Brain–Language Model Alignment: Insights into the Platonic Hypothesis and Intermediate-Layer Advantage

Angela Lopez-Cardona[1,2] **Sebastian Idesis**[1] **Mireia Masias-Bruns**[1] **Sergi Abadal**[2]
Ioannis Arapakis[1]
[1] Telefónica Research, Barcelona, Spain
[2] Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

Do brains and language models converge toward the same internal representations of the world? Recent years have seen a rise in studies of neural activations and model alignment. In this work, we review 25 fMRI-based studies published between 2023 and 2025 and explicitly confront their findings with two key hypotheses: (i) the Platonic Representation Hypothesis—that as models scale and improve, they converge to a representation of the real world, and (ii) the Intermediate-Layer Advantage—that intermediate (mid-depth) layers often encode richer, more generalizable features. Our findings provide converging evidence that models and brains may share abstract representational structures, supporting both hypotheses and motivating further research on brain–model alignment.

1 Introduction

An emerging field at the intersection of neuroscience and Artificial Intelligence (AI) seeks to address the question: do brains and models converge toward the same internal representation of the world? It studies whether biological brains and artificial models create similar representations when exposed to the same stimuli [1], while the AI field offers new explanations for why such convergence might occur. **Language Models (LMs)** have become central to this hypothesis space, as their emerging capabilities raise questions about whether their internal representations parallel human language processing [2]. Numerous studies have shown structural similarities between brain activations and those of LMs [1]. Investigating these similarities requires a shared representational domain, typically established through partial linear mappings between features extracted from neural recordings—often via functional magnetic resonance imaging (fMRI)—and activations from computational models exposed to the same stimulus [1, 3, 4].

These findings naturally raise a deeper question: what drives this alignment? [3]. Current studies explore, for instance, how alignment varies with LMs performance [5], the relative contribution of linguistic features and perceptual features [6], and the impact of model scale, architecture, or dataset size [7]. A more recent line of research examines whether task-specific fine-tuning, multimodal integration, or human alignment data can systematically enhance brain–model correspondence [8, 9].

One theoretical perspective relevant to this question is the **Platonic Representation Hypothesis (PRH)**, introduced by Huh et al. [10]. It suggests that as neural networks scale and improve, their internal representations converge toward a shared statistical model of reality. This convergence may extend beyond artificial systems: biological systems, such as the human brain, might also share aspects of this abstraction, as both face the same fundamental challenge of efficiently extracting and understanding the underlying structure in images, text, sounds, and other modalities. Building on this idea, we seek evidence in the most recent related works on representation alignment that, if models

Preprint.

are converging toward a representation of the real world, and the brain also represents that same world, then both should converge toward each other.

In theoretical AI, increasing interest in Large Language Models (LLMs) has revealed that the final layers are not always the most informative. For example, Skean et al. [11] presented a comprehensive, layer-wise analysis across architectures, model sizes, and tasks. Their findings suggest that intermediate layers consistently outperform final layers, especially in autoregressive models. Interestingly, intermediate layers in LMs have been found to exhibit greater similarity to brain representations [12].

Bringing together these perspectives, and considering the rapid evolution of the field, we review **25 studies**, all published since 2023, that investigate similarity between brain representations and those of LMs, summarized in Section 3, and comparing our work to other reviews in Section 2. Our goal is to assess whether the evidence supports qualitatively two key hypotheses proposed in prior work: (i) the Platonic Representation Hypothesis (Section 4), and (ii) the Intermediate-Layer Advantage (Section 5). Finally, in Section 6, we summarise the main insights and discuss open questions and limitations.

2 Related work

Karamolegkou et al. [3] reviewed over 30 studies published until 2023, comparing evidence on the similarity between LM representations and brain activity. Their analysis suggests that, while the evidence remains inconclusive, correlations with model size and quality offer cautious optimism. More recently, Oota et al. [1] provided a comprehensive survey of brain-model alignment across modalities and tasks, covering encoding/decoding pipelines, datasets, and evaluation choices in depth.

In comparison, our review is intentionally more focused: we exclusively review the most recent fMRI-based studies (2023-2025), and provide more up-to-date insights. We restrict our analysis to fMRI as it is the most widely used modality in this field, enabling easier comparison across studies, and offering higher spatial resolution. From this filtering process, we derived a list of 25 works, which form the basis of our review. Rather than treating alignment as an empirical curiosity, we explicitly use it to test two theoretical emerging AI hypotheses: (i) the Platonic Representation Hypothesis, and (ii) the Intermediate-Layer Advantage. This framing shifts the focus from *"what has been observed?"* to *"do the data support these specific hypotheses?"*

3 Summary of Reviewed Works: Data, Models, and Methods

In this section, we provide an overview of the reviewed works, focusing on three key aspects: the research questions driving each study (Subsection 3.1), the datasets and models employed (Subsection 3.2), and the experimental methods used to assess similarity between brain activity and model representations (Subsection 3.3). This synthesis highlights common patterns and methodological variations across the literature. In Section 4 and Section 5, we build upon this overview to present the main conclusions and supporting evidence, mapping each work to the specific hypothesis it addresses.

3.1 Main Research Directions

Several studies have investigated brain-LM alignment to address different research questions. We categorize these works according to the specific questions they target, as indicated in Table 1.

3.2 Datasets and models

Datasets. In Table 2, we detail the datasets used and the models evaluated in each study. The choice of dataset and model varies depending on the research question. Datasets may capture brain activity during natural language processing, visual perception, auditory experiences, or a combination of the above, labeled as **R**(eading), **V**(iewing), and **L**(istening) respectively (see Table 2). Some examples interpolate these datasets with **Eye-tracking (ET)** data to trace the word-to-word transitions during reading, then sum the corresponding fMRI signal over each transition, as in St-Laurent et al. [33].

Models. Several studies rely on **text-based** Transformer [34] architectures. These include (i) encoder-only models like BERT [35], which are bidirectional; (ii) decoder-only models, such as GPT [36] and LLaMA [37, 38], and (iii) encoder–decoder models, such as BART [39] or T5 [40]. Other studies use

Table 1: Thematic categorization of reviewed works.

Theme	Representative question	Works
Information content in representations	Which linguistic/stimulus features (lexical, syntactic, semantic, stimulus-driven) drive brain–model alignment?	[5, 6, 13–16]
Scaling laws and architecture size	How do parameter count, data scale, and architectural choices affect alignment?	[7, 17–20]
Task-specific training effects	Do models trained for specific objectives (e.g., moral reasoning, speech) align better with brain data?	[21–24]
Instruction-tuning and human alignment	Does instruction tuning change the correspondence between model representations and neural activity?	[8, 17, 19, 25]
Cross-lingual and multilingual effects	Do different languages converge to a shared conceptual space in the brain?	[26]
Brain-informed tuning	Does fine-tuning on brain/behavioral signals improve neural predictivity?	[21–23, 27]
Modality differences	How do audio-based vs. text-based models compare in predicting brain signals?	[7, 28]
Multimodal vs. unimodal models	Do multimodal models predict brain activity better than unimodal ones?	[8, 9, 13, 29–32]

audio-based models, including Wav2Vec 2.0 [41] and HuBERT [42], which learn unimodal audio representations. For **vision**-only models, variants of ViT [43] are common, often compared against multimodal models such as CLIP [44], trained to align images and text.

Multimodal Large Language Models (MLLMs), which are LLMs that extend their functionality beyond textual data by training on heterogeneous datasets [4], are also increasingly used. Specially Vision Large Language Models (VLLMs) such as InstructBLIP [45] and mPLUG-Owl [46], which generally follow a LLM architecture augmented with an additional visual encoder. In both LLMs and MLLMs, versions are used both before and after undergoing **human alignment** techniques. These post-training methods aim to align model outputs with human expectations [4]. Both aligned and non-aligned variants are used to study how this post-processing affects correlation with brain representations. The category for each specific model is in Appendix A.1 (Table 4).

3.3 Methods for Brain–Model Alignment

Approaches to assessing brain–model similarity vary across studies. The majority of recent studies (22 of 25) rely on the **encoding model** [1], in contrast to earlier work that examined alternative approaches (see Karamolegkou et al. [3]). As shown in Figure 1, the encoding framework predicts brain activity (e.g., fMRI responses) directly from model representations: the same stimuli are presented to both systems, a linear mapping is trained, and its accuracy (typically measured by correlation) defines the **brain score** [7, 47]. This method underlies many recent studies [8, 18–20, 25, 32].

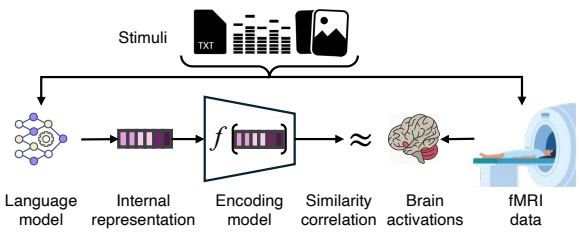

Figure 1: Encoding model framework for brain–model alignment. Model activations are linearly mapped to fMRI responses, and alignment is quantified by correlation.

One methodological variation accounts for the **temporal** nature of the neural signal, modeling the delay of the hemodynamic response. In the temporal approach, model representations are temporally aligned with brain recordings using methods such as Lanczos interpolation [12] together with a

finite impulse response (FIR) model, with multiple lags (e.g., 2–8 s). This is used in works such as [5–7, 16, 17, 21, 23, 24, 26, 27, 29–31]. The **residual approach**, proposed by Toneva et al. [48], predicts brain activity from a baseline model, derives residuals, and evaluates whether an alternative model explains variance in them, thereby isolating unique predictive power. This approach facilitates more targeted hypotheses adopted in [9, 15, 22, 28].

Other variations alter the model input instead of the **model representation** or hidden states. For instance, Gao et al. [19] leverage attention weights to capture word-to-word relations rather than isolated representations, while Rahimi et al. [16] use attribution-based features from explainable AI methods to quantify each preceding word's impact on next-word prediction. Finally, several studies replace encoding with representational similarity analysis (RSA) [49], which compares pairwise representational distances between the two systems [1, 13, 32].

4 The Platonic Representation Hypothesis

Huh et al. [10] introduced the **Platonic Representation Hypothesis**, which assumes the existence of an abstract, ideal representation space that reflects the true statistical structure of the world. Observable data, including images (X), texts (Y), sounds, etc. are regarded as projections or partial observations of this latent reality. The authors hypothesise that as models scale and improve, their internal representations converge toward this space across architectures, tasks, and modalities. To test this, they use a kernel-based alignment metric to measure whether models place similar data points close together in feature space. If two models place similar data points close together in their feature spaces, they are considered aligned. Rather than coincidence, convergence across modalities suggests that models approximate the world's underlying structure.

The authors provide theoretical and empirical evidence of convergence and propose different explanations for its occurrence. Our review centers on model scale, competence, training, and the stronger convergence observed in multimodal models. Building on this, we consider **brain–model alignment**, which the authors note only briefly as additional support for their hypothesis. Unlike Huh et al. [10], who mention this connection in passing, we directly test whether the factors proposed to enhance convergence likewise predict stronger alignment with biological brains. In other words, we reverse the logic: if the hypothesis is correct, and these factors push models toward an *ideal* representation of reality, and if biological brains share such a representation, then we should observe a positive relationship between each factor and brain-model alignment, Figure 2.

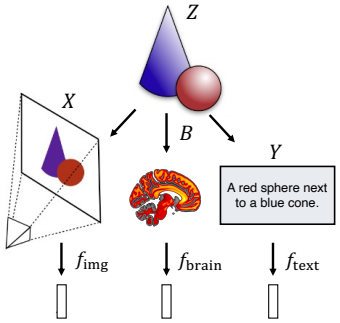

Figure 2: Platonic Representation Hypothesis (adapted from original [10]): Images (X), text (Y), and brain activity (B) are projections of a common underlying reality (Z).

Below, we describe these factors and the evidence found in the works we covered. Concretely, larger models should align better with brain data (Subsection 4.1), models trained on a broader set of tasks should do the same (Subsection 4.2), and multimodal models should align more closely than unimodal ones (Subsection 4.3).

4.1 Performance and Scaling Drives Convergence

Huh et al. [10] show that convergence increases with model competence, as higher-performing models show more similar internal structures. Deep networks have an inductive bias toward simple, statistically efficient solutions, and scaling in parameters, data, or compute amplifies this bias, reducing the space of possible representations. Larger models therefore tend to converge on shared representations that mirror the structure of the data. Although scaling alone can induce alignment, different architectures vary in how effectively they exploit it. Overall, growing scale leads models toward increasingly aligned representations that approximate a shared statistical model of the world which leads us to propose Hypothesis 1.

Hypothesis 1 *Larger and more capable models should align more strongly with brain activity.*

Table 2: Overview of datasets and models employed across reviewed studies. Rows are grouped and color-coded by modality (speech, text, speech, text, images, text, video, text, and multimodal)

Reference	Dataset	Model(s)
[21]	Passive natural language listening [50] (L)	Wav2Vec2.0 [41] and HuBERT [42]
[22]	Podcast Stories [51] (L)	Wav2Vec2.0 [41] and HuBERT [42], Whisper [52]
[28]	Subset Moth Radio Hour [53] (R)	BERT [35], GPT-2 [36], T5 Flan [54], Wav2Vec2.0 [41], Whisper [52]
[7]	Podcast Stories [51] (L)	OPT [55], LLaMA [37], HuBERT [42], WavLM [56], Whisper [52]
[26]	The Little Prince [57] (L)	Monolingual, multilingual, untrained BERT [35], Whisper [52]
[24]	Harry Potter Dataset [58] (R)	BART [39], LED [59], BigBird [60] and LongT5 [40]
[15]	Narratives [61] (L)	BERT [35], GPT2 [36]
[25]	Pereira [62] (R), BLANK2014 [63] (L), Harry Potter Dataset [58] (R)	GPT2 [36], T5 [40], LLaMa 2 [37], Vicuna, Alpaca [64], T5 Flan [54]
[5]	Harry Potter Dataset [58] (R)	GPT-2 [36]
[20]	Pereira [62] (R+V)	GPT-2 [36]
[14]	Pereira [62] (R)	GPT-2-XL [36]
[23]	Moral judgement [65]	BERT [35], DeBERTa [66](T), RoBERTa [67]
[6]	Podcast Stories [51] (L)	OPT[55], Pythia [68]
[17]	The Little Prince [57] (L)	Llama 3 [37], Gemma [69], Baichuan2 [70], DeepSeek-R1 [71], GLM [72], Qwen2.5 [73], OPT [55], Mistral [74], BERT [35]
[18]	Natural Stories fMRI [75] (L), Pereira [62] (R)	GPT-2 [36], GPT-Neo [76], OPT [55], and Pythia [68]
[27]	Moth Radio Hour [77] (R)	Monolingual (text english, chinese), multilingual BERT [35], XLM-R, XGLM, LLaMA-3.2 [38])
[16]	Narratives [61] (L)	GPT-2 [36], LLaMA 2 [37], and Phi-2 [78]
[19]	Reading Brain [79] (R)	LLaMA [38], GPT [36], Mistral [74], Alpaca [64], Gemma [69]
[29]	Sherlock clips [29] (L+V)	ViT [43], Word2Vec [80], GPT2 [36]
[8]	Natural Scenes Dataset [81] (V)	InstructBLIP [45], mPLUG-Owl [46], IDEFICS [82], ViT-H [43], and CLIP [44]
[32]	Pereira [62] (R+V)	GPT-2 [36] , Qwen-2.5 [73], Vicuna-1.5 [83], FLAVA [84], LLaVA [85], Qwen2.5-VL [86]
[30]	Moth Radio Hour [77] (R), Movie watching [87] (L+V)	BridgeTower [88], RoBERTa [67] and ViT [43]
[31]	Japanese movie [31] (L+V)	Word2Vec [80], BERT [35], GPT2 [36], OPT [55], Llama 2 [37], CLIP [44], GIT [89], BridgeTower [88], LLaVA [85]
[13]	BOLD Moments Dataset [90] (V)	ResNet-50 [91], ViViTB [92], CodeLlama-7B, Llama3-8B [38], BLIP-L [93], LLaVA-OV-7B [94]
[9]	Movie10 [33] (L+V)	ImageBind [95], TVLT [96], Wav2Vec2.0 [41], ViT-B [43], ViViTB [92], VideoMAE [97]

The review by Karamolegkou et al. [3], focusing on earlier work, reported that brain-model similarity increases modestly with model size. Building on this, we examine recent studies for additional evidence. Antonello et al. [7] investigated how model architecture, size, and training data influence the ability of LLMs to predict human brain activity, showing that the brain prediction performance scales logarithmically with model size. Their results further indicate that increases in training data volume and downstream task performance correlate with improved neural alignment, whereas increasing hidden state size without corresponding performance gains can actually degrade encoding quality. Consistent with this trend, both Lei et al. [17] and Gao et al. [19] found that larger models exhibit better alignment with brain responses.

Challenging this view, Lin et al. [18] tested whether brain alignment reflects true linguistic learning or simply the artifact of larger vector dimensionality. Comparing trained models to untrained but dimension and architecture-matched counterparts, and modelling only residual variance, they found that controlling for dimensionality diminishes or even reverses the apparent scaling benefit: larger models do not align better, and the unique contribution of training may even decrease with size. In contrast, Oota et al. [15], through residual analysis (Subsection 3.3), showed that specific linguistic properties make a genuine contribution to brain alignment.

In parallel, Merlin and Toneva [5] demonstrated that brain alignment cannot be explained by next-word prediction performance alone. Even when controlling for word-level information and prediction accuracy, residual alignment persists in language regions, suggesting that models capture additional properties relevant to brain responses. Complementing these perspectives, Antonello and Cheng [6] provided evidence that alignment increases with training through a two-phase abstraction process, in which intermediate layers construct higher-dimensional, compositional representations. Similarly, Hosseini et al. [20] showed that models trained on developmentally realistic data volumes (100M words) already achieve near-maximal alignment, comparable to models trained on billions of words, with further gains in prediction accuracy failing to enhance alignment. Taken together, these findings support Hypothesis 1: Although alignment tends to increase with model scale and performance, neither scaling nor predictive coding performance fully account for it, suggesting that alignment relies on richer representational mechanisms beyond word-level prediction.

4.2 Task Expansion Drives Convergence

Another dimension of convergence highlighted by Huh et al. [10] is task diversity. Training on a broader set of tasks forces models to find representations that satisfy multiple objectives. As task diversity increases, the space of viable representations shrinks, pushing models toward shared, general-purpose representations, since every new task or dataset adds constraints. The paper visualises this as the intersection of constraint regions in representation space. Based on this, we propose Hypothesis 2

Hypothesis 2 *Models trained on a broader set of tasks should align more strongly with brain activity.*

Aw and Toneva [24] demonstrated that fine-tuning on a narrative summarization task, BookSumf [98], produces richer and more brain-like representations than training models exclusively on generic web data, despite not improving language modeling performance.

Another line of work focused on **instruction fine-tuning**, where models are trained on many heterogeneous tasks phrased as natural language instructions (e.g., summarization, Question Answering (QA), classification). Such training is thought to push models closer to an idealised Platonic representation, by optimizing under diverse constraints. Aw et al. [25] evaluated 25 models and reported that instruction-tuned models improve alignment by an average of 6%. Consistently, Lei et al. [17] found that instruction-tuned models outperform their base counterparts, whereas Gao et al. [19] showed that, when controlling for size, instruction-tuning yields no significant benefit.

An increasing number of studies have explored **brain-tuning**, i.e., fine-tuning models with fMRI data. Conceptually, this adds an additional constraint, forcing models to align not only with external objectives (e.g., language modelling, image classification) but also with biological representations. Within the framework of the PRH, such constraints further narrow the representational space, pushing models toward a shared, modality-independent structure. For example, Meek et al. [23] tested whether fine-tuning encoder-based LLMs on moral reasoning data or directly on fMRI recordings improves neural alignment, using the ETHICS benchmark [99] and the Moral Judgments dataset [65]. Their results indicate that neither strategy consistently improves brain alignment or task performance, suggesting that targeted fine-tuning alone may be insufficient.

Regarding text and audio models, Oota et al. [28] investigated alignment during reading and listening, emphasizing the contribution of low-level features. Their findings indicate that text-based models achieve stronger alignment overall, particularly in late language regions where alignment reflects semantic rather than stimulus-driven processing. Moreover, text-based models exhibited superior cross-modal transfer (e.g., to visual and auditory regions), implying that they encode richer and more generalisable representations than the speech-based ones. Building on this work, Moussa et al. [22] applied brain-tuning to speech models, improving alignment with semantic regions, reducing reliance on low-level features, and enhancing downstream semantic performance without impairing speech abilities. Moussa and Toneva [21] confirmed and extended these results, showing that brain-tuning

reorganizes representations into a clearer progression from acoustics to semantics. Together, these findings suggest that while text models initially align more closely with brain processing and may approximate the modality-independent space proposed by the PRH, brain-tuning moves speech models closer to this representation.

Finally, Negi et al. [27] showed that multilingual LLMs outperform monolingual ones in brain alignment and cross-lingual transfer, even without fine-tuning. Fine-tuning on brain data yields only small, inconsistent gains, suggesting that the main advantage stems from multilingual pre-training rather than brain-based fine-tuning.

4.3 Cross-Modal Training Drives Convergence

The authors show that better LMs tend to align more strongly with vision models like DINOv2 [100], and that multimodally trained models such as CLIP exhibit higher vision–language alignment, which drops when fine-tuned on a single modality. Cross-modal training encourages models to learn representations that are not tied to a single modality but capture a shared, abstract structure—bringing them closer to a *Platonic* representation. Using diverse data types, such as image–text pairs, constrains models to discover representations valid across modalities. This motivates the following Hypothesis 3.

Hypothesis 3 *Models trained on more modalities should align more strongly with brain activity.*

Most prior studies have compared **VLLMs** with LLMs. Oota et al. [8] showed that instruction-tuned MLLMs align slightly better with brain activity than CLIP, with both outperforming vision-only models, thus supporting the benefit of cross-modal integration. Similarly, Tang et al. [30] found that multimodal transformers generalise across modalities (e.g., trained on movies and applied to speech), with multimodal features showing stronger alignment in higher-level cortical regions. Small et al. [29] further reported that multimodal embeddings outperform unimodal ones in language and social brain regions but not in visual areas, suggesting non-uniform convergence. Along these lines, Nakagi et al. [31] showed that multimodal vision–semantic models better explain high-level narrative regions, with PCA isolating variance linked to semantic features such as background story. Finally, Ryskina et al. [32] confirmed the advantage of VLLMs for cross-modal conceptual meaning and introduced novel conceptual ROIs-voxels that respond consistently to the same concepts across modalities, best predicted by both LLMs and MLLMs. This supports the view that models capture modality-independent conceptual information, approximating a *real world* representation.

Zada et al. [26] provided evidence from languages and **audio**, showing with fMRI from monolingual speakers that unilingual BERT embeddings are similar but rotated, aligning more closely for related languages and predicting comprehension activity across languages with minimal loss. In contrast, multilingual and multimodal models captured more abstract, language-independent concepts: their mid-layers are less language-specific, and alignment is stronger for languages closer to the native tongue. These results suggest that broader linguistic and modality exposure yields richer conceptual spaces that better align with the brain. Extending beyond language and audio, Han et al. [13] introduced **video** as a modality and reported that image–language and video–LMs show stronger alignment in higher-level brain regions, with predictivity rising from early to later layers, especially when models integrate predictive processing across modalities. Similarly, Oota et al. [9] showed that video–audio models outperform unimodal ones across language, visual, and auditory regions. Importantly, this effect extends beyond low-level sensory features: cross-modal training enhances alignment in language areas, reinforcing the view that multimodal exposure yields more brain-like, world-grounded representations. Overall, these findings support Hypothesis 3, indicating that cross-modal training pushes models toward modality-independent conceptual representations.

5 The Intermediate Layer Advantage Hypothesis

Skean et al. [11] build on prior work showing that linguistic and semantic features often emerge in middle layers, while final layers become overly tuned to pretraining in LMs. On the one hand, they introduce a unified framework—combining information-theoretic, geometric, and invariance-based metrics (e.g., DiME [101], curvature [102], InfoNCE [103])—to evaluate layer-wise representation quality across architectures, model sizes, and tasks. On the other hand, their empirical results show that intermediate layers consistently outperform final ones on 32 MTEB benchmark tasks [104], sometimes by up to 16%, a trend observed in both Transformers (Pythia [68], Llama3 [38], BERT [35])

and State Space Models (SSMs) (Mamba [105]). Importantly, their metrics not only correlate strongly with downstream performance but also peak at the same intermediate layers, thereby providing both an empirical and theoretical demonstration that these layers yield the most robust and generalizable representations. They also identify architecture-specific patterns: autoregressive decoders show a pronounced mid-layer compression valley, while bidirectional encoders remain more uniform. Extending to vision, only autoregressive image transformers, like AIM [106], display the same mid-depth bottleneck, suggesting the training objective, not the modality, drives this effect.

Hypothesis 4 *If intermediate layers of LMs encode the most robust and generalizable linguistic and semantic features, then these layers should also show the strongest alignment with brain activity.*

Toneva and Wehbe [12] first demonstrated that middle layers of LMs show the strongest alignment with brain language regions, a finding repeatedly replicated in subsequent work. This observation aligns with recent evidence that middle layers encode richer and more generalizable linguistic representations [11]. Here, we assess whether the studies under review provide additional support for this pattern of layer-wise convergence across model classes.

Several studies focused on **decoder**-based **text** LMs. Of particular relevance, Antonello and Cheng [6] examined a question directly related to the hypothesis in Skean et al. [11]. They found evidence that LMs undergo a two-phase abstraction process during training, an early composition phase, and a later prediction phase, reflected in how well the model layers align with brain activity. Results show that layers with higher intrinsic dimensionality exhibit stronger brain alignment and that such brain-aligned representations emerge predominantly in the middle layers. Moreover, as training advances, the composition phase becomes compressed into fewer layers, suggesting that training simultaneously improves task performance and sharpens the emergence of cognitively relevant representations.

Similar evidence is described in Antonello et al. [7], comparing LLaMA and OPT. The LLaMA models are marginally better at encoding than the OPT models and reach peak performance in relatively early layers followed by a slow decay. In contrast, OPT models achieve their maximum performance in layers that are roughly three-quarters into the model, which mirrors results observed in other decoder models. They propose the larger training set of the LLaMA models as an explanation for both their superior encoding performance and their different layer-wise pattern. Other studies (e.g., Lei et al. [17], Kauf et al. [14]) find that many models also peak in intermediate layers.

Complementing this, Rahimi et al. [16] reached a similar conclusion, focusing on importance rather than representations (i.e., how much each word in context contributes to next-word prediction). Early layers correspond to initial stages of language processing in the brain, while later layers align with more advanced stages. Layers with higher attribution scores, i.e., more influential for prediction, also show stronger alignment with neural activity.

In **encoder–decoder** text models, Aw and Toneva [24] evaluated four architectures and found that improvements emerge in intermediate layers or are distributed across depth before and after fine-tuning, but never peak in the final layer. Similarly, Oota et al. [28] showed that text-based models follow a clearer progression from low- to high-level representations: early layers encode superficial textual features (e.g., number of letters or words), which diminish in deeper layers, where alignment with later-stage brain regions strengthens. For these models, alignment peaks in mid-to-late layers, consistent with the emergence of abstract, semantically relevant representations.

Additionally, several studies investigated **encoder**-based text model architectures. Specifically, Oota et al. [15] found that in all cases performance is strongest in the middle layers. Similarly, Zada et al. [26] analysed monolingual and multilingual models—for monolingual models, correlations between unimodal embeddings increased through the late-intermediate layers and dropped in the final layer, while for multilingual models, cross-language correlations grew until roughly three-quarters depth before declining. These findings suggest that the first and last layers are associated with language-specific processes, whereas intermediate layers capture more conceptual representations.

For **audio** models, Oota et al. [28] analysed layer-feature correlations and found that speech-based models follow a different pattern from text-based ones. Low-level features such as Mel spectrograms, and phonological information are strongly encoded in the early and intermediate layers and persist even in deeper layers. These features drive alignment with early auditory cortices, but do not support robust alignment with late language regions once removed, indicating that speech models retain a sensory-phonological focus and lack abstract semantic representations in deeper layers. Building on this, Moussa and Toneva [21] showed that brain-tuning reshapes this hierarchy: early layers remain

acoustic, while late layers align strongly with higher language regions and capture complex semantic information, closely mirroring the brain's progression from acoustics to semantics.

In a different vein, Antonello et al. [7] reported that upper-middle and uppermost layers generally yield the best performance, except in Whisper, where performance increases steadily with depth, likely due to the use of only the encoder. Complementing this, Zada et al. [26] showed that when both Whisper modules are used, correlations between embeddings of different languages peak in the encoder's final layers and in the decoder's mid-to-late layers.

Comparing text and **vision** models, Han et al. [13] observed that text models show weak alignment in early layers but stronger alignment in middle and final layers, mainly in language and higher cognitive regions. In contrast, vision models align strongly with early and mid-visual areas but alignment diminishes at later stages. Multimodal models demonstrate broad alignment from the outset, sustain it across middle layers, and in final layers LLaVA (a predictive multimodal model) achieves strongest overall alignment, while BLIP-L shows reduced alignment in vision. Training objective plays a significant role: classification and captioning models align better at early layers but weaken at depth, whereas predictive models improve in middle and late layers, maintaining or enhancing alignment in higher cognitive regions, especially under multimodality. Building on these findings, Tang et al. [30] reported that BridgeTower achieves peak accuracy in intermediate layers associated with cortical conceptual regions, while Oota et al. [8] showed that InstructBLIP and IDEFICS align with higher visual regions in middle layers and early visual regions in later layers, whereas mPLUG-Owl reaches maximal alignment in late layers across both high- and low-level regions.

Finally, Oota et al. [9] extended this comparison to **video** models, reporting consistent declines in performance from early to deeper layers, for both multimodal and unimodal models. Their key finding is that joint video-audio embeddings achieve superior brain alignment across all layers relative to unimodal video or speech embeddings.

6 Conclusions

This field has gained momentum, as reflected in the surge of studies over the past two years. Multiple factors have been examined as potential drivers of brain-language model alignment, with growing attention to multimodal models, particularly VLLMs. According to the PRH, such alignment arises because improving models approximate the same underlying representation of the world that the brain interprets. The studies we reviewed generally support this view: alignment tends to be stronger for larger, higher-performing models and those trained on more tasks (Subsection 4.1). Fine-tuning likewise improves alignment, consistent with the claim that adding data, tasks, or constraints drives representational convergence. Evidence that brain-tuning or human alignment brings models closer to this representation, however, appears weaker (Subsection 4.2). Cross-modal training, by contrast, reliably enhances alignment by pushing models toward modality-independent conceptual representations. Overall, higher-performing models align more broadly across the brain, including in regions not tied to their training modality. By testing evidence from the original hypothesis against an additional line suggested by Huh et al. [10], brain–model alignment, we find converging support for the existence of such an underlying representation, though the evidence is not uniformly positive.

At the same time, evidence shows that final layers are not those that align most strongly with the brain (Section 5), in line with Hypothesis 4. Across architectures and brain regions, studies vary, but a general trend emerges: intermediate layers yield better alignment with relevant regions. This too resonates with the PRH: intermediate layers appear to encode representations closer to the world's underlying structure, while final layers are increasingly specialised to pre-training objectives and less universally aligned. Finally, Table 3 provides a summary of the qualitative analysis of how the reviewed works support (or contradict) the hypotheses proposed in this review. We considered a study to show agreement or disagreement when its findings could be qualitatively related to a given hypothesis. Cases labelled as neutral correspond to studies whose results could not be clearly interpreted as either supporting or contradicting the hypothesis. Framing alignment through the dual lenses of the PRH and the Intermediate-Layer Advantage provides a theoretical scaffold for designing future experiments, benchmarks, and models that more directly probe shared human–artificial representational structure.

Limitations. The studies we analysed differ in metrics, datasets, and alignment protocols, as also noted by Karamolegkou et al. [3]. Such heterogeneity implies that effects of scale or layer depth

Table 3: Overview of the reviewed studies classified by modality (speech, text, speech, text, images, text, video, text, multimodal). This classification follows the same organisational structure as Table 2 to facilitate direct comparison. Each column represents one of the four hypotheses introduced in Section 4 and Section 5. Cell colours convey the qualitative degree of support: strong disagreement, disagreement, neutral, agreement, and strong agreement.

Reference	Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 4
[21]				
[22]				
[28]				
[7]				
[26]				
[24]				
[15]				
[25]				
[5]				
[20]				
[14]				
[23]				
[6]				
[17]				
[18]				
[27]				
[16]				
[19]				
[29]				
[8]				
[32]				
[30]				
[31]				
[13]				
[9]				

should be treated as qualitative patterns rather than quantitative laws. More broadly, interpretation is complicated by choices of stimuli, resolution, recorded brain regions, and neural preprocessing, all of which capture different aspects of cognition. Architectural details (e.g., tokenization in language, visual preprocessing in vision) may further interact with the data. Our review emphasises general trends across 25 language-related studies without attempting to parse finer-grained regional or processing differences. Vision-specific work, such as Gifford et al. [107], thus lies outside our scope.

Acknowledgments and Disclosure of Funding

This research is supported by Horizon Europe's European Innovation Council through the Pathfinder program (SYMBIOTIK grant 101071147) and by the Industrial Doctorate Plan of the Department of Research and Universities of the Generalitat de Catalunya Grant AGAUR 2023 DI060.

References

[1] Subba Reddy Oota, Zijiao Chen, Manish Gupta, Bapi Raju Surampudi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey). *Transactions on Machine Learning Research*, July 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=YxKJihRcby.

[2] Yuhong Zhang, Shilai Yang, Gert Cauwenberghs, and Tzyy-Ping Jung. From Word Embedding to Reading Embedding Using Large Language Model, EEG and Eye-tracking. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, July 2024. doi: 10.1109/EMBC53108.2024.10781627. URL https://ieeexplore.ieee.org/document/10781627. ISSN: 2694-0604.

[3] Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping Brains with Language Models: A Survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9748–9762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.618. URL https://aclanthology.org/2023.findings-acl.618.

[4] Angela Lopez-Cardona, Sebastian Idesis, and Ioannis Arapakis. Integrating Cognitive Processing Signals into Language Models: A Review of Advances, Applications and Future Directions, April 2025. URL http://arxiv.org/abs/2504.06843. arXiv:2504.06843 [cs].

[5] Gabriele Merlin and Mariya Toneva. Language models and brains align due to more than next-word prediction and word-level information. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18431–18454, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1024. URL https://aclanthology.org/2024.emnlp-main.1024/.

[6] Richard Antonello and Emily Cheng. Evidence from fMRI Supports a Two-Phase Abstraction Process in Language Models. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, October 2024. URL https://openreview.net/forum?id=VZipjFlBpl#discussion.

[7] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, 36:21895–21907, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4533e4a352440a32558c1c227602c323-Abstract-Conference.html.

[8] Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi Gnvv, Manish Shrivastava, Maneesh Kumar Singh, Bapi Raju Surampudi, and Manish Gupta. Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain). In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=xkgfLXZ4e0.

[9] Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Maneesh Kumar Singh, Manish Gupta, and Bapi Raju Surampudi. Multi-modal brain encoding models for multi-modal stimuli. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=0dELcFHig2.

[10] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: the platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 20617–20642, Vienna, Austria, July 2024. JMLR.org.

[11] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by Layer: Uncovering Hidden Representations in Language Models. In *Forty-second International Conference on Machine Learning*, June 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

[12] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html.

[13] Hyewon Willow Han, Ruchira Dhar, Qingqing Yang, Maryam Hoseini Behbahani, María Alejandra Martínez Ortiz, Tolulope Samuel Oladele, Diana C. Dima, Hsin-Hung Li, Anders

Søgaard, and Yalda Mohsenzadeh. Investigating the role of modality and training objective on representational alignment between transformers and the brain. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, October 2024. URL https://openreview.net/forum?id=t4CnKu6yXn#discussion.

[14] Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *bioRxiv: The Preprint Server for Biology*, page 2023.05.05.539646, May 2023. ISSN 2692-8205. doi: 10.1101/2023.05.05.539646.

[15] Subbareddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36: 18001–18014, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3a0e2de215bd17c39ad08ba1d16c1b12-Abstract-Conference.html.

[16] Maryam Rahimi, Yadollah Yaghoobzadeh, and Mohammad Reza Daliri. Explanations of Deep Language Models Explain Language Representations in the Brain, February 2025. URL http://arxiv.org/abs/2502.14671. Issue: arXiv:2502.14671 arXiv:2502.14671 [cs].

[17] Yu Lei, Xingyang Ge, Yi Zhang, Yiming Yang, and Bolei Ma. Do Large Language Models Think Like the Brain? Sentence-Level Evidence from fMRI and Hierarchical Embeddings, May 2025. URL http://arxiv.org/abs/2505.22563. arXiv:2505.22563 [cs].

[18] Yi-Chien Lin, Hongao Zhu, and William Schuler. Vectors from Larger Language Models Predict Human Reading Time and fMRI Data More Poorly when Dimensionality Expansion is Controlled, May 2025. URL http://arxiv.org/abs/2505.12196. arXiv:2505.12196 [cs].

[19] Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain. *bioRxiv*, April 2025. doi: 10.1101/2024.08.15.608196.

[20] Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. *Neurobiology of Language*, 5(1):43–63, April 2024. ISSN 2641-4368. doi: 10.1162/nol_a_00137. URL https://doi.org/10.1162/nol_a_00137.

[21] Omer Moussa and Mariya Toneva. Brain-tuned Speech Models Better Reflect Speech Processing Stages in the Brain, June 2025. URL http://arxiv.org/abs/2506.03832. arXiv:2506.03832 [cs].

[22] Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving Semantic Understanding in Speech Language Models via Brain-tuning. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=KL8Sm4xRn7.

[23] Austin Meek, Artem Karpov, Seong Hah Cho, Raymond Koopmanschap, Lucy Farnik, and Bogdan-Ionut Cirstea. Inducing Human-like Biases in Moral Reasoning Language Models. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, October 2024. URL https://openreview.net/forum?id=OuIGwpTQic#discussion.

[24] Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, February 2023. URL https://openreview.net/forum?id=KzkLAE49H9b.

[25] Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning Aligns LLMs to the Human Brain. In *First Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=nXNN0x4wbl#discussion.

[26] Zaid Zada, Samuel A. Nastase, Jixing Li, and Uri Hasson. Brains and language models converge on a shared conceptual space across different languages, June 2025. URL http://arxiv.org/abs/2506.20489. arXiv:2506.20489 [q-bio].

[27] Anuja Negi, Subba Reddy Oota, Manish Gupta, and Fatma Deniz. Brain-Informed Fine-Tuning for Improved Multilingual Understanding in Language Models, July 2025. URL https://www.biorxiv.org/content/10.1101/2025.07.07.662360v1. ISSN: 2692-8205 Pages: 2025.07.07.662360 Section: New Results.

[28] Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8528, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.462. URL https://aclanthology.org/2024.acl-long.462/.

[29] Hannah Small, Haemy Lee Masson, Stewart Mostofsky, and Leyla Isik. Vision and language representations in multimodal AI models and human social brain regions during natural movie viewing. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, October 2024. URL https://openreview.net/forum?id=pS1UjuYuJu#discussion.

[30] Jerry Tang, Meng Du, Vy A. Vo, Vasudev Lal, and Alexander G. Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 29654–29666, Red Hook, NY, USA, December 2023. Curran Associates Inc.

[31] Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling Multi-level and Multi-modal Semantic Representations in the Human Brain using Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20313–20338, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.1133.

[32] Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, and Evelina Fedorenko. Language models align with brain regions that represent concepts across modalities, August 2025. URL http://arxiv.org/abs/2508.11536. arXiv:2508.11536 [cs].

[33] Marie St-Laurent, Basile Pinsard, Oliver Contier, Katja Seeliger, Valentina Borghesani, Julie Boyle, Pierre Bellec, and Martin Hebart. cneuromod-things : a large-scale fMRI dataset for task- and data-driven assessment of object representation and visual memory recognition in the human brain. *Journal of Vision*, 23(9):5424, August 2023. ISSN 1534-7362. doi: 10.1167/jov.23.9.5424. URL https://doi.org/10.1167/jov.23.9.5424.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI*, 2019.

[37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

[38] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and et al. The Llama 3 Herd of Models, August 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[39] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1):140:5485–140:5551, January 2020. ISSN 1532-4435.

[41] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 12449–12460, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

[42] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. ISSN 2329-9304. doi: 10.1109/TASLP.2021.3122291. URL https://ieeexplore.ieee.org/document/9585401.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html. ISSN: 2640-3498.

[45] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International*

Conference on Neural Information Processing Systems, NIPS '23, pages 49250–49267, Red Hook, NY, USA, May 2024. Curran Associates Inc.

[46] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, March 2024. URL http://arxiv.org/abs/2304.14178. arXiv:2304.14178.

[47] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/full/10.1073/pnas.2105646118. Publisher: Proceedings of the National Academy of Sciences.

[48] Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2 (11):745–757, November 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00354-6. URL https://doi.org/10.1038/s43588-022-00354-6.

[49] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, November 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full. Publisher: Frontiers.

[50] Amanda LeBel, Shailee Jain, and Alexander G. Huth. Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual. *Journal of Neuroscience*, 41 (50):10341–10355, December 2021. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0118-21.2021. URL https://www.jneurosci.org/content/41/50/10341. Publisher: Society for Neuroscience Section: Research Articles.

[51] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, August 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02437-z. URL https://doi.org/10.1038/s41597-023-02437-z.

[52] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 28492–28518, Honolulu, Hawaii, USA, July 2023. JMLR.org.

[53] Fatma Deniz, Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *Journal of Neuroscience*, 39(39):7722–7736, September 2019. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0675-19.2019. URL https://www.jneurosci.org/content/39/39/7722. Publisher: Society for Neuroscience Section: Research Articles.

[54] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, December 2022. URL http://arxiv.org/abs/2210.11416. arXiv:2210.11416 [cs].

[55] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and

Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, June 2022. URL http://arxiv.org/abs/2205.01068. arXiv:2205.01068 [cs].

[56] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022. ISSN 1932-4553, 1941-0484. doi: 10.1109/JSTSP.2022.3188113. URL http://arxiv.org/abs/2110.13900. arXiv:2110.13900 [cs].

[57] Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data*, 9(1):530, August 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01625-7. URL https://www.nature.com/articles/s41597-022-01625-7. Publisher: Nature Publishing Group.

[58] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, 9(11):e112575, November 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112575. Publisher: Public Library of Science.

[59] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. URL http://arxiv.org/abs/2004.05150. arXiv:2004.05150 [cs].

[60] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

[61] Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1):250, September 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-01033-3. URL https://doi.org/10.1038/s41597-021-01033-3.

[62] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, March 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03068-4. URL https://doi.org/10.1038/s41467-018-03068-4.

[63] Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, 112(5):1105–1118, September 2014. ISSN 0022-3077. doi: 10.1152/jn.00884.2013. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4122731/.

[64] Stanford. Alpaca, 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html.

[65] Jorie Koster-Hale, Rebecca Saxe, James Dungan, and Liane L. Young. Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110(14):5648–5653, April 2013. doi: 10.1073/pnas.1207992110. URL https://www.pnas.org/doi/10.1073/pnas.1207992110. Publisher: Proceedings of the National Academy of Sciences.

[66] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=XPZIaotutsD.

[67] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

[68] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR, July 2023. URL https://proceedings.mlr.press/v202/biderman23a.html. ISSN: 2640-3498.

[69] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, and et al. Gemma: Open Models Based on Gemini Research and Technology, April 2024. URL http://arxiv.org/abs/2403.08295. arXiv:2403.08295.

[70] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. Baichuan 2: Open Large-scale Language Models, April 2025. URL http://arxiv.org/abs/2309.10305. arXiv:2309.10305 [cs].

[71] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, and et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

[72] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools, July 2024. URL http://arxiv.org/abs/2406.12793. arXiv:2406.12793 [cs].

[73] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs].

[74] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile

Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

[75] Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, February 2020. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2019.107307. URL https://www.sciencedirect.com/science/article/pii/S0028393219303495.

[76] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://zenodo.org/records/5297715.

[77] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, April 2016. ISSN 1476-4687. doi: 10.1038/nature17637. URL https://www.nature.com/articles/nature17637. Publisher: Nature Publishing Group.

[78] Alyssa Hughes. Phi-2: The surprising power of small language models, December 2023. URL https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

[79] Ping Li, Chun-Ting Hsu, Ben Schloss, Anya Yu, Lindsey Ma, Marissa Scotto, Friederike Seyfried, and Chanyuan Gu. The Reading Brain Project L1 Adults, 2022. URL https://openneuro.org/datasets/ds003974/versions/3.0.0.

[80] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[81] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x.

[82] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 71683–71702, Red Hook, NY, USA, December 2023. Curran Associates Inc.

[83] Vicuna team. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org. URL https://lmsys.org/blog/2023-03-30-vicuna.

[84] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language And Vision Alignment Model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, June 2022. doi: 10.1109/CVPR52688.2022.01519. URL https://ieeexplore.ieee.org/document/9880206. ISSN: 2575-7075.

[85] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[86] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, October 2023. URL http://arxiv.org/abs/2308.12966. arXiv:2308.12966 [cs].

[87] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224, December 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.10.014. URL https://doi.org/10.1016/j.neuron.2012.10.014. Publisher: Elsevier.

[88] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. BridgeTower: building bridges between encoders in vision-language representation learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, volume 37 of *AAAI'23/IAAI'23/EAAI'23*, pages 10637–10647. AAAI Press, February 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i9.26263. URL https://doi.org/10.1609/aaai.v37i9.26263.

[89] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research*, August 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=b4tMhpN0JC.

[90] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N. Apurva Ratan Murty, Kendrick Kay, Aude Oliva, and Radoslaw Cichy. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications*, 15(1):6241, July 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-50310-3. URL https://doi.org/10.1038/s41467-024-50310-3.

[91] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. doi: 10.1109/cvpr.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

[92] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, October 2021. doi: 10.1109/ICCV48922.2021.00676. URL https://ieeexplore.ieee.org/document/9710415. ISSN: 2380-7504.

[93] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, June 2022. URL https://proceedings.mlr.press/v162/li22n.html. ISSN: 2640-3498.

[94] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, October 2024. URL http://arxiv.org/abs/2408.03326. arXiv:2408.03326 [cs].

[95] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.01457. URL https://ieeexplore.ieee.org/document/10203733/.

[96] Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. TVLT: Textless Vision-Language Transformer. *Advances in Neural Information Processing Systems*, 35:9617–9632, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3ea3134345f2e6228a29f35b86bce24d-Abstract-Conference.html.

[97] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 10078–10093, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-71387-108-8.

[98] Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.488. URL https://aclanthology.org/2022.findings-emnlp.488/.

[99] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

[100] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

[101] Oscar Skean, Jhoan Keider Hoyos Osorio, Austin J. Brockmeier, and Luis Gonzalo Sanchez Giraldo. DiME: Maximizing Mutual Information by a Difference of Matrix-Based Entropies, July 2023. URL http://arxiv.org/abs/2301.08164. arXiv:2301.08164 [cs].

[102] Eghbal Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36:43918–43930, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/88dddaf430b5bc38ab8228902bb61821-Abstract-Conference.html.

[103] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

[104] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148.

[105] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *First Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[106] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Vaishaal Shankar, Alexander Toshev, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 12371–12384, Vienna, Austria, July 2024. JMLR.org.

[107] A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes, July 2023. URL http://arxiv.org/abs/2301.03198. arXiv:2301.03198 [cs].

A Appendix

A.1 Complementary details of the reviewed works

Table 4: Models grouped by modality.

Modality	Models
Text	BART [39], LED [59], BigBird [60], LongT5 [40], Word2Vec [80], GPT-2 [36], OPT [55], LLaMA 2 [37], GPT-Neo [76], T5 [40], Vicuna [83], Alpaca [64], T5 Flan [54], GPT-2-XL [36], DeBERTa [66], RoBERTa [67], Pythia [68], LLaMA 3 [37], Gemma [69], Baichuan2 [70], DeepSeek-R1 [71], GLM [72], Qwen2.5 [73], Mistral [74], BERT [35] (monolingual and muntilingual), XLM-R, XGLM, LLaMA-3.2 [38]), Phi-2 [78]
Speech	Wav2Vec2.0 [41], HuBERT [42], WavLM [56]
Image	ViT-H [43], ViT [43], ResNet-50 [91]
Video	VideoMAE [97], ViViTB [92]
Image + Text	InstructBLIP [45], mPLUG-Owl [46], IDEFICS [82], CLIP [44], BLIP-L [93], BridgeTower [88], GIT [89], LLaVA [85], FLAVA [84], Qwen2.5-VL [86].
Video + Text	LLaVA-OV-7B [94]
Video + Speech	TVLT [96]
Speech to Text	Whisper [52]
Multimodal	ImageBind [95]