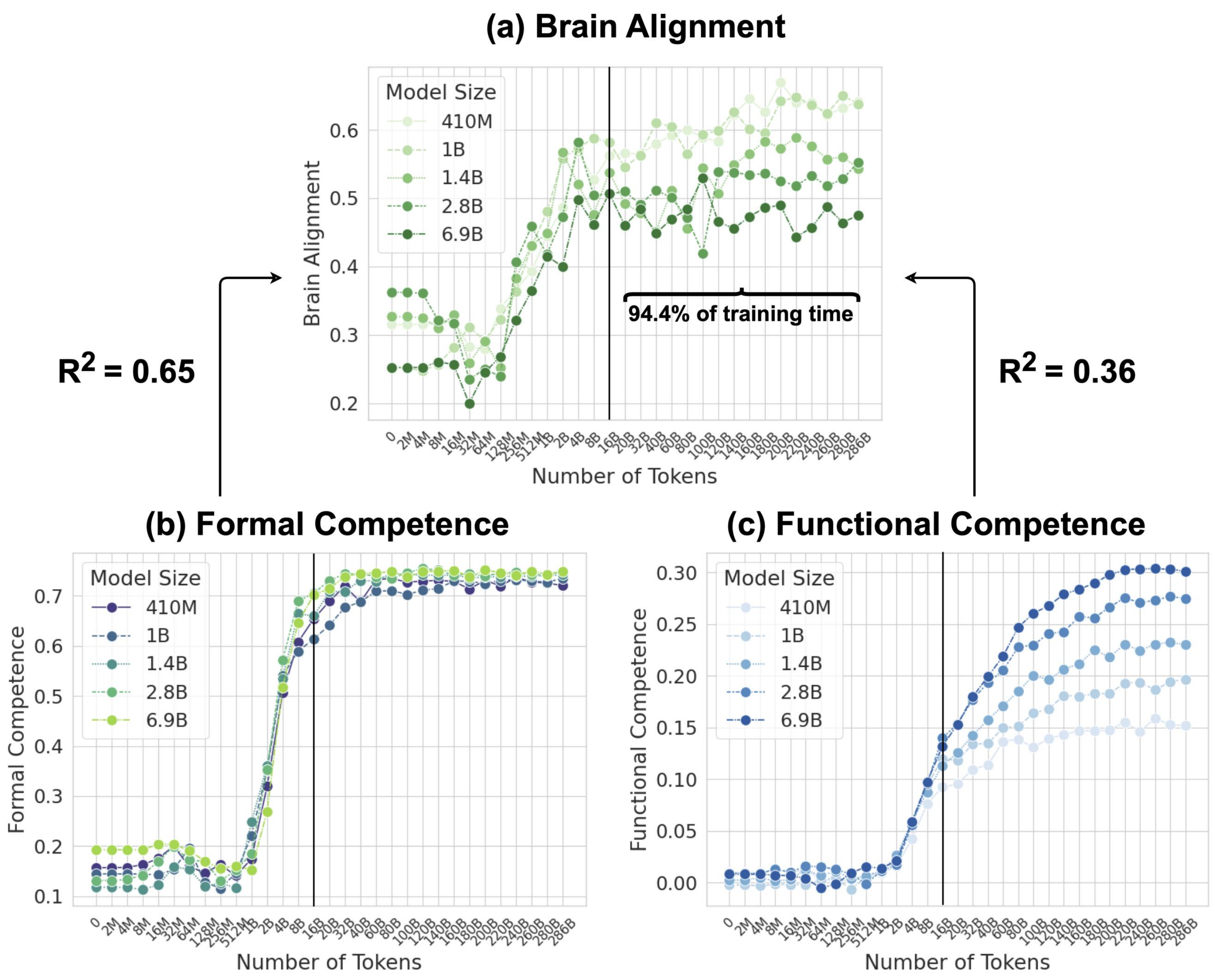
The LLM Language Network:
How LLMs Outgrow the Human Language Network

Badr AlKhamissi[1] Greta Tuckute[2] Yingtian Tang[1] Taha Binhuraib[3]
Antoine Bosselut*,[1] Martin Schrimpf*,[1]
[1] EPFL [2] MIT [3] GT
*Equal Supervision
NeuroAI Lab nlp lab
@bkhmsi

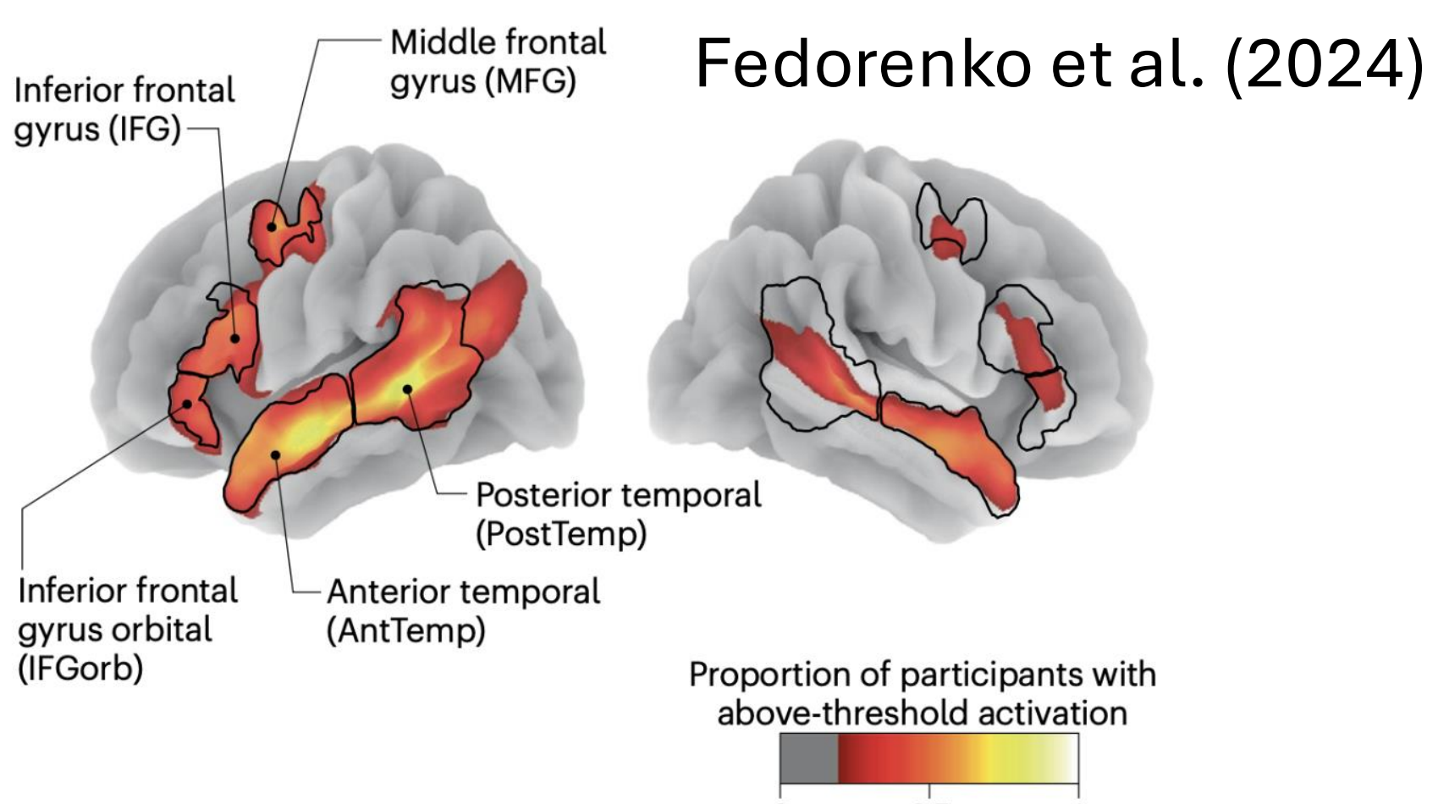
* Highlights

1. **Untrained models** align with brain via **context integration**
2. **Formal linguistic** competence drives alignment early, **saturates ~4B tokens**
3. **Functional competence** emerges later, with **weaker brain correlation**
4. **Correlation** between models' **brain alignment** and their **next-word-prediction** performance, as well as their **behavioral alignment fades over time.**
5. **Model size ≠ better brain alignment** (when controlling features).

1 Brain Alignment Tracks Formal More Than Functional Competence



* Human Language Network


Inferior frontal gyrus (IFG)
Middle frontal gyrus (MFG)
Posterior temporal (PostTemp)
Inferior frontal gyrus orbital (IFGorb)
Anterior temporal (AntTemp)
Proportion of participants with above-threshold activation
0
0.5
1

Fedorenko et al. (2024)

Specialized area within the brain responsible for understanding and producing language.

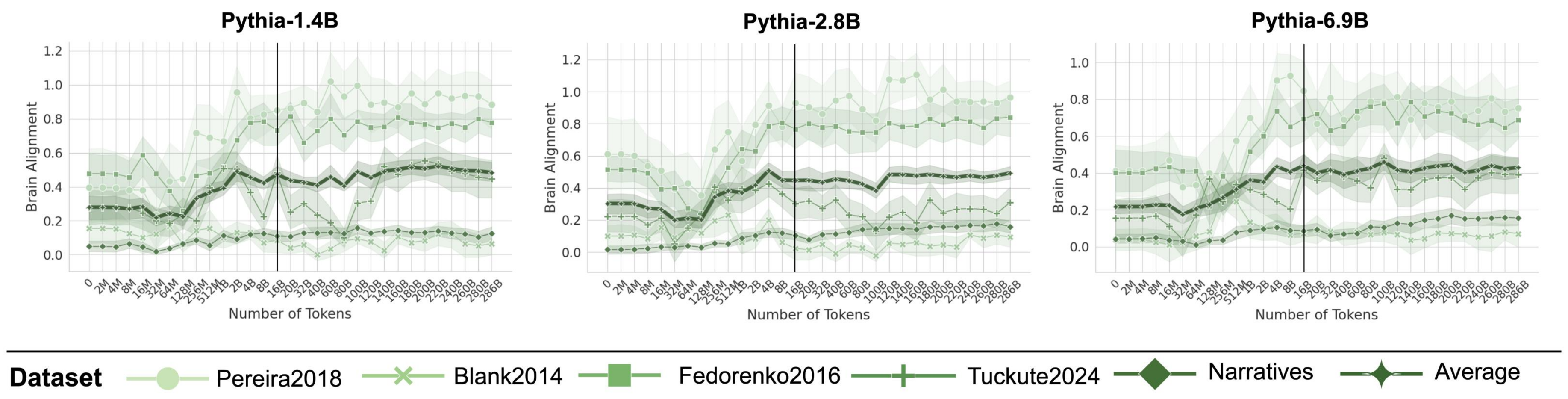
2 Methods

1. Benchmarked **34** checkpoints
2. Spanning ~**300B** tokens
3. Across **8** different model sizes
4. On **5** brain-recording datasets, and **1** behavioral dataset
5. And on **2** formal linguistic benchmarks and **6** functional

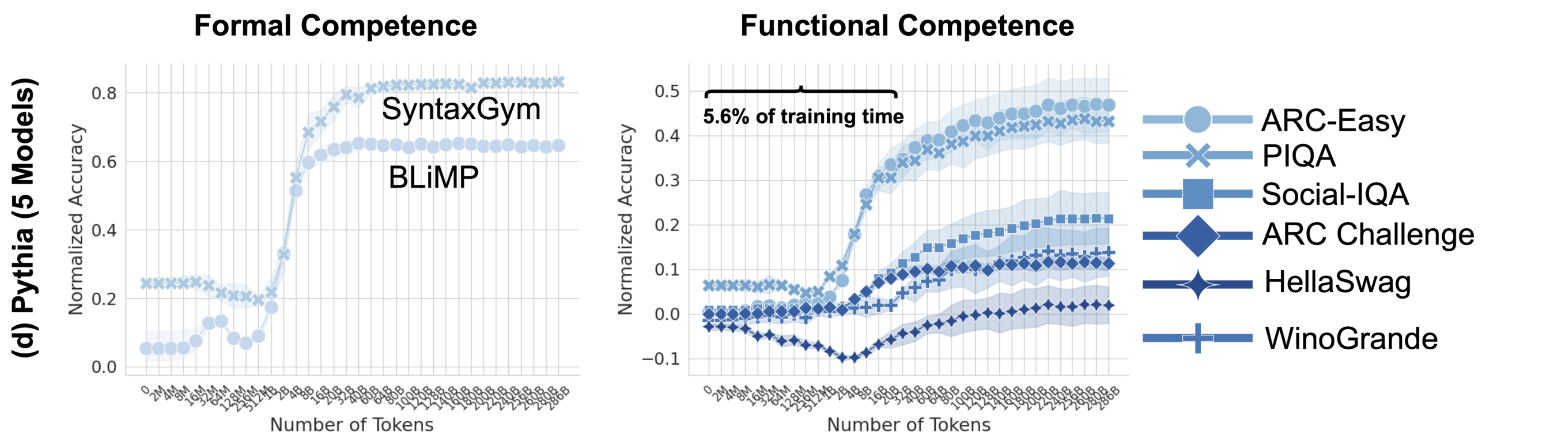
Research Questions

What drives brain alignment of LLMs?
Is it primarily linked to formal or functional linguistic competence?
Do LLMs diverge from humans as they surpass human-level prediction?

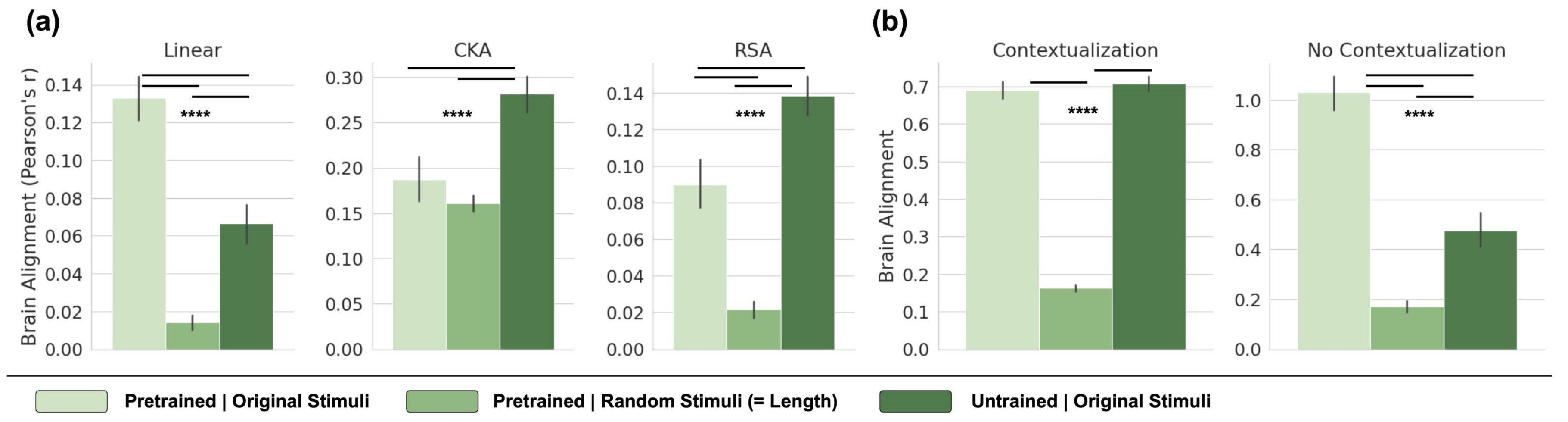
3 Brain Alignment Over Training Per Dataset



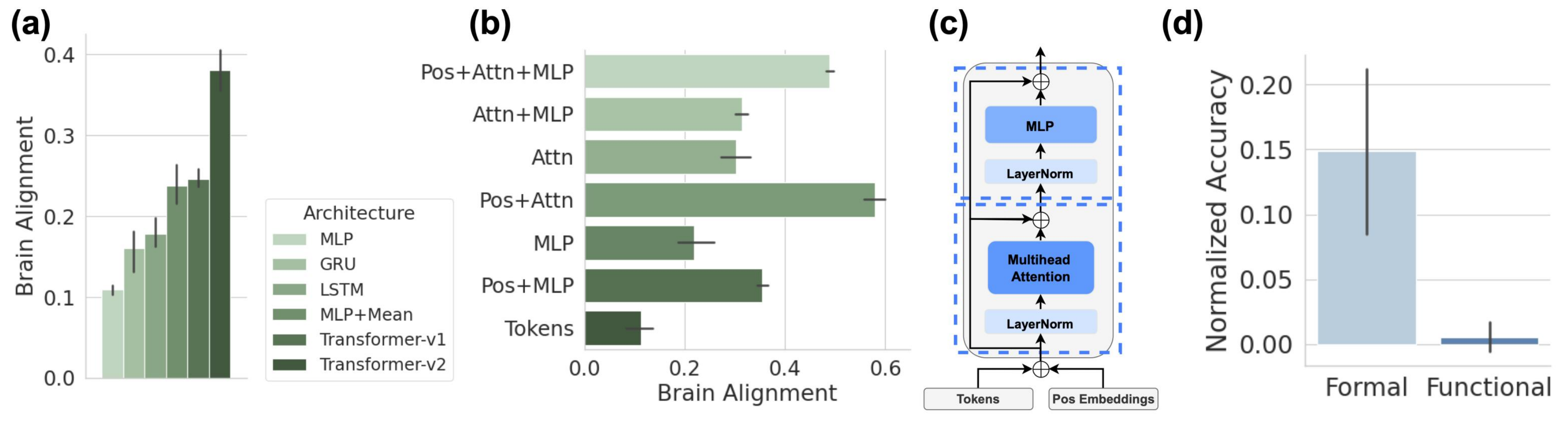
4 Formal & Functional Competence Over Training Per Dataset



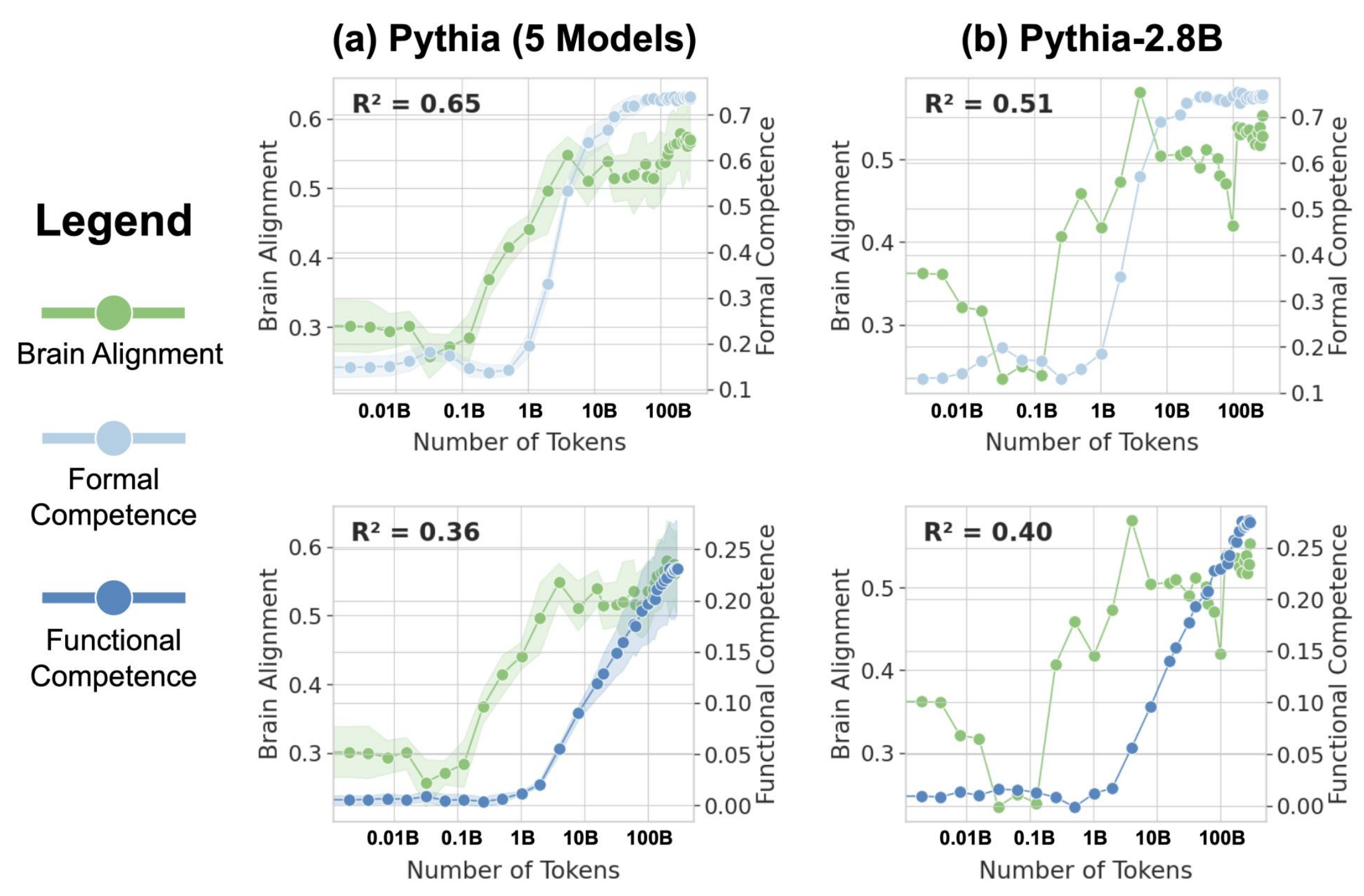
5 Rigorous Brain-Scoring



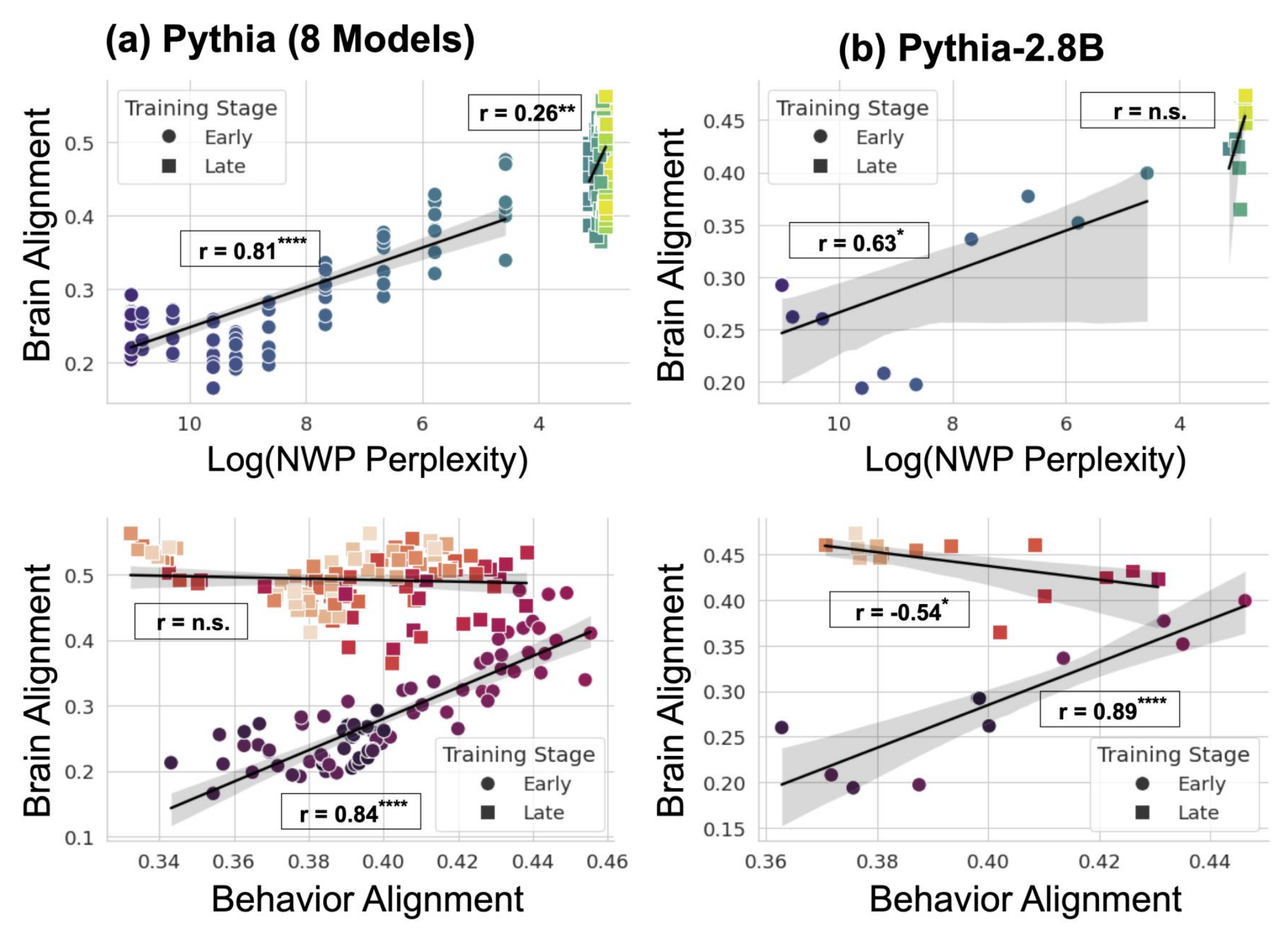
6 Context Integration drives Brain Alignment of Untrained Models



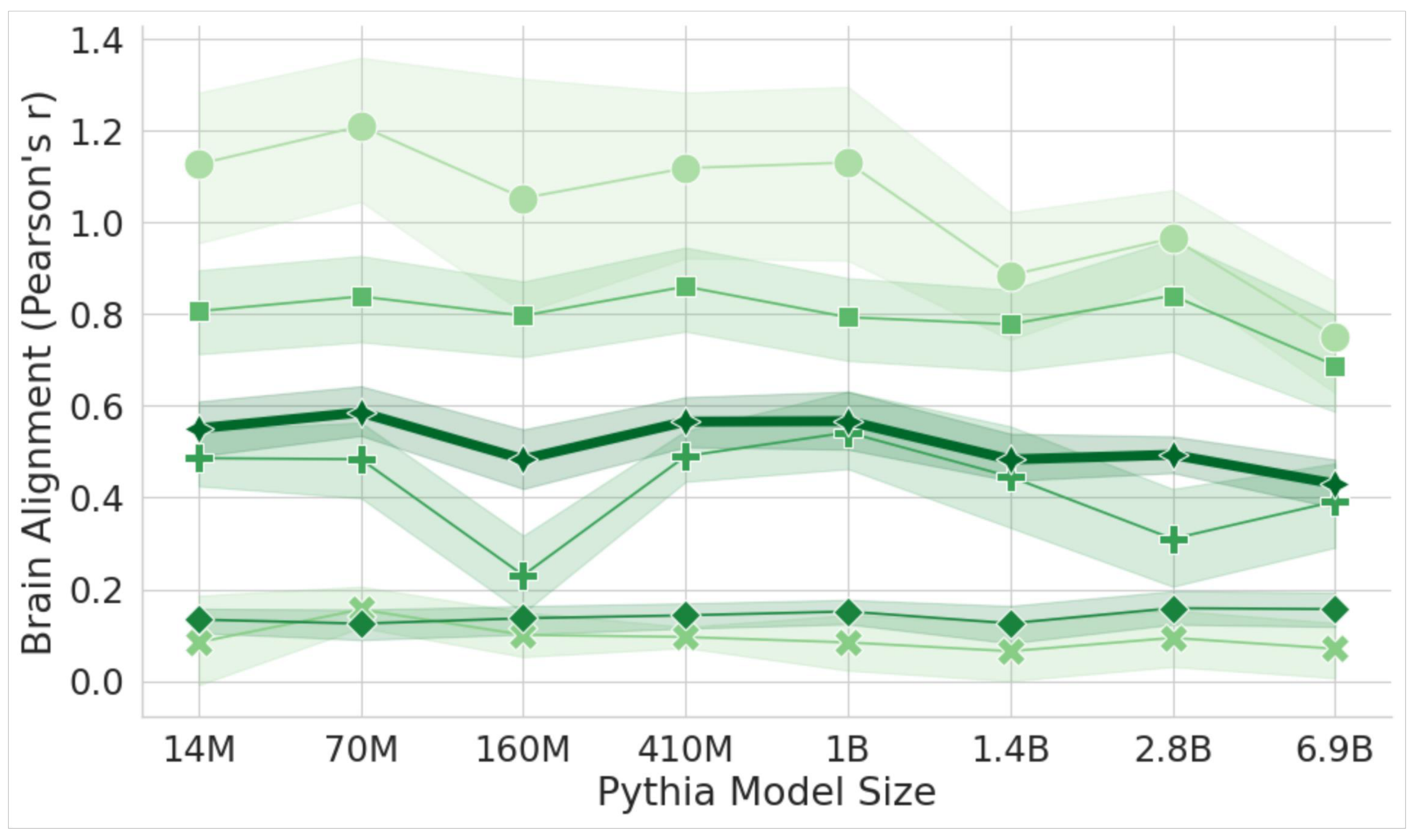
7 Brain Tracks Formal > Functional Competence



8 Brain Aligns with NWP & Behavior Early



9 Model Size ≠ Better Brain Alignment


Brain Alignment (Pearson's r)
14M
70M
160M
410M
1B
1.4B
2.8B
6.9B
Pythia Model Size