Multimodal Foundation Agents Should Use Brain Data as Privileged Supervision

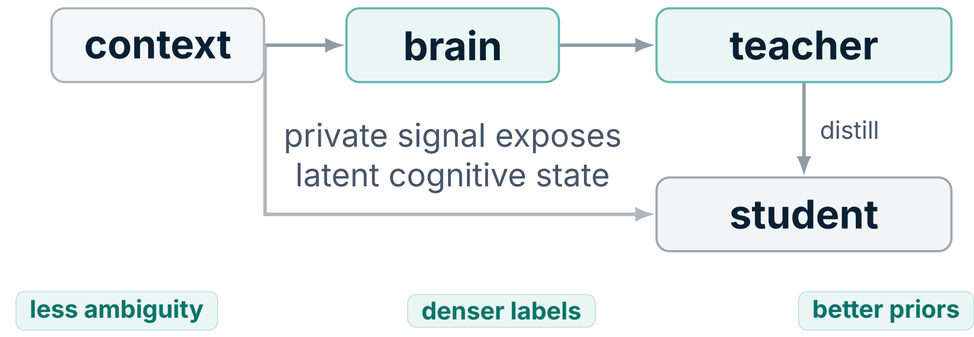
Richard Csaky Barcelona Computational Foundation | Foresight Institute

richard.csaky@gmail.com | richardcsaky.notion.site/main | ricsinaruto.github.io/docs/FHM.pdf

Behavior is not cognition

External behavior is a many-to-one readout of hidden cognitive state. **Foundational Human Modeling** (**FHM**) uses neural recordings to supervise uncertainty, attention, memory, intention, and action preparation while the model is trained.

Mechanism



Brain tokens act as privileged information during training, absent at inference.

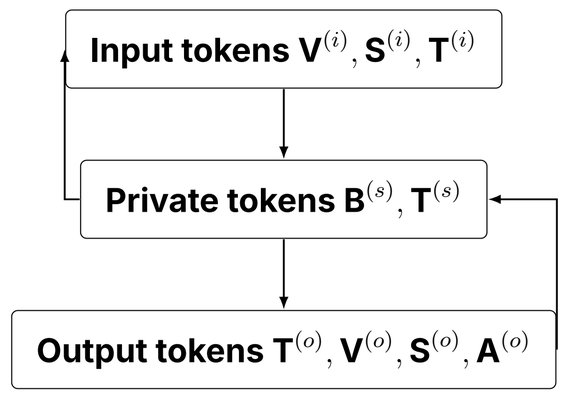
FHM recipe

1. Align context, brain, behavior streams
2. Train brain-augmented teacher
3. Distill ordinary-input student
4. Deploy without brain tokens

Criterion: gains must remain after removing the privileged brain channel, at matched data and compute.

Token stream approach

One typed sequence contains ordinary input tokens, private brain-tokens, and output tokens.



How it works

- **Input/output roles:** vision, audio, text, and action tokens.
- **Private role:** encoding state, uncertainty, internal reasoning, and action preparation.

Why it matters

- Inserts internal compute between perception and action while preserving causality.
- Supports predicted brain tokens and student-only ablations.

Research pillars

Does brain-based supervision help the ordinary-input student after the privileged channel is removed?

1. Long-horizon generative brain models
Scale next-token neural modeling to full sessions, long context, and self-generated rollouts. Design axes: neural tokenization, slow-fast latent structure, exposure-bias reduction, stability under recursive generation.

2. Shared latent geometry
Learn a common representation across EEG, MEG, fMRI, and invasive data so that modality, device, and subject become mutually beneficial. Key tests include cross-modal retrieval, transfer, and forecasting.

3. Typed token interleaving
Treat brain activity as a private typed token stream that can appear alongside vision, audio, text, and action. The sequence model learns when to observe, pause, internally compute, and emit actions while preserving temporal causality.

4. Privileged distillation
Train a teacher with privileged brain tokens, then distill an ordinary-input student. Use matched data and compute and test whether robustness, calibration, or task performance survive removal of the brain channel.

Evaluation

Matched baselines
brain-augmented vs. stimulus-only and behavior-only

Deployment input
ordinary sensory, language, action streams

Stress tests
distribution shift, sensor dropout, long rollouts

Acceptance test
student improves after brain-token removal

Predictions

1. Better robustness and calibration under shift.
2. Stronger inference of human goals, beliefs, and uncertainty.
3. Better transfer across subjects and devices.

Takeaway

FHM treats brain recordings as privileged training-time supervision for frontier AI systems, turning latent human states such as attention, uncertainty, memory, intent, and action preparation into denser learning signals. This can improve learning efficiency by reducing the ambiguity of behavioral data, while also supporting alignment by distilling more human-like priors into ordinary-input agents after the brain channel is removed.