
Multimodal Foundation Agents Should Use Brain Data as Privileged Supervision

Richard Csaky
richard.csaky@gmail.com

Abstract

This position paper argues that multimodal foundation models should use brain recordings as privileged internal-state supervision during training. External behavior alone leaves human-like cognition, robustness, and reasoning weakly constrained. Neural recordings can serve as a training-time modality that shapes representations, latent-state dynamics, private computation policies such as reasoning tokens, and robustness before distillation into agents that operate from ordinary sensory, language, and action streams. We develop this position through *Foundational Human Modeling* (FHM), a program centered on scaling generative brain models, aligning heterogeneous neural modalities through shared latent geometry, inserting brain tokens into unified multimodal token streams, and testing whether the privileged channel improves distilled agents after removal. The argument is grounded in privileged-information theory and prior work showing that neural or human-derived targets can improve robustness, semantic transfer, data efficiency, and behavioral similarity after auxiliary signals are removed.

1 Introduction

Foundation agents now combine language, perception, tool use, and action. Their reliability remains uneven across task distributions and time horizons [Liu et al., 2025, Dell’Acqua et al., 2026, Ord, 2025]. One important limitation is that training data usually contains external observations or behavior traces such as text, images, videos, actions, and tool calls. It rarely contains the internal states that generated those traces. Two humans can produce the same answer with different uncertainty, attention, memory, intention, or latent belief. An agent trained only on the answer must infer these hidden variables indirectly.

Position: Brain recordings should become a training-time privileged modality for multimodal foundation agents. They can regularize internal representations, supervise latent cognitive states, and support distillation into more reliable and human-like behavior.

This position concerns human-compatible multimodal agents that must predict, communicate with, assist, and coordinate with people. Behavior and observation data alone can limit learning efficiency and leave prediction ungrounded. Neural recordings provide temporally resolved measurements of internal dynamics during perception, language, decision making, and action. They can therefore serve as privileged information in the sense of learning using privileged information (LUPI), where a training-time signal guides the model toward latent structure that is unavailable at test time [Vapnik and Vashist, 2009].

Modern multimodal models already translate heterogeneous signals into tokens and train shared sequence models over text, vision, audio, and action streams [Reed et al., 2022, Zhou et al., 2025, Wang et al., 2024, Cui et al., 2025, Bai et al., 2025b,a]. Foundational Human Modeling (FHM) extends this interface by treating neural activity as a private

training-time token stream. During training, brain tokens can be interleaved with stimulus tokens, language tokens, and output behavior tokens. Privileged distillation can then train a deployable agent that relies only on ordinary input streams. A trained brain model can also act as an offline teacher by forecasting likely neural trajectories for new stimuli, assigning uncertainty to ambiguous situations, and providing auxiliary targets for unlabeled perception or dialogue data. The deployable agent receives this benefit as a stronger human-state prior. Deployment uses ordinary sensory, language, and action streams.

If this position is accepted, multimodal foundation-model research should treat neural recordings as a useful training signal with value beyond neuroscientific analysis. That shift would affect dataset priorities, evaluation design, architectures, and training methods. It would also create a new benchmark axis for measuring whether human internal-state supervision improves agents after removal of the privileged channel and how those gains scale with the amount and type of brain recordings.

1.1 External behavior underconstrains AI agents

Behavior is an indirect and many-to-one observation of cognition. For human-aligned agents, the latent state producing the behavior is often the task-relevant target. An assistant must often infer whether a person is confused, uncertain, exploring, remembering, or planning. A model trained only on external traces may learn output correlations and miss the internal causal variables that support coordination.

This underconstraint appears in several familiar failure modes. Language can hide epistemic state because identical surface forms can arise from different underlying states, including knowledge and guessing. Current language models struggle with distinctions between belief, knowledge, and fact even when their answers are fluent [McCoy et al., 2024, Suzgun et al., 2025]. Long-horizon agent behavior creates credit-assignment ambiguity. When an agent succeeds or fails after many steps, the final trace leaves ambiguous which intermediate latent states were most useful and where additional computation would have helped. Multimodal observations confound perception, memory, language, motor planning, and social interpretation. A video of a person completing a task contains rich behavior. It leaves attention shifts, uncertainty, internal processes, and goals largely unlabeled. Inferring human biological and cognitive complexity from video alone requires stronger supervision than physical prediction from interacting objects.

Brain recordings are useful because they occupy a distinct observational position. They measure the subject’s internal response to the stimulus, task, and context. Temporally resolved electrophysiology can expose representational changes, increases in uncertainty, and response preparation. fMRI can provide slower and more spatially structured measurements of semantic and task-state variables. Both modalities provide imperfect access to cognition and can supervise hidden variables that behavior alone identifies weakly.

This argument is aligned with neuroscience-inspired critiques of pure learning. The relevant design target is human-compatible training supervision guided by useful measurements of internal state and independent of biological mechanisms or brain optimality assumptions [Zador, 2019]. Agents built to work with humans should use the best available measurements of human internal state during training. The deployed system can remain an ordinary multimodal foundation agent. The privileged channel has value when it improves the learned internal geometry, robustness, or policy after removal.

2 Related work

Learning from privileged internal signals. The FHM approach aligns with LUPI because an auxiliary channel is observed during training and removed at deployment [Vapnik and Vashist, 2009, Vapnik and Izmailov, 2015]. Brain data is a natural privileged channel because it measures internal processing induced by stimuli, goals, and actions. Prior work shows that brain-derived objectives can regularize vision representations [Fong et al., 2018, Li et al., 2019, Federer et al., 2020, Safarani et al., 2021, Shao et al., 2025, Lu et al., 2026]. Fine-tuning speech or language models with neural objectives can also improve semantic representations and data efficiency for brain alignment [Moussa et al., 2024, Moussa and

Toneva, 2025]. These results support the narrower claim that brain recordings can shape useful representations during training and improve models after the brain channel is removed.

Brain foundation models. Large neural models increasingly rely on scale, shared pre-training, tokenization, and cross-task transfer, the same ingredients that made foundation models practical elsewhere. LaBraM learns transferable EEG representations from large-scale pretraining [Jiang et al., 2024]. POYO provides a unified framework for neural population decoding [Azabou et al., 2023]. BrainOmni explicitly handles heterogeneous EEG and MEG devices through sensor-aware modeling [Xiao et al., 2025]. Generative brain modeling has begun to move from supervised decoding toward next-token prediction over quantized neural activity, including GPT2MEG, MEG-GPT, and RVQ-style multichannel timeseries tokenization [Csaky et al., 2024, Huang et al., 2025, Barmpas et al., 2025, Csaky, 2026]. Scaling-law studies for non-invasive decoding and image reconstruction suggest that data quantity, recording modality, and model size can be studied systematically as core design choices [Sato et al., 2024, Banville et al., 2025].

Brain-language alignment and decoding. Brain-aligned language modeling shows that neural signals can be connected to semantic representations learned by large models. Work on speech perception decodes linguistic content from non-invasive recordings [Défossez et al., 2023]. Language reconstruction from brain recordings demonstrates increasingly expressive generative interfaces [Ye et al., 2025]. NeuroLM, NeuGPT, and fMRI-LM learn neural tokenizers or neural-language objectives across EEG and fMRI settings [Jiang et al., 2025, Yang et al., 2024, Wei et al., 2025]. Complementary cognitive NLP work studies alignment between language-model representations and neural responses, including the effects of instruction tuning and scaling [Aw et al., 2024, Gao et al., 2025]. FHM adapts these results to training-time internal-state supervision for agents that ultimately act from ordinary input streams.

Multimodal token streams with reasoning. Generalist agents and native multimodal models show that a single sequence model can support perception, generation, and action over typed tokens [Reed et al., 2022, Zhou et al., 2025, Wang et al., 2024, Cui et al., 2025, Bai et al., 2025b,a]. Recent work on pause tokens, concurrent thinking and speaking, and latent chain-of-thought suggests that useful deliberation can occur outside the text domain [Xie et al., 2025, Kim et al., 2025, Wang et al., 2026]. Brain tokens fit this line of work as a private training-time modality that shapes latent computation before distillation.

Modality fusion. Brain recordings differ sharply in temporal resolution, spatial resolution, invasiveness, noise, and cost. This makes modality fusion a central technical problem. Sensor-aware models can handle variable EEG and MEG layouts [Xiao et al., 2025]. 3D electrode-position architectures support decoding across surface and depth electrodes [Chen et al., 2025]. Simultaneous MEG-EEG-SEEG recordings provide direct supervision for aligning invasive and non-invasive signals [Dubarry et al., 2014]. Joint MEG-fMRI modeling frames electrophysiology and hemodynamics as multi-rate observations of related latent processes [Jin and Wehbe, 2025]. Together, this work motivates a shared latent geometry as the standard interface for neural modalities.

Human-derived supervision for human-likeness. Human-likeness can be evaluated through neural predictivity, behavioral similarity, calibration, and generalization under task conditions where humans are the reference system. Brain-Score established a useful style of evaluation by combining neural and behavioral benchmarks for object recognition [Schrimpf et al., 2020]. Human behavioral supervision can also reshape foundation-model representations. Centaur fine-tunes a language model on large-scale trial-level human behavior and improves prediction of held-out human choices across experiments. Its representations also become more aligned with human neural activity [Binz et al., 2025]. Human similarity-judgment distillation improves alignment with human behavior, uncertainty, generalization, and out-of-distribution robustness in vision foundation models [Muttenthaler et al., 2025]. These results motivate brain recordings as a denser internal-state target for agent training.

3 Why brain supervision improves efficiency and human-likeness

3.1 Sample efficiency from dense hidden-state labels

Let z denote the latent cognitive state that mediates between context x and action a . This state may include attention, task set, uncertainty, memory retrieval, intention, and response preparation. A behavior-only learner observes pairs (x, a) and must infer z indirectly. A brain-augmented learner observes (x, b, a) during training, where b is a noisy measurement of neural activity generated by z . When

$$I(z; b \mid x, a) > 0,$$

each training example supplies additional information about the latent variables that produced the action. This extra information can reduce the number of behavioral examples needed to learn the same state-conditioned policy. It can also reduce credit-assignment ambiguity in long interactions by distributing useful training signal across the episode.

The learning-theoretic case follows the LUPi and generalized distillation literature. Privileged information can accelerate learning by controlling the learner’s notion of similarity and by transferring teacher-side structure into the deployed decision rule [Vapnik and Vashist, 2009, Vapnik and Izmailov, 2015, Lopez-Paz et al., 2016]. Deep LUPi experiments show larger gains in low-data regimes and report increased sample efficiency for CNNs and RNNs when training-time privileged information controls model uncertainty [Lambert et al., 2018]. Brain recordings fit this role naturally. For a single stimulus or interaction segment, they provide a time-indexed target whose geometry can encode stimulus similarity, semantic access, surprise, uncertainty, and preparation before the final response.

Brain models add a second route to efficiency. A generative or predictive brain model $q_\psi(b \mid x)$ can create auxiliary targets for unlabeled stimuli, estimate likely human internal-state trajectories, and identify cases where the predicted human response is uncertain. A teacher policy can use measured or predicted brain tokens during training. A student policy can then be distilled into the ordinary input space. At inference, the benefit appears as an amortized human-state prior. The agent should need fewer examples, or fewer exploration rollouts to infer a person’s likely goal, uncertainty, or interpretation from ordinary context. Direct evidence for open-ended foundation agents remains limited.

3.2 Evidence from neural and human-derived supervision

The strongest empirical support comes from convergent results across vision, language, and embodied navigation. Early human fMRI work showed that neurally weighted classifiers could improve object recognition and operate after the neural data stream was removed [Fong et al., 2018]. Mouse and macaque visual-cortex objectives improved robustness to noise, adversarial perturbations, and common distortions in vision models [Li et al., 2019, Safarani et al., 2021]. Further work found that training networks to mimic neural statistical structure improved object recognition and robustness to label corruption [Federer et al., 2020]. Human ventral-stream guidance produced hierarchical robustness gains and more human-like decision patterns in deep vision models [Shao et al., 2025]. A human EEG alignment framework produced vision models with stronger similarity to human EEG, cross-modal fMRI alignment, and human behavioral similarity [Lu et al., 2026]. Embodied navigation provides an agent-level example, where a brain encoder and brain-visual fusion improved robustness under visual corruptions [Peng et al., 2025].

Language results provide a complementary signal. Brain-tuning speech models with fMRI responses to stories improved alignment with new brain recordings, reduced reliance on low-level acoustic features, improved semantic downstream tasks, and increased semantic preference in the representation space [Moussa et al., 2024]. A multi-participant extension reported a five-fold decrease in the amount of fMRI data needed for new participants, up to a 50% increase in overall brain alignment, generalization to unseen datasets, and improvements on semantic tasks [Moussa and Toneva, 2025]. This is direct evidence that brain-derived objectives can improve data efficiency for brain alignment and transfer to ordinary semantic evaluation.

Human-likeness should be treated as an empirical property of behavior and internal geometry. Brain-Score evaluates models against neural measurements, behavioral measurements, and task accuracy [Schrumpf et al., 2020]. EEG-aligned ReAlnets improved behavioral similarity on object-recognition benchmarks, which links brain-supervised representation learning to human-like external behavior [Lu et al., 2026]. Centaur shows that large-scale human behavioral supervision can improve prediction of held-out human choices across domains and increase neural alignment of internal representations [Binz et al., 2025]. Human similarity-judgment distillation shows that a small amount of human-derived structure can be expanded through a teacher model and transferred into vision foundation models, improving human-aligned judgments and robustness [Muttenthaler et al., 2025]. Together these studies support a concrete prediction for FHM. Brain signals should improve human-likeness most when they expose internal states that behavior labels leave ambiguous.

4 Brain data supervision

Brain data should primarily supervise internal variables that are hard to identify from behavior alone.

Representations. Brain recordings can regularize the geometry of latent representations during perception, language understanding, and action planning. This objective requires model states that preserve task-relevant distinctions visible in human neural responses and weakly labeled in external traces. For example, a video-language agent trained with aligned video and neural responses can learn to predict both the next frame or caption and the subject’s evolving internal response to the scene. If the resulting representation transfers better to downstream tasks, the brain signal has acted as useful privileged supervision.

Uncertainty and calibration. Human internal state contains information about uncertainty, conflict, surprise, and attention allocation. Brain-supervised models can therefore be evaluated on calibration under distribution shift, and separation of unsupported guesses from well-supported inferences. These properties are especially useful for agents that interact with people in open-ended tasks.

Private computation policy. Current reasoning systems often externalize deliberation through language tokens. This supports interpretability and represents one compute-allocation mechanism among several. Brain tokens can provide a training signal for private latent computation during information integration, pausing, revision, and action preparation. Brain-token models can also supply a private deliberation channel during training. The key test measures gains in reliability, sample efficiency, and robustness against a text-only chain-of-thought baseline.

Cross-modal invariance. The same cognitive state may be observed through EEG, MEG, fMRI, ECoG, behavior, language, and task context. Each view has different noise and sampling properties. Brain data should help agents learn invariances across these views.

4.1 Unified token streams

Let $\mathcal{T}$ be the union of typed token sets across modalities and roles. For each modality $m \in \{V, S, T, B, A\}$, the roles are input $m^{(i)}$, private $m^{(s)}$, and output $m^{(o)}$. Autoregressive modeling predicts both the next role-type and the token value.

$$p_{\theta}(\tau_t \mid \tau_{<t}) = p_{\theta}(k_t \mid \tau_{<t}) p_{\theta}(\tau_t \mid k_t, \tau_{<t}),$$

where k_t indexes modality and role, such as $T^{(i)}$ and $B^{(s)}$. This lets the model learn ordering and internal compute allocation (figure 1). In continuous settings such as listening, movie watching, dialogue, visual search, and interactive control, perception, private state updates, and action can co-occur in time.

This interface supports conditioning adapters, unified interleaving, and privileged distillation. *Conditioning adapters* can map brain activity into the embedding space of a fixed language

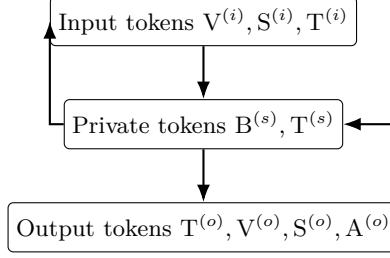


Figure 1: Typed token roles for multimodal streams. The model learns the placement of private tokens during perception and output, including brain (B), vision (V), audio (S), text (T), and action (A) tokens.

or vision-language model [Ye et al., 2025]. *Unified interleaving* can fine-tune a multimodal backbone to model joint streams containing B alongside vision, audio, text, and action tokens [Reed et al., 2022, Zhou et al., 2025]. *Privileged distillation* can train an ordinary-input policy to match an augmented policy trained with brain tokens.

$$\min_{\phi} \mathbb{E} \left[\text{KL} \left(p_{\theta}(\mathbf{a} \mid \mathbf{x}, \mathbf{z}) \parallel \pi_{\phi}(\mathbf{a} \mid \mathbf{x}) \right) \right],$$

where $\mathbf{z}$ denotes privileged neural tokens and π_{ϕ} is the deployable policy.

5 Research pillars

5.1 Long-horizon brain generation

The first priority is stable generative modeling of neural recordings. A model that produces realistic full-session rollouts is useful because rollout stability tests for durable latent dynamics and filters out models driven by short-window correlations [Csaky, 2026, Huang et al., 2025]. Key design axes include data scale and diversity, long-context sequence modeling for high-rate electrophysiology, hierarchical slow-fast latents, exposure-bias reduction, and neural tokenization [Hawthorne et al., 2022, Hiller et al., 2024, Gu and Dao, 2024, Poli et al., 2023, Bengio et al., 2015]. Discrete VQ/RVQ tokenization has been effective for audio and video generation [van den Oord et al., 2017, Défossez et al., 2024, Yan et al., 2021, Lee et al., 2022]. Neural models now need similar tokenizer approaches, potentially including soft tokenizers for mixed continuous-discrete streams [Tschannen et al., 2025, 2024, Zhou et al., 2025].

Long-horizon generation is a key target including whether generated neural trajectories preserve subject-specific, task-specific, and stimulus-conditioned structure over time. This calls for evaluations that include next-token likelihood, spectral fidelity, cross-spectral structure, drift, collapse, phase relationships, state-transition statistics, and calibration under self-generated context. The evaluation should include continuation tasks in which a model receives a long prefix from a subject and generates plausible future responses to a stimulus or task segment.

The rollout requirement also enables later “brain-space deliberation” approaches. A model that cannot generate stable latent neural dynamics is likely to turn private brain-token computation into an arbitrary latent code. Stable neural rollouts that improve downstream distillation would make long-horizon generation a foundation for agent training.

5.2 Shared geometry across recording modalities

EEG, MEG, ECoG/sEEG, and fMRI are noisy, rate-distorted measurements of related cortical processes. FHM should map sensor-space signals into a shared latent geometry whenever possible. For electrophysiology, sensor metadata such as position, orientation, type, and subject anatomy can condition models on the measurement process [Xiao et al., 2025, Chen et al., 2025]. For fMRI, slow hemodynamic observations can supervise low-frequency cognitive-state variables that fast electrophysiology models are trained to predict. Let x

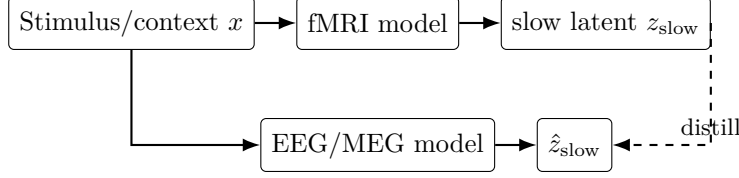


Figure 2: Multi-rate distillation learns slow cognitive states from fMRI and transfers them to faster predictors.

denote the stimulus or task context, $e_1, \dots, e_T$ the fast electrophysiology sequence, and $f_1, \dots, f_U$ the slow fMRI sequence, with $T \gg U$.

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

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

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

This multi-rate view (figure 2) connects electrophysiology and hemodynamics through separate sampling rates [Jin and Wehbe, 2025].

5.3 Scaling laws and acquisition decisions

Brain data is expensive to acquire which motivates treating scaling as an optimization problem. Let modalities $\mathcal{M}$ have per-hour acquisition costs c_m and collected hours h_m . For a budget C , the field should ask which mixture maximizes a chosen downstream utility.

$$\max_{\{h_m\}} \text{Utility}(\{h_m\}) \quad \text{s.t.} \quad \sum_{m \in \mathcal{M}} c_m h_m \leq C.$$

The utility should include neural decoding accuracy, rollout stability, cross-subject transfer, modality robustness, and downstream agent performance after distillation. Existing scaling-law work suggests that modality, data volume, and model size can be studied quantitatively [Sato et al., 2024, Banville et al., 2025]. FHM should extend this analysis to acquisition choices, including low-cost EEG across many subjects, high-quality MEG sessions, invasive recordings in clinical settings, simultaneous multimodal sessions, and fMRI for slow semantic-state supervision.

Proxy experiments can reduce wasted collection. Before acquiring a new modality at scale, studies can be run to corrupt or downsample existing modalities, inject controlled noise into token streams, or simulate missing channels. If model utility is sensitive to a controlled proxy for signal quality, cleaner or richer acquisition may be justified. Fast saturation of utility would point to bottlenecks in architecture, tokenization, task design, or distillation.

5.4 Typed interleaving and privileged distillation

Our central claim is that brain data can be incorporated into unified multimodal token streams as a private, typed, training-time modality and then distilled into deployable agents. Static datasets may look like stimulus $\rightarrow$ brain response. Interactive settings differ because video, speech, gaze, action, and brain state evolve together. Typed interleaving preserves temporal causality by allowing input, private, and output tokens to alternate.

Large behavior-rich brain datasets remain limited. Stimulus-driven recordings such as listening and movie watching are more common and potentially useful in the short-term. A staged path can start with streams such as

$$[T_i] \rightarrow [B_e] \rightarrow [B_r] \rightarrow [B_d] \rightarrow [T_o],$$

where T_i and T_o are input and output language tokens, B_e are encoding-state tokens, B_r are private reasoning-state tokens, and B_d are decoding-state tokens. Existing listening datasets can train the encoding and forecasting parts even when overt language production is absent. Video-first corpora provide a near-term vision-language route through naturalistic video

viewing, stimulus-to-brain forecasting, cross-modal retrieval, and privileged distillation into video-language agents.

Interleaving also clarifies dataset and task design. A useful dataset preserves causal timing across perception, neural-state change, action, and external feedback. Dialogue and assistance datasets should include context, user behavior, and neural response at compatible temporal resolution. Naturalistic video datasets should align frames, audio, gaze when available, and neural activity across the recording.

During training, the augmented model may use brain signal b as an extra input.

$$p_{\theta}(a \mid x, b).$$

The deployable model then approximates the useful behavior from context alone.

$$\pi_{\phi}(a \mid x) \approx \bar{p}_{\theta}(a \mid x), \quad \bar{p}_{\theta}(a \mid x) := \mathbb{E}_{b \sim p(b \mid x)} [p_{\theta}(a \mid x, b)] = \int p_{\theta}(a \mid x, b) p(b \mid x) db.$$

A clean ablation compares a distilled brain-augmented policy with a matched policy trained on the same non-neural data and compute budget.

6 Evaluation and predictions

FHM should be judged by tests that isolate the value of privileged internal-state supervision. We propose four evaluation families covering likelihood and long-rollout stability for neural generation, stimulus-conditioned forecasting and calibration, cross-modal translation under sensor dropout and device shift, and downstream agent robustness on long-horizon tasks after removal of the privileged channel [Xie et al., 2024, Yue et al., 2025]. Splits should include unseen subjects, unseen devices, unseen tasks, and corrupted or missing modalities.

The main predictions are as follows.

1. Brain-token supervision should improve robustness under distribution shift relative to matched stimulus-only training.
2. Distilled brain-augmented policies should outperform non-augmented policies on tasks requiring human goal, belief, or uncertainty inference.
3. Shared-geometry modality fusion should improve cross-device and cross-subject generalization relative to modality-isolated models.
4. Private-token ablations should reveal whether neural tokens reduce reliance on visible text-token reasoning loops.
5. Proxy corruption experiments should forecast returns from cleaner neural modalities before running expensive acquisitions.

A rigorous benchmark should include matched baselines. The comparison should use brain-augmented training, equally scaled stimulus-only training, behavior-only training, and multimodal training under the same compute budget. Agent evaluations should emphasize tasks where human internal state is especially useful. Examples include following ambiguous instructions, recognizing when a user lacks knowledge, coordinating in partially observed environments, predicting goal changes from context, and maintaining reliability over long software or web tasks [Collins et al., 2024, Ying et al., 2025, Xie et al., 2024]. For each task, the key comparison is whether privileged supervision improves the deployable model after removal of the privileged channel. For an initial brain datasets registry please see Appendix A.

7 Alternative views

Behavioral data may be enough. A reasonable objection is that internet-scale behavior already contains the cognitive traces needed for agents, including explanations, demonstrations, preference labels, and interaction logs. Behavior is a high-level output channel. It often omits intermediate state, uncertainty, timing, and latent dynamics. It can also collapse distinct cognitive processes into the same observable action.

Brain data is noisy, private, expensive, and hard to scale. This objection is correct and should shape research paths. FHM uses neural data as training-time privileged information. Deployment uses ordinary input streams. Staged distillation, subject/device holdouts, modality-dropout tests, and proxy scaling experiments can estimate when extra acquisition is justified.

Current neural decoding and generation are too weak. Current systems are early, especially for long-horizon generation and open-ended behavior, in part due to data scale. That is why FHM emphasizes intermediate milestones such as rollout stability, cross-subject generalization, sensor dropout, modality fusion, and downstream robustness after distillation. The program should be evaluated by improvements under scale.

Brain tokens may create false alignment or privacy risks. Brain grounding alone cannot establish alignment. Matching neural dynamics can coexist with harmful objectives or leakage of sensitive information. Brain-token models therefore need explicit consent, purpose limitation, licensing restrictions, de-identification, re-identification testing, privacy-preserving training where possible, and privacy evaluation before public release of model weights or derived representations.

8 Discussion

8.1 Ethics and privacy

Neurophysiological data can carry subject-identifying signatures and unintended correlates of health, identity, affect, or behavior. FHM should therefore be pursued only under ethics oversight, explicit informed consent, and release practices that classify neural data as sensitive human-subject data. Models should be evaluated for memorization and re-identification risk before release. Where possible, public benchmarks should expose privacy-screened evaluation interfaces and avoid releasing raw recordings or unrestricted checkpoints.

Consent must also cover derived uses. A participant who agrees to a decoding study may have agreed only to that study, with no consent for training a general-purpose agent or releasing a neural tokenizer. Release decisions should distinguish raw data, tokenized data, trained encoders, synthetic neural rollouts, and downstream distilled agents. Each artifact has different privacy and misuse risks.

There is also an equity concern. If high-quality brain data is collected from narrow demographic groups, brain-supervised agents may inherit a narrow model of human internal state. This leaves room for restricted release and carefully scoped datasets. Privacy may require restricted access. Claims about human-compatible agents should be tested under demographic, clinical, linguistic, and cultural variation where consent and governance permit.

8.2 Conclusion

Multimodal foundation agents should use brain recordings as privileged internal-state supervision during training. This position turns brain data from a simple decoding target into a useful signal for learning latent cognitive structure. FHM makes the claim actionable through long-horizon brain generation, shared latent geometry across recording modalities, typed interleaving of brain, stimulus, language, and action tokens, and privileged distillation into agents that operate from ordinary input streams at test-time. The decisive empirical questions are whether the privileged channel improves robustness, calibration, cross-subject transfer, and long-horizon task performance.

References

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns LLMs to the human brain. In *Conference on Language Modeling (COLM)*, 2024. doi: 10.48550/arXiv.2312.00575. URL <https://openreview.net/forum?id=moBXgN6J6V>.

- Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL <https://openreview.net/forum?id=sw2Y0sirtM>.
- 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, 2025a. URL <https://arxiv.org/abs/2511.21631>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report, 2025b. URL <https://arxiv.org/abs/2502.13923>.
- Hubert Banville, Yohann Benchetrit, Stéphane d’Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity, 2025. URL <https://arxiv.org/abs/2501.15322>.
- 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>.
- Marcel Binz, Elif Akata, Matthias Bethge, et al. A foundation model to predict and capture human cognition. *Nature*, 644:1002–1009, 2025. doi: 10.1038/s41586-025-09215-4. URL <https://www.nature.com/articles/s41586-025-09215-4>.
- 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.
- 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>.
- Richard Csaky, Mats W. J. van Es, Oiwi Parker Jones, and Mark Woolrich. GPT2MEG: Quantizing MEG for autoregressive generation, 2024. URL <https://arxiv.org/abs/2404.09256>.
- 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, 2025. URL <https://arxiv.org/abs/2510.26583>.
- Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023. doi: 10.1038/s42256-023-00714-5. URL <https://www.nature.com/articles/s42256-023-00714-5>.

- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2210.13438. URL <https://openreview.net/forum?id=ivCd8z8zR2>. Featured Certification at the ICLR 2024 Journal Track.
- Fabrizio Dell’Acqua, III McFowland, Edward, Ethan R. Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Kraye, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423, 2026. doi: 10.1287/orsc.2025.21838. URL <https://doi.org/10.1287/orsc.2025.21838>.
- 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.
- Callie Federer, Haoyan Xu, Alona Fyshe, and Joel Zylberberg. Improved object recognition using neural networks trained to mimic the brain’s statistical properties. *Neural Networks*, 131:103–114, 2020. doi: 10.1016/j.neunet.2020.07.013. URL <https://doi.org/10.1016/j.neunet.2020.07.013>.
- Ruth C. Fong, Walter J. Scheirer, and David D. Cox. Using human brain activity to guide machine learning. *Scientific Reports*, 8:5397, 2018. doi: 10.1038/s41598-018-23618-6. URL <https://www.nature.com/articles/s41598-018-23618-6>.
- Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 5(11):1080–1090, 2025. doi: 10.1038/s43588-025-00863-0. URL <https://www.nature.com/articles/s43588-025-00863-0>.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. URL <https://openreview.net/forum?id=tEYskw1VY2>.
- Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, Hannah Sheahan, Neil Zeghidour, Jean-Baptiste Alayrac, Joao Carreira, and Jesse Engel. 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 8535–8558. PMLR, 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*, volume 37, 2024. URL <https://openreview.net/forum?id=5sm8YDnWvC>. NeurIPS 2024.
- Rukuang Huang, Sungjun Cho, Chetan Gohil, Oiwi Parker Jones, and Mark Woolrich. MEG-GPT: A transformer-based foundation model for magnetoencephalography data, 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. URL <https://openreview.net/forum?id=jfkFW2gCr6>.

- 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>.
- 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. doi: 10.18653/v1/2025.findings-acl.1217. URL <https://aclanthology.org/2025.findings-acl.1217/>.
- John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Lambert_Deep_Learning_Under_CVPR_2018_paper.html.
- 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, 2025. URL <https://arxiv.org/abs/2504.01990>.
- David Lopez-Paz, Leon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL <https://arxiv.org/abs/1511.03643>.
- Zitong Lu, Yile Wang, and Julie D. Golomb. Achieving more human brain-like vision via human EEG representational alignment. *Communications Biology*, 9:463, 2026. doi: 10.1038/s42003-026-09685-w. URL <https://www.nature.com/articles/s42003-026-09685-w>.
- R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, 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(41): e2322420121, 2024. doi: 10.1073/pnas.2322420121. URL <https://www.pnas.org/doi/10.1073/pnas.2322420121>.
- Omer Moussa and Mariya Toneva. Brain-tuning improves generalizability and efficiency of brain alignment in speech models. *arXiv preprint arXiv:2510.21520*, 2025. doi: 10.48550/arXiv.2510.21520. URL <https://arxiv.org/abs/2510.21520>. NeurIPS 2025.
- 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.
- Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647:349–355, 2025. doi: 10.1038/s41586-025-09631-6. URL <https://www.nature.com/articles/s41586-025-09631-6>.
- Toby Ord. Is there a half-life for the success rates of ai agents?, 2025. URL <https://arxiv.org/abs/2505.05115>.

- Jie Peng, Changde Du, Kaicheng Fu, and Huiguang He. Brainav: Incorporating human brain activity to enhance robustness in embodied visual navigation. *Science China Technological Sciences*, 68:2120405, 2025. doi: 10.1007/s11431-025-3007-6. URL <https://doi.org/10.1007/s11431-025-3007-6>.
- 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. *Transactions on Machine Learning Research*, 2022. doi: 10.48550/arXiv.2205.06175. URL <https://openreview.net/forum?id=1ikK0kHvjv>.
- Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago A. Cadena, Kelli Restivo, George Denfield, Andreas S. Tolias, and Fabian H. Sinz. Towards robust vision by multi-task learning on monkey visual cortex. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/06a9d51e04213572ef0720dd27a84792-Abstract.html>.
- 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, 2024. URL <https://arxiv.org/abs/2407.07595>.
- Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like?, 2020. URL <https://www.biorxiv.org/content/10.1101/407007v2>.
- Zhenan Shao, Linjian Ma, Yiqing Zhou, Yibo Jacky Zhang, Sanmi Koyejo, Bo Li, and Diane M. Beck. Probing human visual robustness with neurally-guided deep neural networks, 2025. URL <https://arxiv.org/abs/2405.02564>.
- 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(11):1780–1790, 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 *Computer Vision – ECCV 2024*, pages 292–309. Springer, 2024. doi: 10.1007/978-3-031-72998-0_17. URL https://doi.org/10.1007/978-3-031-72998-0_17.
- 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.
- Aaron 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 Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL <https://jmlr.org/papers/v16/vapnik15b.html>.

- 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.
- Jiecong Wang, Hao Peng, and Chunyang Liu. Latent chain-of-thought as planning: Decoupling reasoning from verbalization, 2026. 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, Qiyong Yu, et al. Emu3: Next-token prediction is all you need, 2024. URL <https://arxiv.org/abs/2409.18869>.
- 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. URL <https://arxiv.org/abs/2511.21760>.
- 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. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2505.18185. URL <https://openreview.net/forum?id=cjHQj0tCy6>. 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. Oworld: 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, 2025. URL <https://arxiv.org/abs/2508.15827>.
- Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 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. URL <https://arxiv.org/abs/2410.20916>. Findings of ACL 2025.
- 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(1):346, 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, 2025. URL <https://arxiv.org/abs/2502.20502>.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui 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 (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.736. 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. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://openreview.net/forum?id=SI2hI0frk6>. ICLR 2025 Oral.

A 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
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

Dataset	Subjects	Hours	Modality	Task Type	Task / Notes	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, 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