Foundational Human Modeling: A roadmap for brain-grounded multi-modal AI

Richard Csaky
richard.csaky@gmail.com

Abstract

Foundation agents excel at language, perception, and tool-use, yet remain brittle across domains and time: their success rates can exhibit "half-lives" under distribution shift, and their reasoning is often forced through language tokens (Liu et al., 2025; Dell'Acqua et al., 2023; Ord, 2025; McCoy et al., 2024). This roadmap argues for *Foundational Human Modeling (FHM)*: scaling generative brain models (EEG/MEG/ECoG/sEEG/fMRI) and integrating *brain tokens* into multimodal token streams to provide privileged internal signals for grounding, regularization, and human-like cognition. We propose: (i) scaling next-brain-token prediction toward stable long-horizon neural rollouts (Csaky, 2026; Csaky et al., 2024; Huang et al., 2025); (ii) unifying heterogeneous brain modalities via shared geometry and adapters (Xiao et al., 2025; Sato et al., 2024); and (iii) treating cognition as an interleaved typed token stream in which brain, vision, language, and action tokens can be produced in a dynamic ordering (Reed et al., 2022; Zhou et al., 2024).

1 Motivation and Scope

Contemporary AI systems are shaped by the objectives they optimize (McCoy et al., 2024). Even as benchmarks improve, key desiderata for human-like intelligence remain open: understanding human goals and beliefs, acting in uncertainty-aware ways people can predict, and maintaining reliability over long-horizon tasks (Collins et al., 2024; Ying et al., 2025; Ord, 2025). Separately, neuroscience-inspired critiques emphasize that animal brains exhibit inductive biases that may not emerge from "pure learning" on external behavior alone (Zador, 2019).

This roadmap treats brain recordings as *privileged information* about internal computation: signals that can supervise latent variables, regularize representations, and ground agents in human-like dynamics. We focus on a concrete program—FHM—that progresses from scalable generative brain modeling to multimodal interleaving with language and action.

We organize FHM around four testable theses.

1. **Long-horizon brain generation is foundational.** A model that produces stable, realistic long trajectories ("full-session" rollouts) is a cornerstone for downstream alignment, decoding, and control (Csaky, 2026; Huang et al., 2025). Accurate rollouts enable counterfactual evaluation of policies under a human-like latent prior.

2. **Brain modalities are multiple views of shared latent dynamics.** EEG, MEG, ECoG/sEEG, and fMRI are noisy, rate-distorted measurements of related underlying processes; joint representation learning should be possible via shared geometry and cross-modal adapters (Xiao et al., 2025; Sato et al., 2024).

3. **Cognition is best modeled as a typed, interleaved token stream.** Models should choose when to emit/consume stimulus, brain, language, and action tokens rather than obeying a fixed order; this generalizes multimodal agent architectures (Reed et al., 2022; Zhou et al., 2024).

4. **Feasibility can be probed via proxy scaling experiments.** Controlled corruption/noise studies can forecast the returns to cleaner modalities and better tokenization prior to expensive data collection efforts (Suzgun et al., 2025; Park et al., 2025).

2 Related Work

Learning using privileged information. Brain recordings provide a privileged view into the internal dynamics that mediate perception, cognition, and action: they expose intermediate variables that are often causally upstream of overt behavior, and they do so at temporal resolutions (EEG/MEG/ECoG) that are difficult to infer from actions alone (Csaky, 2026). This connects naturally to the *learning using privileged information* (LUPI) paradigm, where additional signals available at training time can improve generalization by constraining a learner toward latent structure that explains the target outputs (Vapnik and Vashist, 2009). Neural measurements can act as such privileged training signals: brain-derived objectives can regularize vision models toward more robust representations (Li et al., 2019), and "brain-tuning" speech-language models on fMRI can induce more brain-relevant semantics while improving downstream performance (Moussa et al., 2024). From this perspective, a powerful *generative* model of brain activity is not merely a decoder for downstream tasks; it is a way to internalize reusable structure about neural dynamics and, indirectly, about the predictive world models that biological systems implement.

Token-stream modeling is becoming a dominant design pattern in multimodal generative systems: high-bandwidth modalities are first mapped to compact sequences of discrete tokens, and a single decoder-only backbone is trained over interleaved multimodal streams with next-token prediction. Emu3, Emu3.5, Qwen2.5-VL, and Qwen3-VL demonstrate that a native multimodal Transformer trained with next-token prediction over unified vision–language (and video) tokens can support both perception and high-fidelity generation, including long-horizon synthesis (Wang et al., 2024; Cui et al., 2025; Bai et al., 2025b,a). These developments suggest an appealing blueprint for brain foundation models: (i) learn modality-specific tokenizers that compress diverse neural recordings into a common discrete space; (ii) train a shared autoregressive backbone over unified token streams; and (iii) expose standardized interfaces for conditioning, prompting, and downstream adaptation.

Brain–language alignment and multimodal neural models. Generative priors over neural activity are also motivated by decoding and alignment: reconstructing stimuli or behavior, building semantically meaningful neural representations, and characterizing what aspects of cognition are captured by current AI systems. Several works treat brain activity as a "foreign language" by

learning neural tokenizers and coupling them to large language model backbones. NeuroLM learns a text-aligned EEG tokenizer and uses multi-task instruction tuning for unified EEG inference (Jiang et al., 2025). NeuGPT and fMRI-LM similarly aim to jointly model neural tokens and text across multiple recording modalities, enabling language-conditioned understanding and generation from neural recordings (Yang et al., 2024; Wei et al., 2025). Orthogonally, work in cognitive NLP studies representational alignment between LLMs and neural responses, including evidence that instruction tuning (Aw et al., 2023) as well as scaling and training choices (Gao et al., 2025) can systematically increase alignment.

3 Typed Tokens and Dynamic Interleaving

All LLM-style (including VLMs and VLAs) models should have reasoning modules after having learnt language. The purpose of a reasoning module however is not to produce a stream-of-words, but rather to allow for deliberation without action (Xie et al., 2025; Kim et al., 2025; Wang et al., 2026). Reasoning can be done in various modalities—including brain tokens.

Such a model might observe various types of tokens: visual, sound, text, brain; and be able to produce and interleave these (plus action tokens relevant to the environment). Importantly during production there is no fixed order, similar to how a brain might pause mid-sentence to do a tiny deliberation loop (with or without explicit language-thoughts). Why shouldn't the model be able to start "thinking" while it has not finished a paragraph, or continue thinking while producing an output—this would help with error recovery / self-monitoring. A plausible scenario during gameplay may be to receive a number of visual tokens and then produce an interleaved stream of brain & text reasoning tokens plus output action tokens. Visual, sound, and action self-communication, i.e. imagery can be similarly conceived in the form of $\mathrm{V}^{(self)}$, $\mathrm{S}^{(self)}$, $\mathrm{A}^{(self)}$.

3.1 Token taxonomy

More formally, let $\mathcal{T}$ be the space of discrete tokens; each modality m has token type $\mathcal{T}_m$ and (optionally) a tokenizer ϕ_m and detokenizer ψ_m. We distinguish *roles* of tokens:

$$m^{(i)} \text{ (input from environment)}, \qquad m^{(s)} \text{ (self / private)}, \qquad m^{(o)} \text{ (output to environment)}.$$

For example, text decomposes into $\mathrm{T}^{(i)}$ (observed prompt), $\mathrm{T}^{(s)}$ (reasoning traces), and $\mathrm{T}^{(o)}$ (emitted response). Brain tokens B can appear as internal cognition ($\mathrm{B}^{(s)}$) and, in experimental settings, as observed streams ($\mathrm{B}^{(i)}$).

3.2 Dynamic ordering as a learned policy

Temporal ordering between the different modalities may be difficult. In VLMs, images usually come first as conditioning for the generated language. What makes sense in the first instance and mostly aligns with "brain model as grounding" and "brain prior" ideas is to construct the flat sequence by putting stimulus/input tokens first, then brain tokens, and then output tokens (i.e. text/actions). Thus the brain tokens are conditioned on the input tokens, and the overall output is conditioned

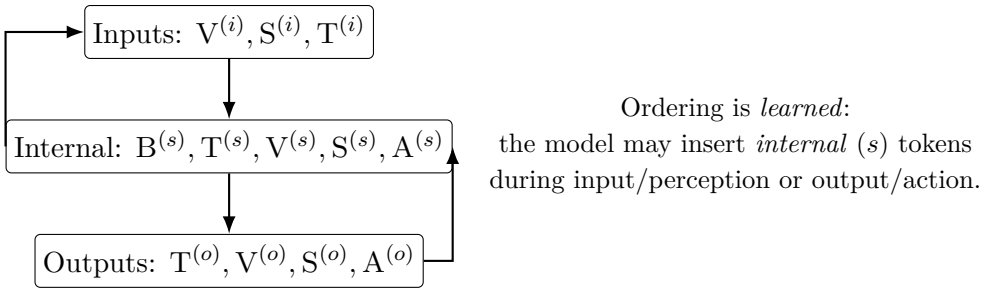

Figure 1: Typed tokens and interleaving. FHM proposes that models learn *when* to allocate internal computation (brain/self tokens) rather than externalizing all deliberation as emitted text. B refers to brain tokens, V to image/video, S to audio, T to text, and A to discrete actions.

both on the input and the brain tokens. However, if the input modality is sequential as well, e.g. listening, this may not be optimal. In the most general case the model should not be constrained by any ordering and decide for itself when to emit certain token types, i.e. mix in various reasoning tokens both during processing of input tokens, and production of output tokens.

Formally, a standard next-token model factors:

$$p_\theta(\tau_{1:N}) = \prod_{t=1}^{N} p_\theta(\tau_t \mid \tau_{<t}),$$

but we do *not* fix τ_t to be text-first, then reasoning/brain, then action. Instead, the model learns to choose token types:

$$p_\theta(\tau_t \mid \tau_{<t}) = \sum_{k \in \mathcal{K}} p_\theta(k \mid \tau_{<t})\, p_\theta(\tau_t \mid k, \tau_{<t}),$$

where k indexes token-type/role (e.g., $T^{(i)}, B^{(s)}, A^{(o)}$). This mixture view makes "when to think" an explicit, learnable decision—without forcing deliberation to be (only) in language (Figure 1).

4 Pillar I: Scaling Brain Foundation Models

4.1 Objective

Train generative models on large-scale neural recordings that can generate stable long trajectories—with *stable full-session rollouts* as a concrete north star (Csaky, 2026; Huang et al., 2025). Accurate rollouts enable counterfactual evaluation of policies under a human-like latent prior.

4.2 Key technical avenues

Data scaling and diversity. Increase coverage across subjects and tasks (Banville et al., 2025; Sato et al., 2024).

Scheduled sampling / rollout robustness. Gradually replace teacher forcing with self-generated tokens to reduce exposure bias (Bengio et al., 2015).

Block-causal masking. Full attention within blocks, causal across blocks; predict all channels at $t+1$ in a masked fashion.

Context scaling curriculum. Train shorter contexts first, ramp to longer; compare long-context designs (Hawthorne et al., 2022; Hiller et al., 2024).

Multi-timescale hierarchy. Introduce slow/fast latent tokens; draw on hierarchical autoregressive designs in video/audio (Kondratyuk et al., 2023; Dhariwal et al., 2020; Défossez et al., 2022).

Efficient architectures. High-rate electrophysiology demands long-context efficiency. Compare Transformer baselines (Vaswani et al., 2017) to long-context attention designs (Hawthorne et al., 2022; Hiller et al., 2024) and linear-time sequence models (Gu and Dao, 2024; Poli et al., 2023; Peng et al., 2024; Zuo et al., 2025).

Better tokenization. Tokenization is a major design axis. Discrete RVQ/VQ-style tokens have enabled strong generative modeling in audio and video (van den Oord et al., 2017; Défossez et al., 2022; Yan et al., 2021; Lee et al., 2022). Brain foundation models increasingly adopt vector quantization or RVQ variants (Jiang et al., 2024; Chen et al., 2024; Barmpas et al., 2025; Csaky et al., 2024; Huang et al., 2025). Soft tokenization can reduce quantization artifacts and may improve gradient flow for mixed modalities (Tschannen et al., 2025, 2024; Zhou et al., 2024).

4.3 Evaluation targets

Distributional fidelity: PSD distance; cross-spectral density similarity; autocorrelation structure; functional connectivity similarity between real and generated sequences.

Long-horizon stability: absence of collapse/drift over long rollouts; stationarity diagnostics over sliding windows.

Evoked activity: Given a long initial input the model should be able to generate varied and specific (to the dataset/subject) task data indefinitely which is useful for downstream decoding.

Generalization: cross-subject and cross-dataset likelihood/perplexity; robustness to missing channels/sensors.

5 Pillar II: Scaling Laws and Modality Fusion

5.1 Objective

Quantify how performance scales with data and modality choice under acquisition cost constraints, then design optimal mixtures.

Let modalities $\mathcal{M}$ have per-hour cost c_m and data hours h_m. Given budget C:

$$\max_{\{h_m\}_{m\in\mathcal{M}}} \quad \mathrm{Perf}(\{h_m\}) \quad \text{s.t.} \quad \sum_{m\in\mathcal{M}} c_m h_m \le C.$$

Empirically, decoding performance shows log-linear scaling across modalities and data regimes (Banville et al., 2025; Sato et al., 2024). FHM intends to extend this to *generative* metrics (rollout stability, likelihood) and *agent utility*.

5.2 Proxy scaling via "simulating brain noise in LLMs"

When neural data is scarce, study controlled noise injection in text/vision to approximate modality corruption. Related work shows that subtle training/objective choices and information leakage can shape model behavior strongly (McCoy et al., 2024; Park et al., 2025). The aim is to forecast whether collecting cleaner modalities (e.g., MEG vs EEG) yields better returns than scaling hours within a modality.

5.3 Modality Fusion

A practical unification path across modalities is to project sensor-space signals into a canonical source-space atlas (e.g., ROI-parcellated cortex), enabling shared tokenizers across devices and modalities. Recent work suggests that device- and modality-heterogeneity can be handled explicitly by conditioning on sensor metadata rather than assuming a fixed channel topology. For example, Xiao et al. (2025) introduce a sensor encoder that leverages physical properties (e.g., positions/orientations/types) to unify heterogeneous EEG and MEG setups, enabling robust fusion across varying channel counts and previously unseen devices. A natural extension is to incorporate invasive electrophysiology (ECoG/sEEG/iEEG) by treating each contact as an additional "sensor type" with 3D cortical coordinates and modality-specific noise/forward assumptions; notably, topology-agnostic transformer architectures that operate over arbitrarily positioned electrodes (surface grids, strips, and depth contacts) have already been demonstrated for multi-subject decoding using 3D positions as the organizing geometry (Chen et al., 2025). Moreover, simultaneous MEG–EEG–SEEG recordings provide direct trimodal supervision for aligning non-invasive and invasive signals within a shared latent space (Dubarry et al., 2014). Finally, bridging electrophysiology and fMRI can be framed as a multirate latent-variable problem with modality-specific observation models: a fast latent cortical process constrained by MEG/EEG, and a slow hemodynamic readout obtained via downsampling and HRF-like temporal filtering for fMRI (Jin and Wehbe, 2025). This suggests that fMRI voxel/ROI representations can be mapped into the same shared latent space used for electrophysiology, yielding hierarchical constraints across temporal scales without forcing a single sampling rate.

All modalities provide observations on the underlying shared latent phenomena, thus learning a joint model or representational space should be very possible, with e.g., adapters. VLMs are able to translate between images and language due to shared latents, so similarly this should be very doable across all brain modalities. For example we could use the same tokenizer for EEG & MEG, and train a different one for fMRI, but the core Transformer model is the same. The Transformer model is already the same over input stimuli (e.g. images) and brain data, so different brain modalities can be treated in the same general manner.

An alternate approach is to use fMRI to define slow cognitive state variables z_{slow} and train fast electrophysiology models to predict these states:

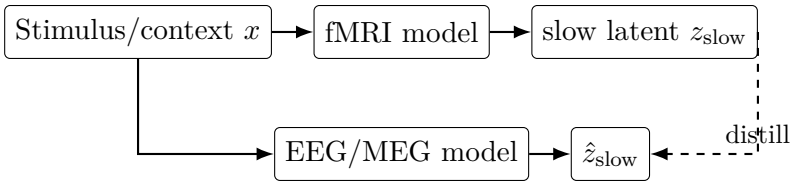

Figure 2: Multi-rate distillation: learn slow cognitive states from fMRI and transfer them to fast electrophysiology predictors.

1. Train an fMRI model to produce a slow latent z_{slow} from stimulus/context or from voxels/ROIs.
2. Train an EEG/MEG model to predict that same z_{slow} from fast signals (distillation / representation matching).

Let x be stimulus/context, fast electrophysiology $e_{1:T}$, and slow fMRI $f_{1:U}$ (with $T \gg U$). A generic hierarchical model:

$$z_{\text{slow}}(u{+}1) \sim p(z_{\text{slow}}(u{+}1) \mid z_{\text{slow}}(u), x)\,, \tag{1}$$

$$e_t \sim p(e_t \mid z_{\text{slow}}(\lfloor t/r \rfloor), z_{\text{fast}}(t))\,, \tag{2}$$

$$f_u \sim p(f_u \mid \text{HRF}(z_{\text{slow}}(u)))\,, \tag{3}$$

where $\text{HRF}(\cdot)$ is a hemodynamic blur (e.g., a learned convolution kernel), and z_{fast} is a fast electrophysiology latent at the sample rate. This connects to joint hemodynamic/electrophysiology modeling with common source mappings (Jin and Wehbe, 2025). Figure 2 gives an overview of the distillation.

6 Pillar III: Interleaved Token-Stream Cognition

6.1 Objective

Model cognition as a sequence over typed tokens:

$$\underbrace{[\text{V}, \text{S}, \text{T}]^{(i)}}_{\text{stimulus / input}} \;+\; \underbrace{[\text{B}, \text{T}, \text{V}, \text{S}, \text{A}]^{(s)}}_{\text{internal / brain \& self}} \;+\; \underbrace{[\text{T}, \text{V}, \text{S}, \text{A}]^{(o)}}_{\text{action / output}},$$

where the ordering is *not fixed* (Figure 1).

6.2 Why interleaving matters

The brain analogy for LLMs is that the input token processing is the understanding part, and then there is a discrete thinking part (with no external stimuli, *brain in a bottle*), which can be stopped arbitrarily to produce behavior — depending on compute budget or learned dynamic policy.

A static, sequential brain dataset may be: see image → elicit brain activity → produce language. A dynamic dataset may be: watch movie → continuous brain activity → talking over it / gaze change. In this case input-brain-output token triplets are continuously interleaved to preserve temporal causality. Most useful cases are probably dynamic.

When input is continuous (e.g., listening, movie watching), enforcing a strict "input then brain then output" layout may be suboptimal. A more general model should decide when to emit certain token types:

- Think during perception (insert brief internal loops while processing input).
- Self-monitor during output (continue deliberation while speaking/acting).
- Dynamically allocate compute and "brain token budget" for a task.

We note that a major challenge in this approach is the availability of varied interleaved modality datasets or an appropriate training paradigm.

6.3 Integration paths with frontier AI

Unified token-stream model. Fine-tune a multi-modal agent to interleave brain tokens, similar in spirit to unified token spaces in multi-modal generation (Zhou et al., 2024; Tschannen et al., 2025).

Brain signals as conditioning. Map brain activity into the embedding space of an LLM/VLM via an adapter and keep the foundation model fixed (adapter-only training). This is analogous to "brain adapter" approaches in neural-to-language reconstruction (Ye et al., 2025).

Distillation. Train core multi-modal AI model (agent) and brain models separately, then distill brain model into the agent policy:

$$\min_{\phi} \; \mathbb{E}\Big[\mathrm{KL}\Big(p_\theta(\mathbf{a} \mid \mathbf{x}, \mathbf{z}) \parallel \pi_\phi(\mathbf{a} \mid \mathbf{x})\Big)\Big], \tag{4}$$

which is an explicit LUPI instantiation, where $\mathbf{a}$ are output actions, $\mathbf{x}$ are inputs, and $\mathbf{z}$ are brain signals. Brain signals can also regularize latent state spaces.

A key question is whether brain tokens reduce the need for language-based deliberation loops, improving reliability and sample efficiency, particularly on long-horizon tasks (Ord, 2025; Kargupta et al., 2025).

6.4 Token-Stream Encoding–Decoding

Behavior-producing brain datasets (free-form language, rich action) are scarce; naturalistic stimuli datasets (e.g., listening, movie watching) are more abundant (Appendix A includes candidates).

For such datasets we propose training a model on streams such as:

[language tokens] → [encoding B] → [reasoning B] → [decoding B] → [decoded language],

with optional separator tokens [SEP] for boundaries. Such a model can be trained on existing listening datasets (no need for language production), where brain tokens can be used both for encoding and decoding within the same model with alternating training batches. Once a base model is trained: for instruction tuning, we can obtain the target sentence's brain tokens using encoding and use that as the decoded brain tokens. This aligns with unified neural GPT approaches that combine brain and language/audio token streams (Yang et al., 2025) and general neural foundation modeling across tasks (Azabou et al., 2023).

We propose the following evaluations.

1. Stimulus→brain prediction quality and calibration.
2. Brain→text decoding (when available) with controlled generalization splits.
3. Ablations on where/when internal tokens are inserted (policy over token types).

7 Benchmarking

Resting-state generation: long-horizon rollouts with spectral/connectivity fidelity metrics.

Stimulus-aligned prediction: naturalistic listening/movie watching; predict future brain tokens given stimulus history.

Cross-modal translation: EEG↔MEG, electrophysiology↔fMRI (where paired), stimulus↔brain.

Behavior decoding: brain→text (speech/typing) and brain→action where available; tie to agent benchmarks that probe reliability in real environments (Xie et al., 2024; , METR; Yue et al., 2025; Laine et al., 2024).

Generalization:

1. **Cross-subject:** train on subjects A, test on unseen subjects B.
2. **Cross-dataset:** train on dataset X, test on dataset Y with aligned tasks/modality.

8 Conclusion

FHM is a concrete path toward scalable modeling of internal human dynamics and toward integrating privileged neural signals into multimodal agents. The program emphasizes (i) long-horizon generative brain modeling, (ii) cost-aware fusion of heterogeneous modalities via shared geometry and latents, and (iii) typed interleaving of stimulus, brain, language, and action tokens. The roadmap aims to evolve from scattered datasets and prototypes into a coherent foundation-model paradigm.

Ethics and privacy. Neurophysiological data are highly sensitive and can carry subject-identifying signatures and unintended correlates of health, identity, or behavior; therefore, FHM should be

pursued only under ethics oversight with explicit informed consent, strong de-identification, and strict licensing compliance. Evaluation and mitigation of re-identification risks should be conducted before any public model release.

References

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain, 2023. COLM 2024.

Mehdi Azabou et al. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ca113d122584f12a6727341aaf58887-Abstract-Conference.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025b.

Hubert Banville, Yohann Benchetrit, Stéphane d'Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*, 2025.

Konstantinos Barmpas, Na Lee, Alexandros Koliousis, Yannis Panagakis, Dimitrios A. Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models, 2025. URL https://arxiv.org/abs/2510.13068.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 1171–1179, 2015. doi: 10.5555/2969239.2969370. URL https://papers.nips.cc/paper/5956-scheduled-sampling-for-sequence-prediction-with-recurrent-neural-networks.

Junbo Chen, Xupeng Chen, Ran Wang, Chenqian Le, Amirhossein Khalilian-Gourtani, Erika Jensen, Patricia Dugan, Werner Doyle, Orrin Devinsky, Daniel Friedman, Adeen Flinker, and Yao Wang. Transformer-based neural speech decoding from surface and depth electrode signals. *Journal of Neural Engineering*, 22(1):016017, 2025. doi: 10.1088/1741-2552/adab21.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. *arXiv preprint arXiv:2401.10278*, 2024.

Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark K. Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024. doi: 10.1038/s41562-024-01991-9. URL https://doi.org/10.1038/s41562-024-01991-9. arXiv:2408.03943.

Richard Csaky. Scaling next-brain-token prediction for MEG, 2026. URL https://arxiv.org/abs/2601.20138. arXiv:2601.20138.

Richard Csaky, Mats WJ van Es, Oiwi Parker Jones, and Mark Woolrich. Gpt2meg: Quantizing meg for autoregressive generation. *arXiv preprint arXiv:2404.09256*, 2024.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. doi: 10.48550/arXiv.2210.13438. URL https://arxiv.org/abs/2210.13438.

Fabrizio Dell'Acqua et al. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. SSRN 4573321, 2023. URL https://www.hbs.edu/faculty/Pages/item.aspx?num=64700. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020. URL https://arxiv.org/abs/2005.00341.

Anne-Sophie Dubarry, Jean-Michel Badier, Agnès Trébuchon-Da Fonseca, Martine Gavaret, Romain Carron, Fabrice Bartolomei, Catherine Liégeois-Chauvel, Jean Régis, Patrick Chauvel, F.-Xavier Alario, and Christian-G. Bénar. Simultaneous recording of MEG, EEG and intracerebral EEG during visual stimulation: from feasibility to single-trial analysis. *NeuroImage*, 99:548–558, 2014. doi: 10.1016/j.neuroimage.2014.05.055.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Increasing alignment between language model and human brain representations. *Nature Computational Science*, 2025. doi: 10.1038/s43588-025-00863-0.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=AL1fq05o7H.

Curtis Hawthorne et al. General-purpose, long-context autoregressive modeling with perceiver AR. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17543–17565, 2022. URL https://proceedings.mlr.press/v162/hawthorne22a.html.

Markus Hiller, Krista A. Ehinger, and Tom Drummond. Perceiving longer sequences with bi-directional cross-attention transformers. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xvhjRjoFCN. NeurIPS 2024.

R. Huang, S. Cho, C. Gohil, O. P. Jones, and Mark Woolrich. Meg-gpt: A magnetoencephalography-focused large language model for the neural dynamics of naturalistic language and music processing, 2025. URL https://arxiv.org/abs/2510.18080.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP. ICLR 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2409.00101. Conference version: ICLR 2025.

Beige Jerry Jin and Leila Wehbe. Estimating brain activity with high spatial and temporal resolution using a naturalistic meg-fmri encoding model, 2025. URL https://arxiv.org/abs/2510.09415.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L Griffiths, Max Kleiman-Weiner, Jiawei Han, et al. Cognitive foundations for reasoning and their manifestation in llms. *arXiv preprint arXiv:2511.16660*, 2025.

Eunki Kim, Sangryul Kim, and James Thorne. Learning to insert [pause] tokens for better reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23760–23777, 2025.

Dan Kondratyuk et al. Videopoet: A large language model for zero-shot video generation, 2023. URL https://arxiv.org/abs/2312.14125.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms, 2024. URL https://arxiv.org/abs/2407.04694.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11523–11532, 2022. doi: 10.1109/CVPR52688.2022.01123. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lee_Autoregressive_Image_Generation_Using_Residual_Quantization_CVPR_2022_paper.html.

Zhe Li, Wieland Brendel, Edgar Y. Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian H. Sinz, Xaq Pitkow, and Andreas S. Tolias. Learning from brains how to regularize machines, 2019.

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*, 2025.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew R. Gormley, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(13):e2322420121, 2024. doi: 10.1073/pnas.2322420121. URL https://www.pnas.org/doi/10.1073/pnas.2322420121.

Model Evaluation & Threat Research (METR). Measuring AI ability to complete long tasks: time to 80% success rate in software tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/, 2025. URL https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning, 2024. Published as a conference paper at ICLR 2025.

Toby Ord. Is there a half-life for the success rates of ai agents?, 2025. URL https://arxiv.org/abs/2505.05115.

Yeji Park, Minyoung Lee, Seonghoon Chun, and Jiyoung Choe. Mitigating cross-image information leakage in mllm-based agentic data analysis, 2025. URL https://arxiv.org/abs/2508.13744.

Bo Peng et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=soz1SEiPeq.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28043–28078, 2023. doi: 10.48550/arXiv.2302.10866. URL https://arxiv.org/abs/2302.10866.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. doi: 10.48550/arXiv.2205.06175. URL https://arxiv.org/abs/2205.06175. Journal reference: Transactions on Machine Learning Research (11/2022).

Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *arXiv preprint arXiv:2407.07595*, 2024.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E. Ho, Thomas Icard, Dan Jurafsky, and James Zou. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, 7:245–255, 2025. doi: 10.1038/s42256-025-01113-8. URL https://www.nature.com/articles/s42256-025-01113-8.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pages 292–309. Springer, 2024.

Michael Tschannen, André Susano Pinto, and Alexander Kolesnikov. Jetformer: An autoregressive generative model of raw images and text. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=sgAp2qG86e. ICLR 2025.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315, 2017. doi: 10.5555/3295222.3295378. URL https://papers.nips.cc/paper/7210-neural-discrete-representation-learning.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. doi: 10.5555/3295222.3295349. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need. arXiv:1706.03762.

Jiecong Wang, Hao Peng, and Chunyang Liu. Latent chain-of-thought as planning: Decoupling reasoning from verbalization. *arXiv preprint arXiv:2601.21358*, 2026. doi: 10.48550/arXiv.2601.21358. URL https://arxiv.org/abs/2601.21358.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding, 2025.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals, 2025. Accepted at NeurIPS 2025.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. URL https://openreview.net/forum?id=tN61DTr4Ed. NeurIPS 2024.

Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models. *arXiv preprint arXiv:2508.15827*, 2025. doi: 10.48550/arXiv.2508.15827. URL https://arxiv.org/abs/2508.15827.

Wilson Yan, Anand Srivastava, Ali Farhadi, and Abhinav Gupta. Videogpt: Video generation using vq-vae and transformers, 2021. URL https://arxiv.org/abs/2104.10157.

Yiqian Yang, Yiqun Duan, Hyejeong Jo, Qiang Zhang, Renjing Xu, Oiwi Parker Jones, Xuming Hu, Chin-Teng Lin, and Hui Xiong. Neugpt: Unified multi-modal neural gpt, 2024.

Yiqian Yang, Yiqun Duan, Hyejeong Jo, Qiang Zhang, Renjing Xu, Oiwi Parker Jones, Xuming Hu, Chin-teng Lin, and Hui Xiong. Neugpt: Unified multi-modal neural GPT. In *International Conference on Machine Learning (ICML)*, 2025. URL https://icml.cc/virtual/2025/50469.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8:438, 2025. doi: 10.1038/s42003-025-07731-7. URL https://www.nature.com/articles/s42003-025-07731-7.

Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025. doi: 10.48550/arXiv.2409.02813. URL https://aclanthology.org/2025.acl-long.736/.

Anthony M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1):3770, 2019. doi: 10.1038/s41467-019-11786-6. URL https://doi.org/10.1038/s41467-019-11786-6.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim Khadraoui, et al. Falcon-h1: A family of hybrid-head language models redefining efficiency and performance. *arXiv preprint arXiv:2507.22448*, 2025.

A Semi-public Dataset Registry

Note. Counts/metadata below are approximate and should be verified against dataset releases and licenses. The intent is to provide a starting atlas for data aggregation. *Task Type* provides a broad categorization of tasks into three types: simple stimulus only (s. stim.), naturalistic stimulus only (nat. stim.), and datasets with behavior production.

Dataset	Subjects	Hours	Modality	Task Type	Notes	URL
TUH EEG Corpus	15000	27000	EEG	s. stim.	Clinical; resting; 20ch.	link
Elmiko EEG Corpus	55000	20000	EEG	s. stim.	Clinical; resting; 20ch.	link
HBN EEG	3000	2000	EEG	nat. stim.	Movies; EGI 128ch; pediatric.	link
AJILE12	12	1280	ECoG	production	Motor; 64ch.	link
UPenn RAM	250	1000	ECoG, sEEG	s. stim.	Language; 150ch.	link
Mayo/UPenn Seizure	82	1000	ECoG	nat. stim.	Motor; 64ch.	
Spacetop	100	600	fMRI	nat. stim.	Movies.	link
Narratives fMRI	350	350	fMRI	nat. stim.	Listening.	link
Tomcat	120	280	EEG, fNIRS	production	Gaming; Acticap 32ch.	link
Camcan	650	210	MEG	nat. stim.	Movies; resting; Elekta 306ch; phase 2 has movie data.	link
Speech EEG	1	175	EEG	production	speech; g.tec 128ch; can request data.	link
MOUS	200	160	MEG	nat. stim.	Listening; reading; CTF 275ch; open access.	link
Alljoined-1.6M	20	160	EEG	s. stim.	Images; Emotiv 32ch.	link
Omega	644	151	MEG	s. stim.	Resting; CTF 275ch.	link
WAND	170	130	MEG	s. stim.	Auditory; visual; CTF 275ch; available via openneuro.	link
HCP	90	120	MEG	s. stim.	Motor; resting; Axial grad 248ch.	link

Dataset	Subjects	Hours	Modality	Task Type	Task / Notes	Link
MEG-MASC	27	108	MEG	nat. stim.	Listening; Axial grad 157ch; publicly available.	link
Cogitate Consortium	38	100	ECoG, sEEG	s. stim.	Images; 100ch.	
HCI-SENSE-42	42	84	EEG	production	Computer-use; 32ch.	link
SMN4Lang	12	72	MEG, fMRI	nat. stim.	Listening; Elekta 306ch.	link
THINGS-EEG	10	66	EEG	s. stim.	Images; Easycap 64ch.	link
Stanford EcoG	34	60	ECoG	s. stim.	Language; motor; visual; 128ch.	link
VR Navigation EEG	60	50	EEG	production	VR; 32ch; openneuro.	link
Libribrain	1	50	MEG	nat. stim.	Listening; Elekta 306ch; public.	link
HD-EEG + mouse-tracking	31	48	EEG	production	Computer-use; 128ch.	link
EAV EEG	42	46	EEG	production	speech; 30ch.	link
cNeuroMod: Shinobi	4	40	fMRI	production	Gaming; production.	link
Tacit communication	60	40	EEG	production	Gaming; Biosemi 128ch.	link
THINGS-MEG	4	37	MEG	s. stim.	Images; CTF 275ch.	link
Moba gaming EEG	23	34	EEG	production	Gaming; 64ch; openneuro.	link
NOD	30	30	EEG, MEG, fMRI	nat. stim.	Images.	link
Narratives MEG	3	30	MEG	nat. stim.	Listening; CTF 275ch; donders on request.	link
Atari fMRI	32	30	fMRI	production	Gaming; production.	link
FilmFestival fMRI	20	26	fMRI	production	speech; spoken recall.	link
BABA MEG	30	20	MEG	nat. stim.	Movies; OPM 64ch.	link
Think-Aloud fMRI	112	20	fMRI	production	speech.	link
Handwriting	1	10	Utah	production	200ch	link