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

Angela Lopez-Cardona [1,2] Sebastian Idesis [1] Mireia Masias [1]
Sergi Abadal [2] Ioannis Arapakis [1]

[1]Telefónica Scientific Research, Barcelona, Spain [2]Universitat Politècnica de Catalunya, Barcelona, Spain

Introduction

- **Research Question:** Do brains and Language Models (LMs) converge to similar internal representations?
- Alignment studied via fMRI 🧠 ↔ LM activations (linear maps).
- Factors: performance, scale, architecture, dataset, modality, fine-tuning.
- **Platonic Representation Hypothesis (PRH)** [1]
- **Intermediate-Layer Advantage** [2]
- We review 25 studies (since 2023) testing these two hypotheses.

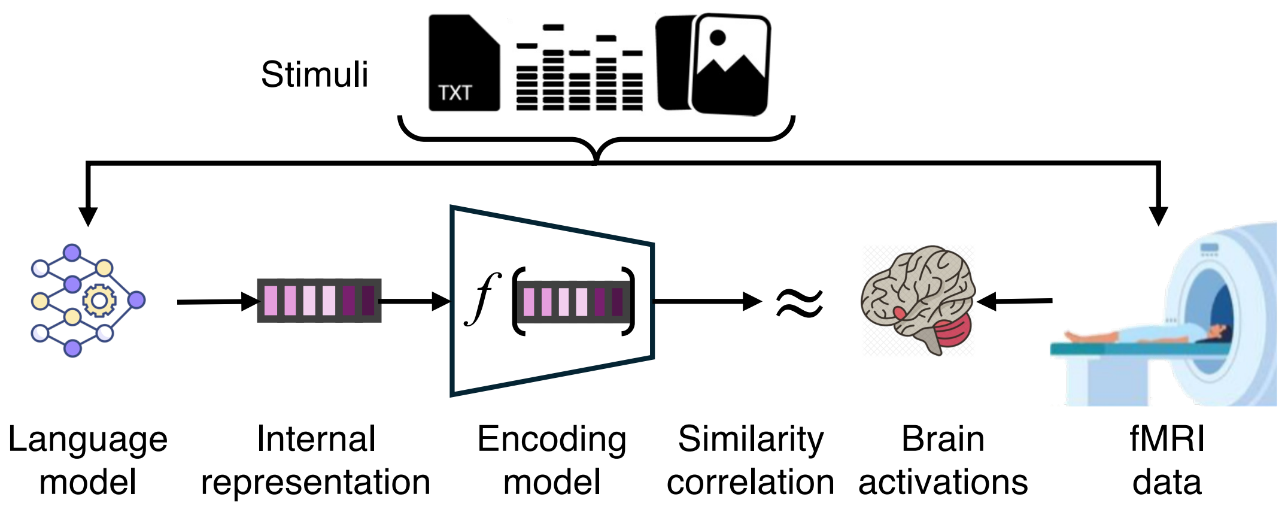

Figure 1.

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

Summary of Reviewed Works: Data, Models, and Methods

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?	[3, 4, 5, 6, 7, 8]
Scaling laws and architecture size	How do parameter count, data scale, and architectural choices affect alignment?	[9, 10, 11, 12, 13]
Task-specific training effects	Do models trained for specific objectives (e.g., moral reasoning, speech) align better with brain data?	[14, 15, 16, 17]
Instruction-tuning and human alignment	Does instruction tuning change the correspondence between model representations and neural activity?	[18, 12, 10, 19]
Cross-lingual and multilingual effects	Do different languages converge to a shared conceptual space in the brain?	[20]
Brain-informed tuning	Does fine-tuning on brain/behavioral signals improve neural predictivity?	[16, 21, 14, 15]
Modality differences	How do audio-based vs. text-based models compare in predicting brain signals?	[22, 9]
Multimodal vs. unimodal models	Do multimodal models predict brain activity better than unimodal ones?	[23, 24, 25, 19, 26, 3, 27]

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)

	Dataset	Model(s)
[14]	Passive natural language listening [28] (L)	Wav2Vec2.0 [29] and HuBERT [30]
[15]	Podcast Stories [31] (L)	Wav2Vec2.0 [29] and HuBERT [30], Whisper [32]
[22]	Subset Moth Radio Hour [33] (R)	BERT [34], GPT-2 [35], T5 Flan [36], Wav2Vec2.0 [29], Whisper [32]
[9]	Podcast Stories [31] (L)	OPT [37], LLaMA [38], HuBERT [30], WavLM [39], Whisper [32]
[20]	The Little Prince [40] (L)	Monolingual, multilingual, untrained BERT [34], Whisper [32]
[17]	Harry Potter Dataset [41] (R)	BART [42], LED [43], BigBird [44] and LongT5 [45]
[5]	Narratives [46] (L)	BERT [34], GPT2 [35]
[18]	Pereira [47] (R), BLANK2014 [48] (L), Harry Potter Dataset [41] (R)	GPT2 [35], T5 [45], LLaMa 2 [38], Vicuna, Alpaca [49], T5 Flan [36]
[8]	Harry Potter Dataset [41] (R)	GPT-2 [35]
[13]	Pereira [47] (R+V)	GPT-2 [35]
[4]	Pereira [47] (R)	GPT-2-XL [35]
[16]	Moral judgement [50]	BERT [34], DeBERTa [51](T), RoBERTa [52]
[7]	Podcast Stories [31] (L)	OPT[37], Pythia [53]
[10]	The Little Prince [40] (L)	Llama 3 [38], Gemma [54], Baichuan2 [55], DeepSeek-R1 [56], GLM [57], Qwen2.5 [58], OPT [37], Mistral [59], BERT [34]
[11]	Natural Stories fMRI [60] (L), Pereira [47] (R)	GPT-2 [35], GPT-Neo [61], OPT [37], and Pythia [53]
[21]	Moth Radio Hour [62] (R)	Monolingual (text english, chinese), multilingual BERT [34], XLM-R, XGLM, LLaMA-3.2 [63])
[6]	Narratives [46] (L)	GPT-2 [35], LLaMA 2 [38], and Phi-2 [64]
[12]	Reading Brain [65] (R)	LLaMA [63], GPT [35], Mistral [59], Alpaca [49], Gemma [54]
[24]	Sherlock clips [24] (L+V)	ViT [66], Word2Vec [67], GPT2 [35]
[19]	Natural Scenes Dataset [68] (V)	InstructBLIP [69], mPLUG-Owl [70], IDEFICS [71], ViT-H [66], and CLIP [72]
[27]	Pereira [47] (R+V)	GPT-2 [35] , Qwen-2.5 [58], Vicuna-1.5 [73], FLAVA [74], LLaVA [75], Qwen2.5-VL [76]
[25]	Moth Radio Hour [62] (R), Movie watching [77] (L+V)	BridgeTower [78], RoBERTa [52] and ViT [66]
[26]	Japanese movie [26] (L+V)	Word2Vec [67], BERT [34], GPT2 [35], OPT [37], Llama 2 [38], CLIP [72], GIT [79], BridgeTower [78], LLaVA [75]
[3]	BOLD Moments Dataset [80] (V)	ResNet-50 [81], ViViTB [82], CodeLlama-7B, Llama3-8B [63], BLIP-L [83], LLaVA-OV-7B [84]
[23]	Movie10 [85] (L+V)	ImageBind [86], TVLT [87], Wav2Vec2.0 [29], ViT-B [66], ViViTB [82], VideoMAE [88]

Acknowledgments

Symbiotik: Horizon Europe's European Innovation Council, under Pathfinder Grant 101071147.
AGAUR: Generalitat de Catalunya, under Grant AGAUR 2023 DI060.

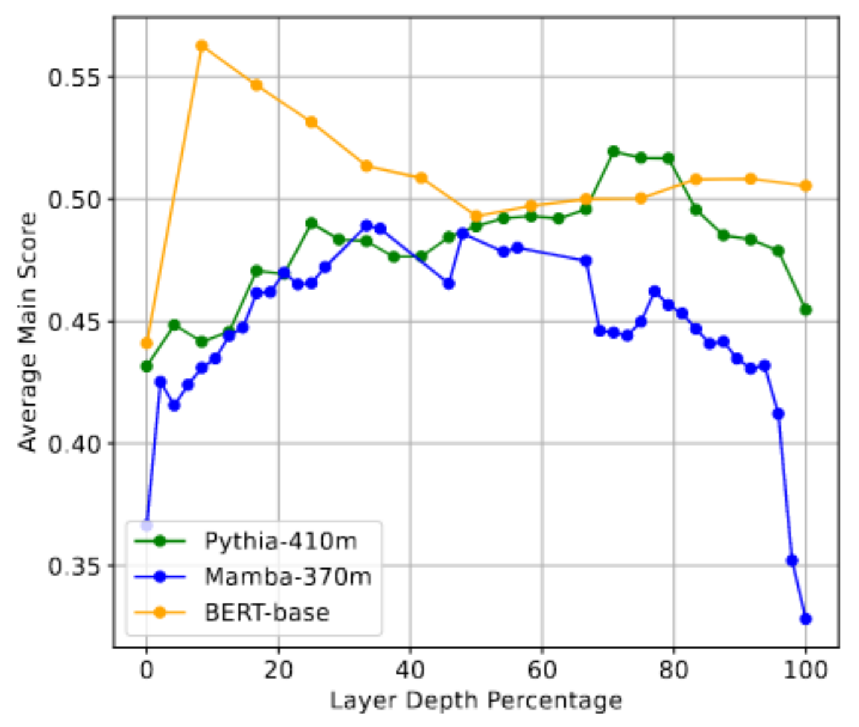

Figure 2. Intermediate layers consistently outperform final layers on downstream tasks [2].

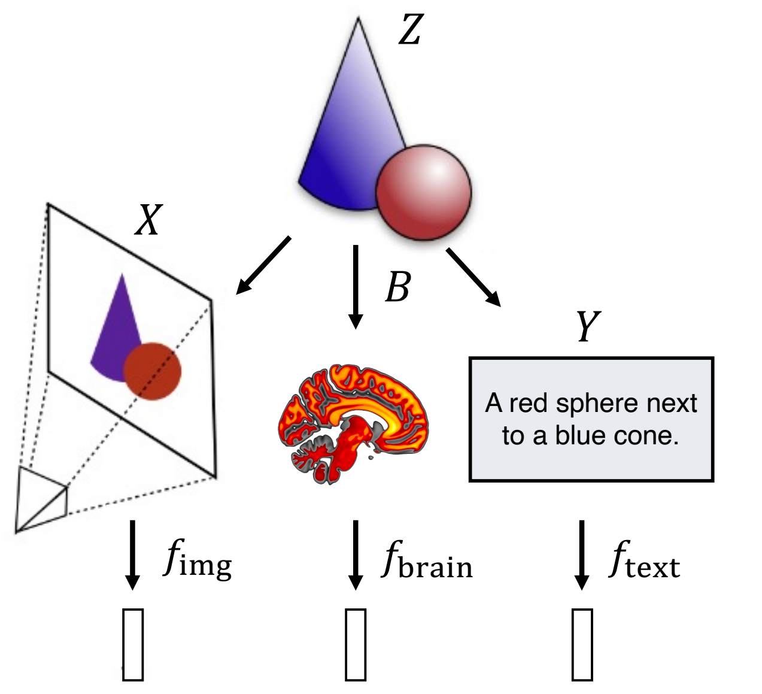

Figure 3. Platonic Representation Hypothesis (adapted from original [1]): Images (X), text (Y), and brain activity (B) are projections of a common underlying reality (Z).

Platonic Representation Hypothesis

Models and brains converge toward shared reality representation.

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

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

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

Intermediate-Layer Advantage

Intermediate layers encode richer and more informative representations, and leveraging them can lead to improved performance.

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.

Table 3. Overview of the reviewed studies classified by modality (speech, text, speech, text, images, text, video, text, multimodal). Each column represents one of the four hypotheses. 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
[14]	neutral	agreement	neutral	strong agreement
[15]	neutral	agreement	neutral	neutral
[22]	neutral	neutral	neutral	agreement
[9]	strong agreement	neutral	neutral	agreement
[20]	neutral	agreement	agreement	agreement
[17]	neutral	agreement	neutral	agreement
[5]	agreement	neutral	neutral	strong agreement
[18]	neutral	agreement	neutral	neutral
[8]	neutral	neutral	neutral	neutral
[13]	agreement	neutral	neutral	neutral
[4]	neutral	neutral	neutral	strong agreement
[16]	neutral	disagreement	neutral	neutral
[7]	agreement	neutral	neutral	strong agreement
[10]	agreement	agreement	neutral	strong agreement
[11]	strong disagreement	neutral	neutral	neutral
[21]	neutral	agreement	neutral	neutral
[6]	neutral	neutral	neutral	agreement
[12]	agreement	disagreement	neutral	neutral
[24]	neutral	neutral	strong agreement	neutral
[19]	neutral	neutral	strong agreement	agreement
[27]	neutral	neutral	agreement	neutral
[25]	neutral	neutral	strong agreement	agreement
[26]	neutral	neutral	agreement	neutral
[3]	neutral	neutral	agreement	agreement
[23]	neutral	neutral	strong agreement	agreement

Discussion

- Alignment tends to be stronger in **larger, multimodal, and instruction-tuned models**, supporting the Platonic Representation Hypothesis.
- **Intermediate layers**, rather than final ones, show the strongest brain alignment across architectures.
- Cross-modal training yields **modality-independent representations** that better match brain activity.
- Overall, evidence suggests brains and models may **converge toward shared abstract representations**, though findings remain qualitative.