal correlation between slices. We identified voxels with scores significantly higher than those expected by chance in the null distribution. We set the threshold for statistical significance to $P < 0.05$, and corrected for multiple comparisons using the FDR procedure. We conducted all encoding analyses using the *himalaya* library[2](la Tour et al., 2022). We will make our code publicly available upon acceptance.

To examine how uniquely each semantic content explained brain activity, we constructed a brain encoding model that incorporated all the semantic features and evaluated the unique variance explained by each semantic feature using variance partitioning analysis (la Tour et al., 2022). In variance partitioning analysis, the unique variance explained is determined by subtracting the prediction performance of a model with a certain feature of interest removed from the prediction performance of the full model that includes all features simultaneously. We spatially interpreted how the five types of semantic contents were represented differently by mapping the unique variance explained onto the cerebral cortex. When the unique variance was mapped explained onto the cortex, only the voxels that were significant in the full model were displayed. To estimate the model weights, we used banded ridge regression (la Tour et al., 2022), which can optimally estimate the regularization parameter for different feature spaces. We used Llama 2 as LLMs, and used the latent representations from the layers that demonstrated the highest accuracy in cross-validation within the training data for different semantic annotations in the previous encoding model analysis. In banded ridge regression we reduced the dimensions of the features using PCA and set the number of dimensions to 1280.

3.5. Principal component analysis

For interpretation purposes, we applied PCA to the weight matrix of the encoding model (Huth et al., 2012) after banded ridge regression. In this analysis, we focused on the representation of the *Story* feature because the representation of such high-level semantic content in the brain was not quantitatively evaluated in previous studies. We used only the voxels with top 5,000 prediction performances.

To interpret the estimated PCs, we projected randomly selected 1,640 *Story* annotations onto each PC, thus acquiring PC scores for the annotation. To further interpret the PC scores, we used GPT-4 (*gpt-4-1106-preview* of the OpenAI API) to classify these annotations into five semantic attributes that were commonly present throughout the annotations. In this classification process, we allowed the annotations to be associated with multiple semantic attributes. Each PC was then evaluated according to the average scores for these five attributes. The five attributes are: attributes depicting tense or peaceful environment ("Tense confrontations or crime", "Everyday interaction or peaceful living"), attributes representing individual or collective decision-making ("Personal growth, change, or determination", "Decision-making or role of leaders"), and attributes illustrating interaction or cooperation with others ("Mutual assistance or cooperation').[3]

3.6. Comparison of different modalities

To investigate the extent to which each semantic content uniquely explained brain activity compared with visual and auditory modalities, we examined the unique variance explained for each semantic content against features extracted from visual and auditory deep neural networks (DNNs). The analysis procedure was almost the same as that described above for the unique variance explained among the five different semantic contents. However, in this analysis, we specifically focused on comparisons between a particular semantic content and other modalities.

4. Results

4.1. Comparison of different language models

We first evaluated how different levels of semantic content explained brain activity independently. Figure 2a shows prediction performance with the latent representations extracted from five language models for five distinct semantic contents. The results showed that *Speech*, *Object*, and *Story* contents predicted brain activity with higher accuracy than *Summary* and *TimePlace* contents. The larger the model, the better the prediction performance, and in particular, Llama 2 consistently achieved higher prediction performance than Word2Vec for all subjects for *Speech*, *Object*, and *Story* ($P < 0.05$, paired t-test), with a notable difference in prediction performance for high-level background *Story* content. See Figure A.1 for the results for all subjects.

Figure 2b shows prediction performance when we used the latent representations of Llama 2 for all five semantic contents simultaneously. It demonstrates that we could predict brain activity across a wide range of brain regions involved in high-level cognition, in addition to sensory areas including vision and audio.

[2]https://github.com/gallantlab/himalaya

[3]These five attributes were also defined by GPT-4 by asking GPT-4 to identify common attributes across annotations.

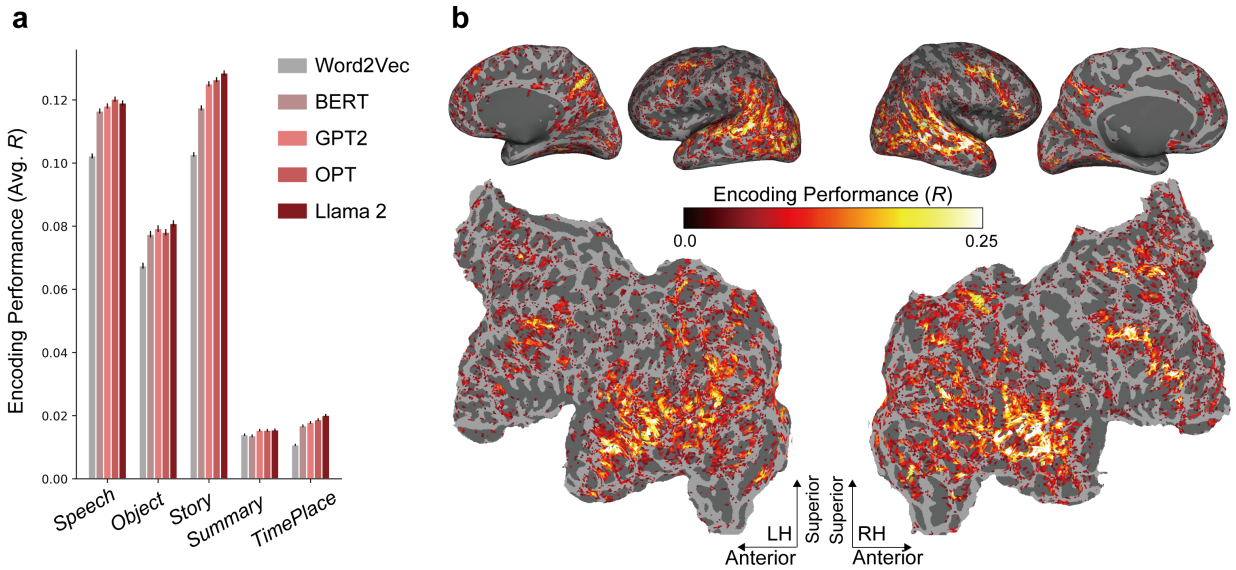

Figure 2. **Encoding model results. a** Prediction performance (measured using Pearson's correlation coefficients) when predicting the brain activity of a single participant (DM01) from the latent representations of different language models for five distinct semantic contents. The figure presents the average prediction performance on the test dataset for the top 5,000 voxels, which were selected within the training cross-validation folds in the layer exhibiting the highest prediction performance. We chose layers and voxels for each semantic content and for each language model, respectively. The error bars indicate the standard error of the mean across voxels. **b** Prediction performance for a single subject (DM01) when all five semantic contents were used simultaneously using Llama 2, projected onto the inflated (top, lateral and medial views) and flattened cortical surface (bottom, occipital areas are at the center), for both the left and right hemispheres. Brain regions with significant accuracy are colored (all colored voxels $P < 0.05$, FDR corrected).

4.2. Variance partitioning analysis

In the previous analysis, we showed that the LLMs' latent representations of *Speech*, *Object*, and *Story* contents predicted brain activity well. However, the unique contribution of each semantic content to explain brain activity remained unclear. Therefore, we used variance partitioning analysis (la Tour et al., 2022) to determine the extent to which the different types of semantic content uniquely accounted for brain activity.

We found that the latent representations of Llama 2 for semantic contents of *Speech*, *Object*, and *Story* exhibit the unique variance explained, whereas *Summary* and *TimePlace* do not (see Figure A.2). Therefore, we mapped the unique variances of these three semantic contents onto the brain surface (Figure 3), which revealed that they correspond to spatially distinct brain regions. Specifically, *Speech* is associated with the auditory cortex, *Object* with the visual cortex, and *Story* with a broader brain region, including the higher visual cortex, precuneus, and frontal cortex. Building on insights from previous studies, we present the first comprehensive results that integrate the various semantic contents into a single study. Figure A.2 shows the results for all subjects for all semantic contents.

4.3. Principal component analysis

In the previous analysis, we demonstrated that different semantic contents uniquely explain spatially distinct brain regions. Next, we analyzed what specific information were captured by our encoding models by applying PCA to the weight of the encoding model for the latent representations of Llama 2 for *Story* content.

The first three PCs explained the weight matrices of the top 5000 voxels, based on prediction performance, with explained variance ratios of $27.2 \pm 4.5\%$, $13.7 \pm 1.3\%$, and $7.6 \pm 1.6\%$ (mean$\pm$s.t.d, N=6), respectively. Figure 4a presents the projection of PC scores computed for each voxel in two example participants (DM03 and DM09). While high-level semantic content within *Story* annotations is presumed to vary significantly between individuals, certain trends were observed across participants, especially for PC1 and PC2. See Figure A.3 for the results for all subjects.

To further interpret information contents in the PCs, we projected *Story* annotations on each PC. Again, we observed consistent trends in PC1 and PC2. Figure 4b shows that PC1 contrasts annotations related to individual or collective decision-making ("Decision-making or role of leaders" and "Personal growth, change, or determination") with annotations depicting environment or background situation ("Tense

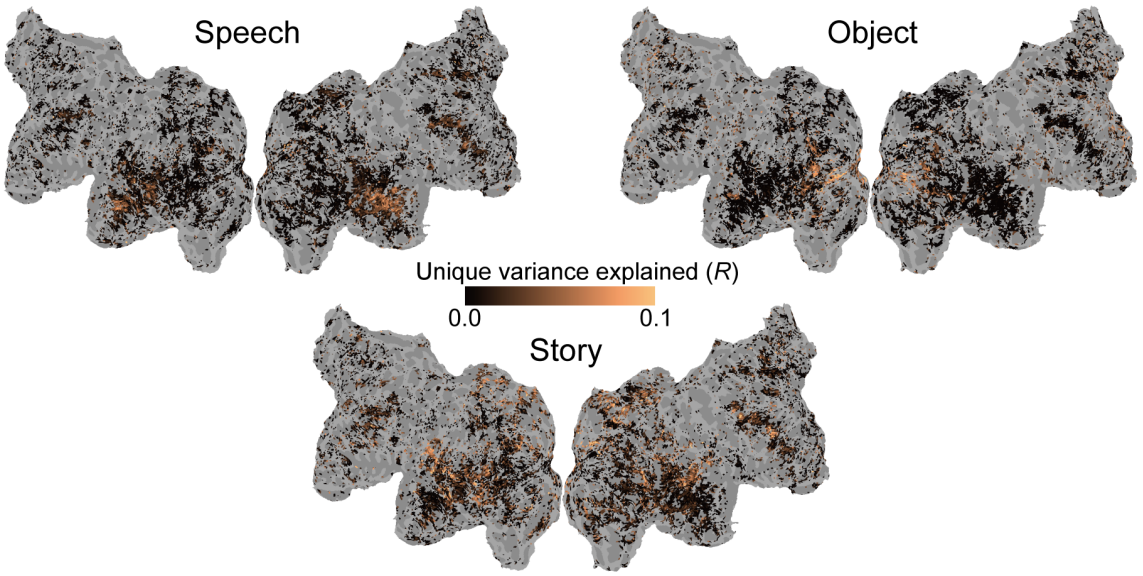

Figure 3. **Variance partitioning analysis.** Unique variance explained by the latent representations of the semantic contents of (**a**) *Speech*, (**b**) *Object*, and **c** *Story* for a single subject (DM01). We used Llama 2 as a language model. Only the voxels that were significant in the full model (Figure 2b) are colored.

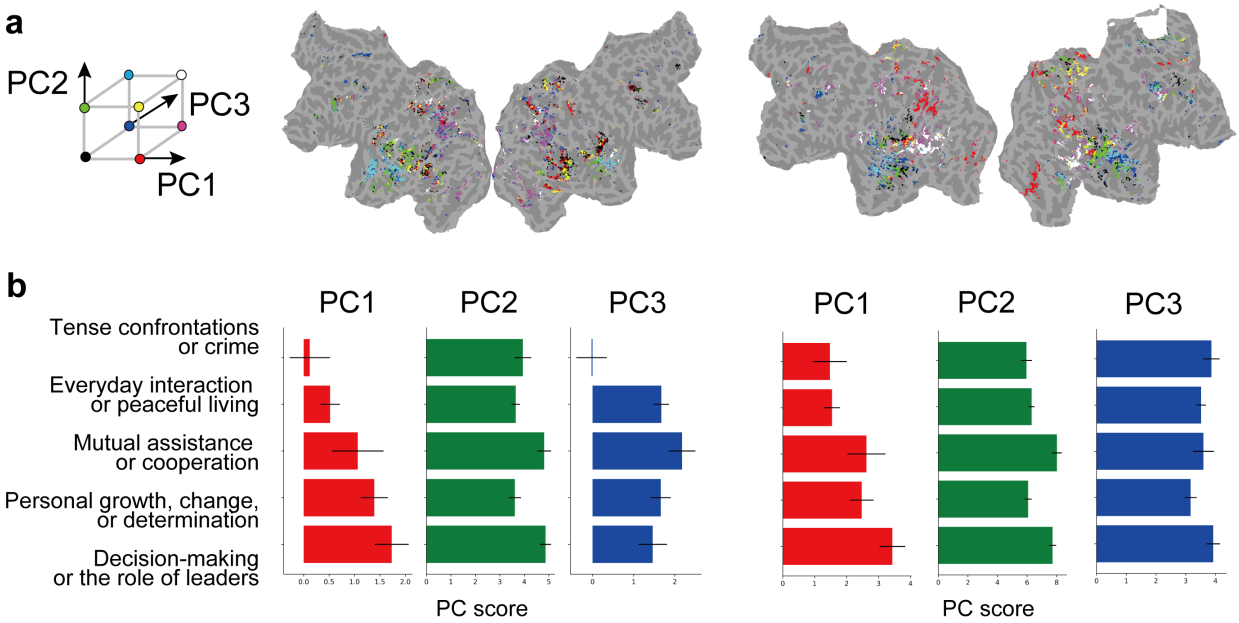

Figure 4. **Principal component analysis. a** PCA on the encoding weight matrix of the latent representations of Llama 2 identified the first three PCs for two subjects (Left, DM03; Right, DM09). Only the voxels with top 5,000 prediction performances were used for PCA. **b** We projected each caption onto the PC space and then calculated the PC scores for each attribute assigned to the caption. The error bars represent the standard error of the mean PC scores across annotations belonging to each attribute.

confrontations or crime", "Everyday interaction or peaceful living"). Regarding PC2, it contrasts annotations indicating cooperation with others ("Decision-making or role of leaders", "Mutual assistance or cooperation") with more personal situations ("Decision-making or role of leaders", "Everyday interaction or peaceful living", "Tense confronta-

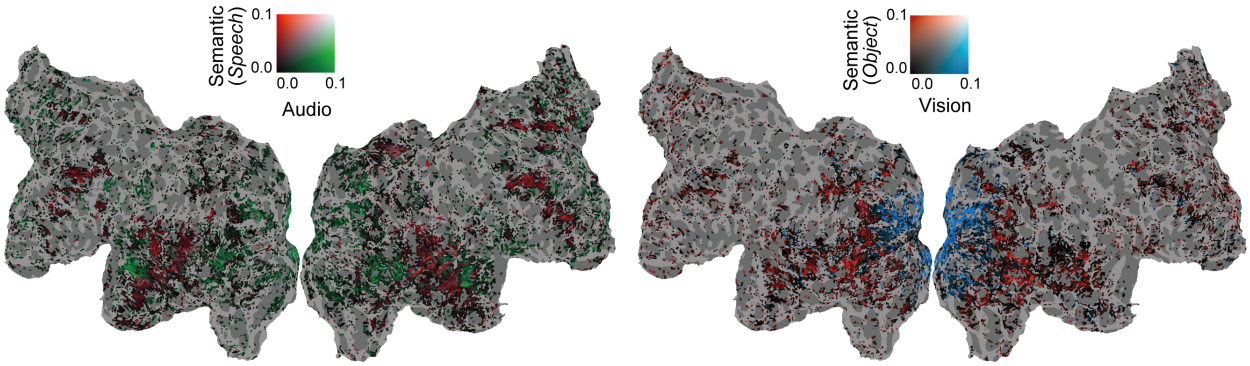

Figure 5. **Comparisons with image and audio modalities. a** Unique variance explained by the semantic contents of *Speech* compared with the audio DNN (AST (Gong et al., 2021)) for a single subject (DM01). **a** Unique variance explained by the semantic contents of *Object* compared with the image DNN (DeiT (Touvron et al., 2021)) for a single subject (DM01). Only the voxels that were significant in the full model are colored. The latent representations of Llama 2 were used for both semantic models.

tions or crime"). As for PC3, although there was a tendency among participants to contrast cooperation ("Mutual assistance or cooperation") with other attributes, the variation across participants was large.

4.4. Comparisons with image and audio modalities

Given the multimodal nature of the stimuli in the study, we were able to quantitatively compare the prediction performance of the latent representations of semantics with other modalities such as vision and audio. Particularly, the strong representation of *Speech* in the auditory region and *Object* in the lower visual region may be attributed to their correlation with audio or visual information rather than the semantic contents themselves.

Therefore, we finally looked unique variance explained by LLMs compared with strong visual (DeiT (Touvron et al., 2021)) and auditory (AST (Gong et al., 2021)) DNN models. Figure 5 shows that the latent representations of Llama 2 for *Speech* and *Object* exhibited unique variance when compared with auditory and visual DNN models, respectively. The semantic content of *Object* was located in the higher visual cortex and the vision DNNs explained the lower visual cortex. Interestingly, the semantic representation of *Speech* was situated between the anterior auditory cortex and the higher visual cortex, which was positioned between the representations related to audio DNNs. Overall, these results support our previous results: our encoding models captured semantics rather than correlated visual or auditory information. See Figure A.4 for the results for all subjects.

5. Discussion and conclusions

In this study, we quantitatively compared the relationship between the different latent representations of LLMs and brain activity, using diverse annotations of semantic contents. To achieve this, we collected a 8.3 hours of fMRI dataset of brain activity recorded while participants watched extensively annotated videos of movies or dramas, which were rich in diverse semantic contents. Our study demonstrated that the LLMs' latent representations explain brain activity particularly well for high-level background story contents compared with traditional language models. Moreover, our results indicated that the latent representations of distinct semantic contents uniquely elucidated different brain regions, beyond what is explained by features from visual and auditory DNNs.

It is crucial to reemphasize that these insights were not apparent from the modeling of individual features, as was performed in previous studies. For instance, the absence of unique variance for *Summary* and *TimePlace* is a significant insight, which indicates that simply including these types of information in encoding analysis might not be sufficient for capturing higher-level semantic representations in the brain. Our results do not indicate that such information is absent in the brain but could be caused by the limitations of modeling approaches or fMRI measurement. Hence, in future work, we need to consider how to model these types of high-level information using the data we will publish as a new biological benchmark for alignment between LLMs and humans.

Broader Impact

We attempted to uncover the neural substrates of human semantic comprehension using the latent representations of LLMs. This study not only enhances our understanding on how the human brain encodes multiple classes of semantic information but also contributes to the quantitative evaluation of how emerging Artificial Intelligence systems, such as LLMs, correspond with the human brain.

It is important to note that the method described in this paper theoretically enables the decoding of high-level story contents from brain activity. However, the method employed in this study requires human volunteers to spend a significant amount of time (over 8 hours) inside a large MRI scanner, rendering it impractical for near-future applications. While this research may not immediately pose privacy concerns, it is imperative that any future research be conducted with explicit consent from the participants, akin to our study. All participants involved in data collection provided informed consent prior to the experiments, and the research procedures received approval from the Institutional Review Board (IRB).

We invite researchers to replicate and move forward our work, such as using the modeling framework as a biological metric to determine the affinity between the brain ans LLMs. To facilitate this, we will publicly release our extensive fMRI dataset, along with the annotations for the 8.3 hours of videos used. Moreover, the source code for predicting fMRI brain activity from the latent representations of LLMs will also be made available.

Acknowledgements

YT was supported by JST, PRESTO Grant Number JPMJPR23I6. SN was supported by MEXT/JSPS KAKENHI JP18H05522 as well as JST CREST JPMJCR18A5 and ERATO JPMJER1801.

Author Contributions

S.N. and Y.T conceived the study. N.K., H.Y., R.K. and S.N. designed and performed the experiments. N.K., H.Y., Y.N., T.M., and Y.T. analyzed the data. Y.T. wrote the original draft. S.N. and Y.T. wrote the manuscript with the consultation by the other authors.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Antonello, R., Vaidya, A., and Huth, A. G. Scaling laws for language encoding models in fmri. *arXiv preprint arXiv:2305.11863*, 2023.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95 (3):709–721, 2017.

Caucheteux, C., Gramfort, A., and King, J.-R. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pp. 1336–1348. PMLR, 2021.

Caucheteux, C., Gramfort, A., and King, J.-R. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327, 2022.

Chang, L. J., Jolly, E., Cheong, J. H., Rapuano, K. M., Greenstein, N., Chen, P.-H. A., and Manning, J. R. Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Science Advances*, 7(17):eabf7129, 2021.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., and Zhou, J. H. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22710–22720, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., and Baker, C. I. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, 7:e32962, 2018.

Güçlü, U. and van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Hasson, U., Furman, O., Clark, D., Dudai, Y., and Davachi, L. Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron*, 57(3):452–462, 2008.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Horikawa, T. and Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

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

Jat, S., Tang, H., Talukdar, P., and Mitchell, T. Relating simple sentence representations in deep neural networks and the brain. *arXiv preprint arXiv:1906.11861*, 2019.

Khosla, M., Ngo, G. H., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, 2021.

la Tour, T. D., Eickenberg, M., Nunez-Elizalde, A. O., and Gallant, J. L. Feature-space selection with banded ridge regression. *NeuroImage*, pp. 119728, Nov 2022.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic resonance in medicine*, 63 (5):1144–1153, 2010.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):250, 2021.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., et al. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Schmitt, L.-M., Erb, J., Tune, S., Rysop, A. U., Hartwigsen, G., and Obleser, J. Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances*, 7(49):eabi6070, 2021.

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

Takagi, Y. and Nishimoto, S. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023.

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

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A. Appendix

A.1. fMRI dataset and preprocessing

A.1.1. STIMULI

In our study, we used multimodal stimuli from nine DVDs, which encompassed 10 episodes of television drama series and a feature film, as detailed in Table A.2. The selection of these nine DVDs adhered to specific criteria: We chose 1) internationally acclaimed dramas and films, premised on the belief that their fame ensured the compelling nature of their content. This was intended to captivate the participants' attention during the narratives in our study. 2) We included a diverse range of genres. Typically, a single story predominantly features certain characters whose dialogue and actions reflect the genre. In such limited scenarios, when deploying algorithms based on machine learning to analyze BOLD signals, ensuring the generalizability of results may prove to be challenging. To cover a broad spectrum of scenarios, we included films from multiple genres; The average total duration for the 10 episodes was 49.98 minutes (Table A.2). Each episode was segmented into two to nine segments for use in our imaging sessions. We chose the segmentation points to maintain segment lengths of approximately 10 minutes and to coincide with transitions between narrative scenes, thereby facilitating the participants' comprehension of each episode. For each segment, except the initial segment, we included the concluding 20 seconds of the preceding segment. The resulting segments varied in length from 512 seconds to 1271 seconds, with an average of 746.8 seconds (Table A.2). Instead of converting each segment into a separate film file, we designated specific playback intervals to the respective DVDs using the intervals as visual stimuli in each imaging session. All films, with the sole exception of "Ghost in the Shell" (originally produced in Japanese), were presented in the Japanese-dubbed version, considering that all participants in our study were Japanese.

A.1.2. PROCEDURES

fMRI BOLD signals were recorded as participants viewed the audiovisual content (the film segments) from 10 episodes across nine DVDs. The visual stimuli were projected centrally with a visual angle of 26.78 $\times$ 15.85 degrees at 25 or 30 Hz. MR-compatible headphones delivered the auditory stimuli. Prior to the viewing sessions, the audio levels were calibrated using non-experimental test clips to ensure clarity and a comfortable volume for participants. Participants were instructed to view the film segments casually, mirroring their everyday television-watching experience. For each participant, fMRI data were acquired over 10 distinct sessions. During each session, participants watched one or two episode segments over three to five segments (each segment lasting approximately 10 minutes) (Table A.2). Because of its extended length (about 2 hours), "Dream Girls" was viewed over two sessions. Each segment's film content was displayed by cueing the respective DVD to play at specific times using the VLC media player's (VideoLAN, France) command line interface. This interface was set up to commence playback coinciding with the scan's start. Manual termination of the scan followed the film segment's conclusion, thereby accommodating the varying durations of playback across segments and sessions. In total, around 9 hours (31905 seconds) of film content were presented to the participants across 10 sessions for fMRI data collection.

A.1.3. PREPROCESSING

For individual preprocessing of EPI data for each participant, the Statistical Parameter Mapping toolbox (SPM8, http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) was used. EPI images were motion-corrected by aligning them to the initial image recorded in the first session for each participant. Voxel responses were standardized by deducting the average response over all time points. Subsequently, prolonged trends in the standardized responses were mitigated by detracting the outcome of median filter convolution with a 120-second time frame. Data captured within the initial 20 seconds of the scan were deemed susceptible to artifacts from startup transients and thus excluded from the analysis. The analysis considered data from 20 seconds post-scan initiation to the conclusion of the film content.

Title	Created by	EAN	Volume #, Episode # (Total duration (s))	Play time for each run (hour: minute: second)
Breaking Bad	Vince Gilligan	4547462088963	Vol. 1, Ep. 1 (3479)	[run1] 00:00:00 – 00:11:19 [run2] 00:10:59 – 00:22:48 [run3] 00:22:28 – 00:34:23 [run4] 00:34:03 – 00:46:05 [run5] 00:45:45 – 00:57:59
Big Bang Theory	Chuck Lorre, Bill Prady	4548967168747	Vol. 1, Ep. 1 (1374)	[run1] 00:00:00 – 00:12:27 [run2] 00:12:07 – 00:22:54
The Crown	Peter Morgan	4547462113825	Vol. 1, Ep. 1 (3507)	[run1] 00:00:00 – 00:10:38 [run2] 00:10:18 – 00:20:43 [run3] 00:20:23 – 00:32:14 [run4] 00:31:54 – 00:45:58 [run5] 00:45:38 – 00:58:27
Heroes	Tim Kring	4988102075101	Vol. 1, Ep. 1 (3194)	[run1] 00:00:00 – 00:10:17 [run2] 00:09:57 – 00:20:36 [run3] 00:20:16 – 00:28:48 [run4] 00:28:28 – 00:41:22 [run5] 00:41:02 – 00:53:14
Suits	Aaron Korsh	4988102341596	Vol. 1, Ep. 1 (2362)	[run1] 00:00:00 – 00:12:34 [run2] 00:12:14 – 00:28:22 [run3] 00:28:02 – 00:39:22
Dream Girls	Bill Condon	4988102646882	N/A (7481)	[run1] 00:00:00 – 00:12:44 [run2] 00:12:24 – 00:26:05 [run3] 00:25:45 – 00:39:22 [run4] 00:39:02 – 00:48:54 [run5] 00:48:34 – 01:07:12 [run6] 01:06:52 – 01:18:23 [run7] 01:18:03 – 01:32:46 [run8] 01:32:26 – 01:43:50 [run9] 01:43:30 – 02:04:41
Glee	Ryan Murphy, Brad Falchuk, Ian Brennan	4988142968722	Vol. 1, Ep. 1 (2877)	[run1] 00:00:00 – 00:12:02 [run2] 00:11:42 – 00:25:14 [run3] 00:24:54 – 00:34:55 [run4] 00:34:35 – 00:47:57
The Mentalist	Bruno Heller	4548967348217	Vol. 1, Ep. 1 (2704)	[run1] 00:00:00 – 00:12:17 [run2] 00:11:57 – 00:22:04 [run3] 00:21:44 – 00:32:44 [run4] 00:32:24 – 00:45:04
Ghost in the Shell: STAND ALONE COMPLEX	Kenji Kamiyama	5022366203746	Vol. 1, Ep. 1 (1510) Vol. 1, Ep. 2 (1505)	[run1] 00:00:00 – 00:13:24 [run2] 00:13:04 – 00:25:10 [run3] 00:00:00 – 00:11:13 [run4] 00:10:53 – 00:25:05

Table A.2. **Videos used for the experiment.**

A.2. Annotation procedure

Each video was annotated for five types of semantic content by annotators employed by external agencies. The annotations were performed by one or several annotators for each type of semantic content. The overview of each annotation is as follows. The details of the annotation procedures will be more thoroughly explained in the documentation accompanying the future release of the dataset.

1. *Speech*: Describe the speech and narration content, specifying the start and end times of each segment.
2. *Object*: Describe the content displayed on the screen every second. The descriptions should be in natural language and range from one to several sentences.
3. *Story*: Approximately every five seconds, describe the story based on the content of the narrative.
4. *Summary*: Approximately every 1–3 minutes, or at points where the overall flow of the story changes, describe the content of each section and provide summaries of the actions performed by each character.
5. *TimePlace*: At each scene transition, describe the time of day and location.

Below are three example annotations from three different scenes in our dataset:

- Scene from *Suits*:
  - (*Speech*) *A novel? A grade schooler?*
  - (*Object*) *The face of a man in a suit is shown large in the center of the screen. The man looks forward, mouth closed, and appears to have a serious expression on his face.*
  - (*Story*) *Harvey expressed surprise that Mike, an elementary school student, was reading an adult novel. Mike replied that he loved books.*
  - (*Summary*) *Harvey Spector is interviewing Mike Ross. He is interested in Mike Ross when he hears how he was able to identify him as a police officer. He gives Mike Ross a question to test his knowledge. Harvey Spector notices that Mike Ross has a good memory and is smart, and decides to hire Mike Ross as an associate.*
  - (*TimePlace*) *Noon. Hotel. Room.*
- Scene from *The Crown*:
  - (*Speech*) *It's okay. It's all right.*
  - (*Object*) *A man is shown in the center looking to the left. Behind the man is a stained glass window, and several figures are vaguely visible.*
  - (*Story*) *Philip was waiting for the woman (Elizabeth) in the church with a nervous look on his face. He told himself over and over again that he would be fine.*
  - (*Summary*) *In front of the church altar, Philip looks nervous as he waits for the ceremony to begin. Former Prime Minister Winston Churchill and his wife arrive at the church, where the attendees stand to greet them.*
  - (*TimePlace*) *Morning. Chapel.*
- Scene from *Ghost in the Shell: STAND ALONE COMPLEX*:
  - (*Speech*) *Hey, what's going on?*
  - (*Object*) *Two men are wearing uniforms and have their hands clasped behind their backs, as if guarding a figure sitting in a chair. Behind them, a sharp-eyed man appears to be on the phone..*
  - (*Story*) *At air traffic control, one man with glasses is impatient and tries to figure out what's going on by whispering on the phone.*
  - (*Summary*) *A tank suddenly starts moving inside the experimental facility of Kenryo Heavy Industries. When a worker calls out to it, it stops, turns around, and attacks the worker and other tanks with a machine gun and cannon. Unrest spreads among the workers and air traffic controllers. Ohba watches the scene from a short distance outside the fence.*
  - (*TimePlace*) *Morning. Inside the Dome. Management Office.*

A.3. Additional results of encoding models

Figures A.1, A.2, A.3, and A.4 show additional results for all subjects for Figures 2, 3, 4, and 5, respectively. They show that our results were robust across subjects.

Finally, Figure A.5 shows additional results for all subjects for the unique variance explained by semantic content compared with other modalities, using other DNNs: MMS for audio and ResNet for vision. They demonstrate that our results remained robust across different DNNs.

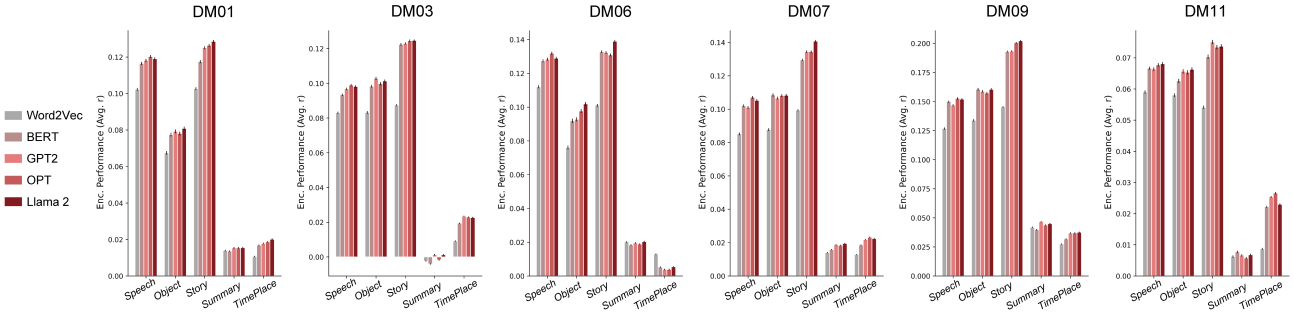

Figure A.1. All subject results for comparing prediction performances among different language models for different semantic contents.

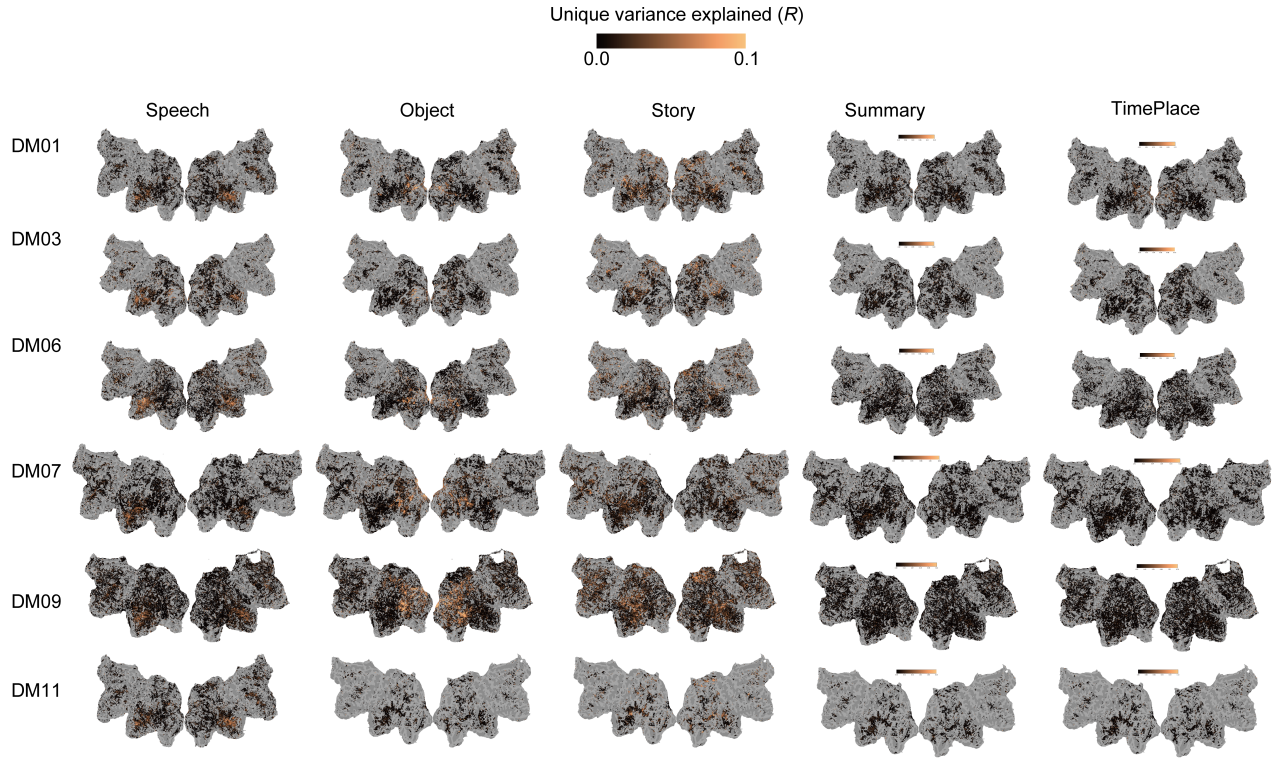

Figure A.2. All subject results for comparing unique variance explained among different semantic contents using variance partitioning analysis.

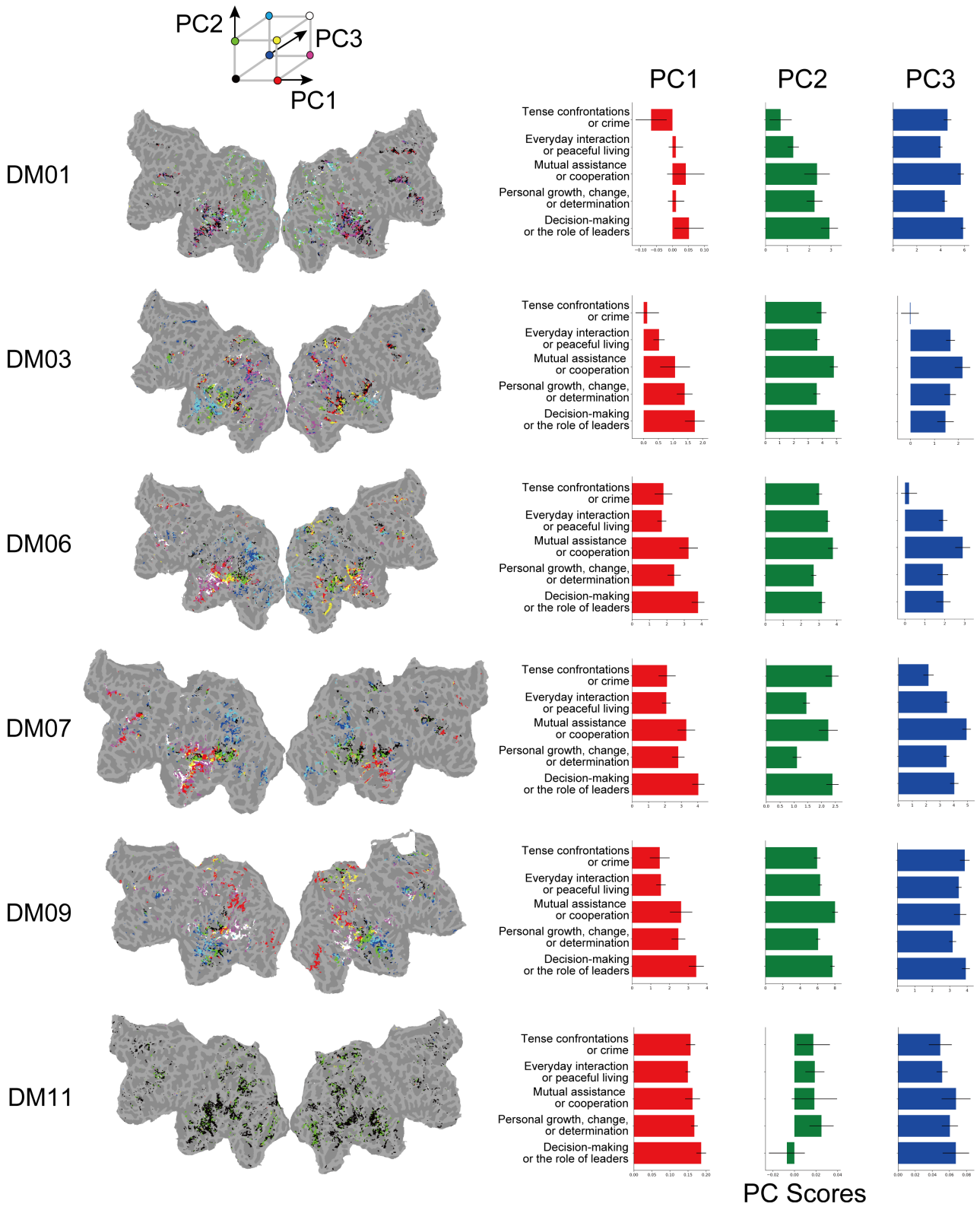

Figure A.3. All subject results for principal components analysis.

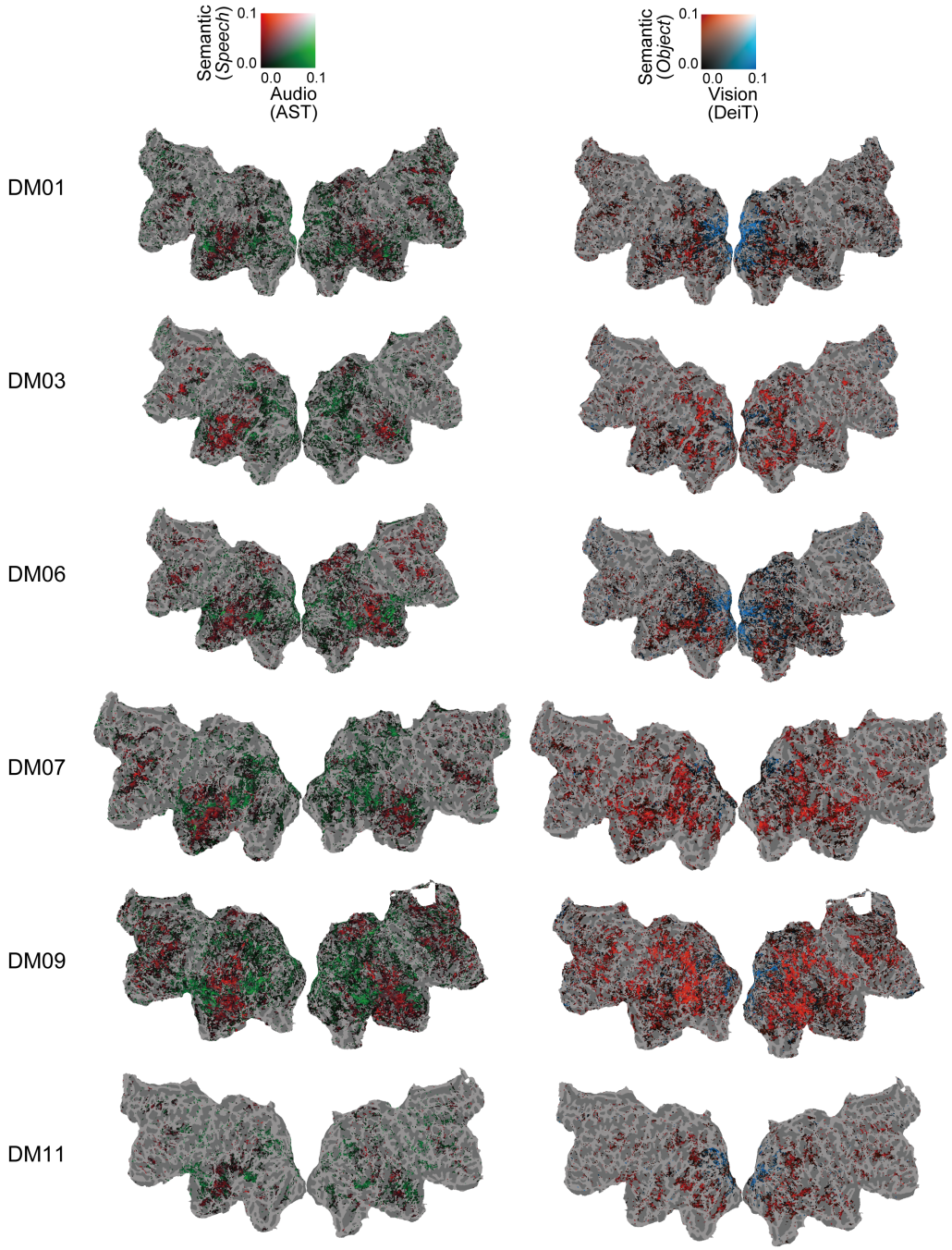

Figure A.4. All subject results for comparing the prediction performance of semantic contents with audio and visual information.

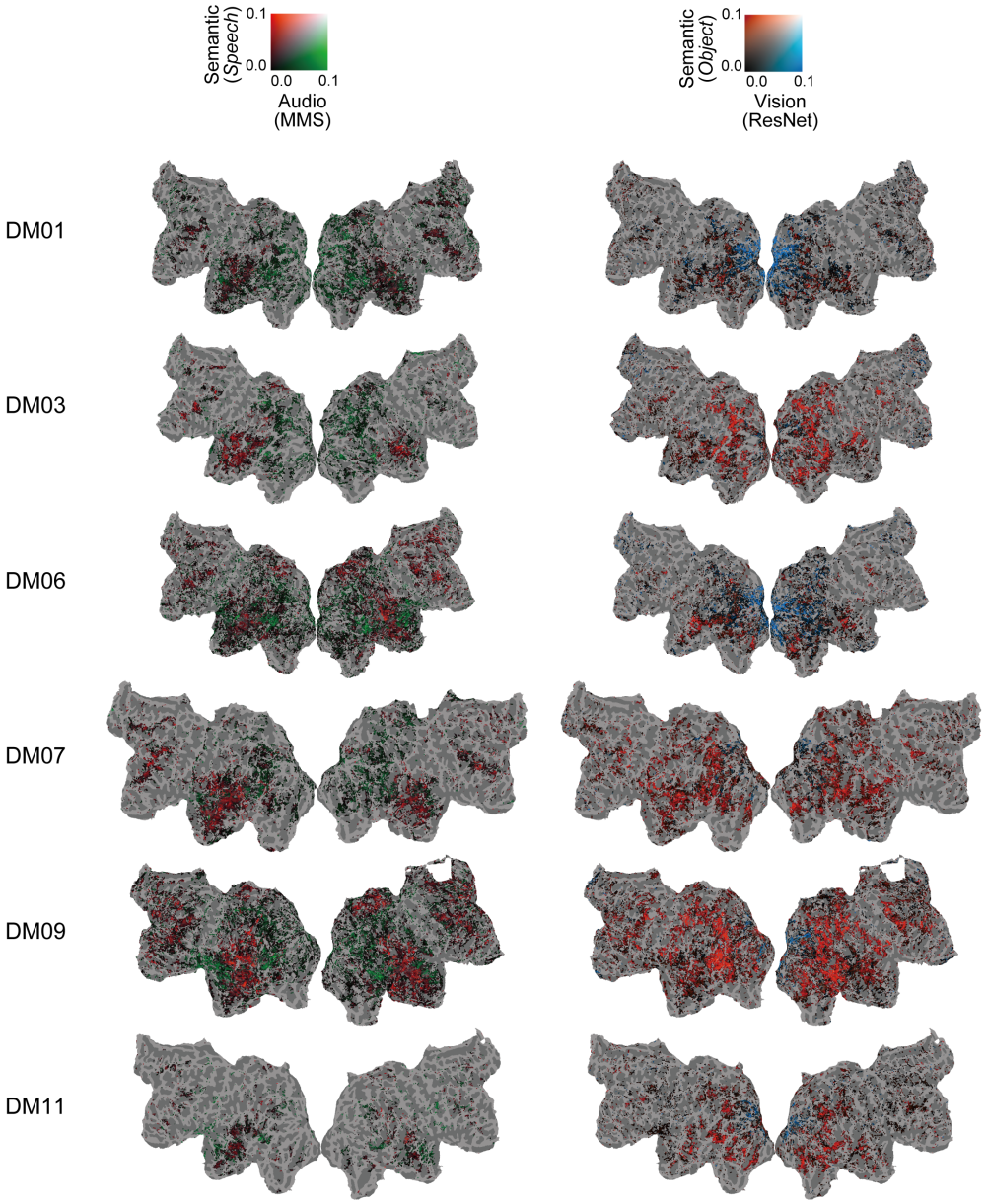

Figure A.5. All subject results for comparing prediction performance of semantic contents with audio and visual information using different DNNs:MMS for audio and ResNet for vision.