Brain-Like Language Processing via a Shallow Untrained Multihead Attention Network

Badr AlKhamissi[1] Greta Tuckute[2] Antoine Bosselut[*,1] Martin Schrimpf[*,1]

[1]EPFL [2]MIT

TL;DR

What makes representations induced by architectural priors alone exhibit reasonable alignment to brain data ??

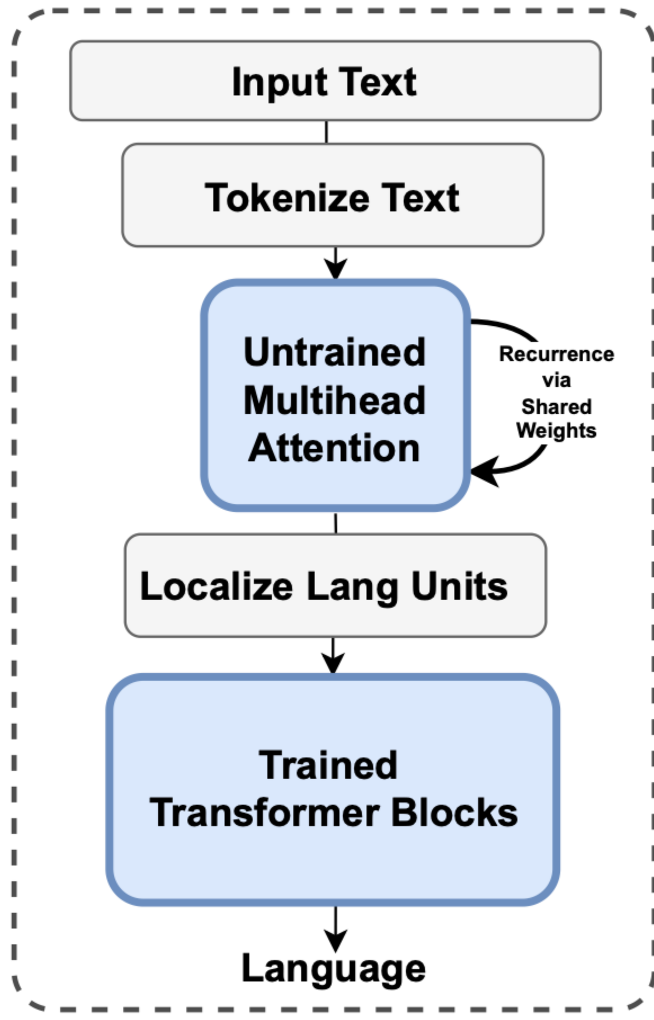
- Token Mixing (e.g., MHA)
- Tokenization Strategy (e.g., BPE)

Highlights

Using language units in a Shallow Untrained Multihead Attention (SUMA) model we achieve:

1. SoTA on brain benchmarks !
2. SoTA on behavioral benchmark !!
3. Efficient language modeling !!!

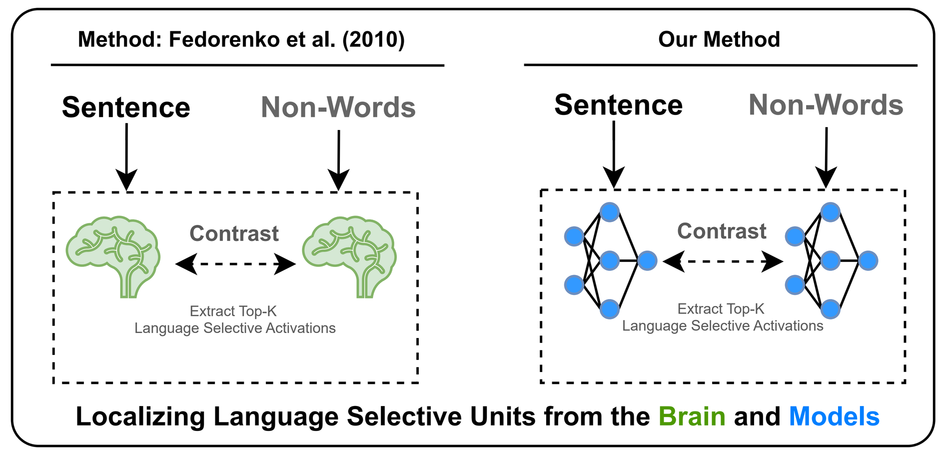
Proposed Model

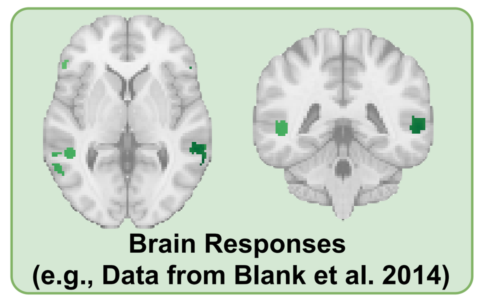


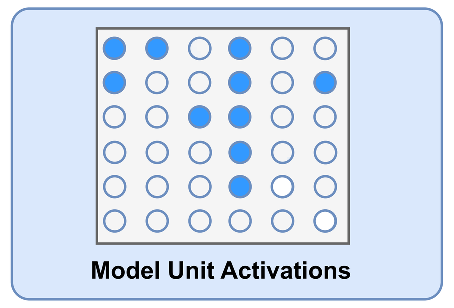
Paper Link

1 Localization

Sentence: THE DOG CHASED THE CAT ALL DAY LONG
Non-Words: BOKER DESH HE THE DRILES LER CICE FRISTY'S

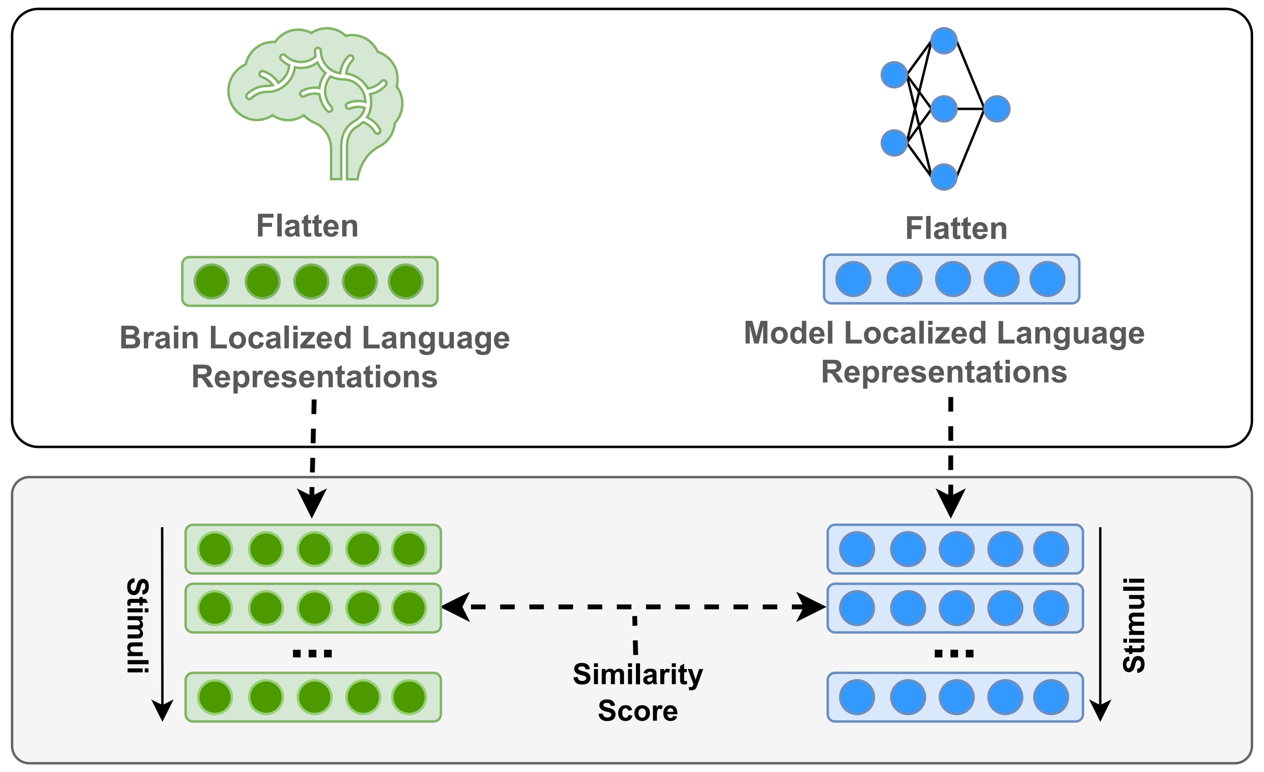




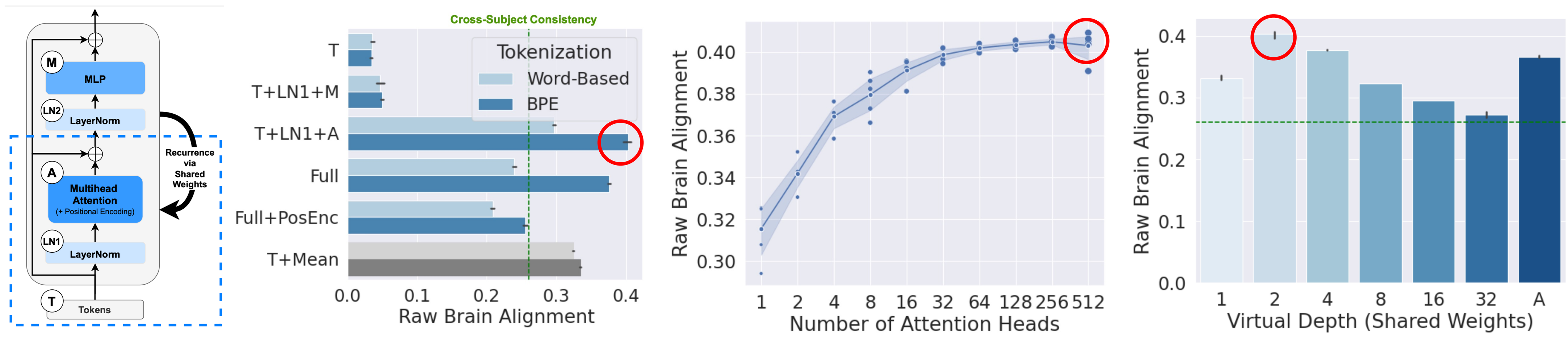


2 Benchmarking

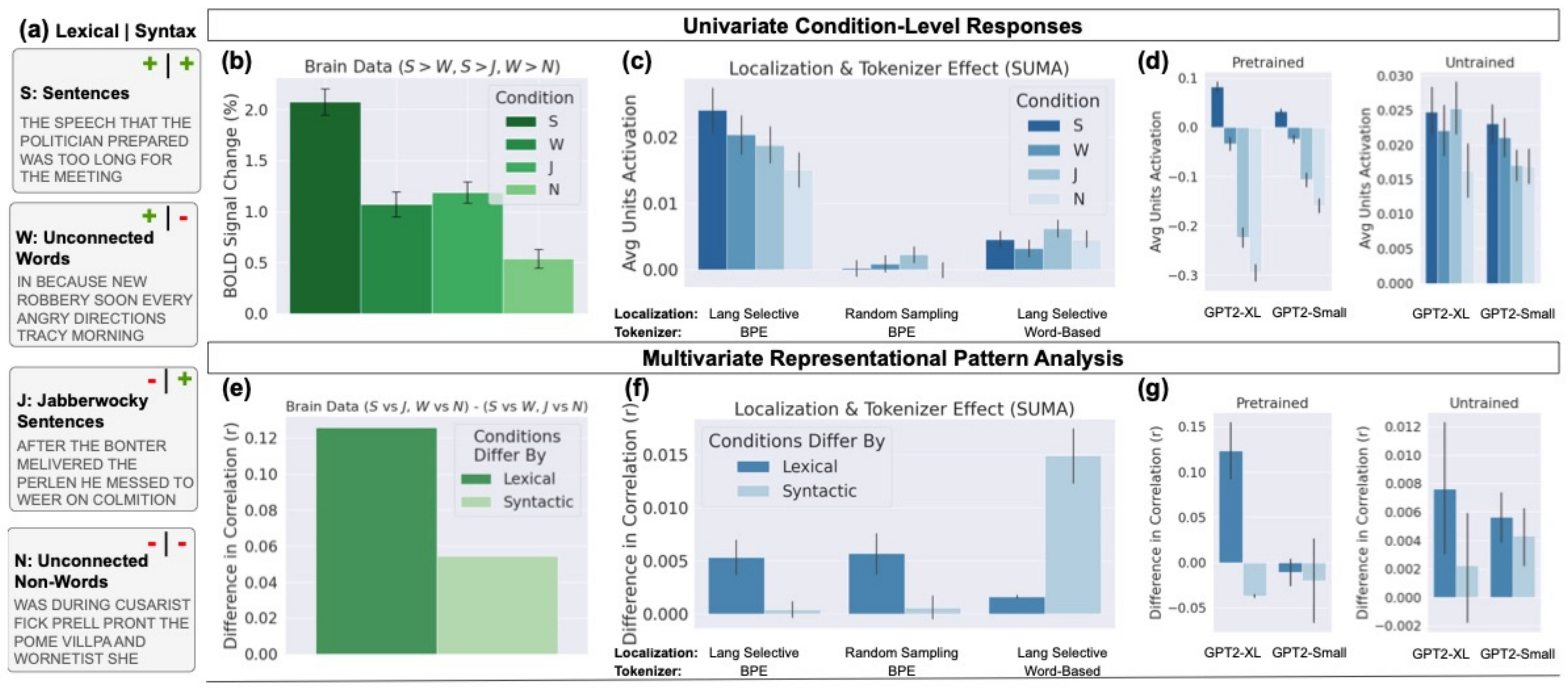
Pereira2018 (fMRI, Reading) | Blank2014 (fMRI, Listening) | Fedorenko2016 (ECoG, Reading) | Tuckute2024 (fMRI, Reading) | Wehbe2014 (fMRI, Reading)



3 Isolating Critical Components of the Transformer Architecture



4 Language Units Exhibit Similar Response Profiles as the Human Language System



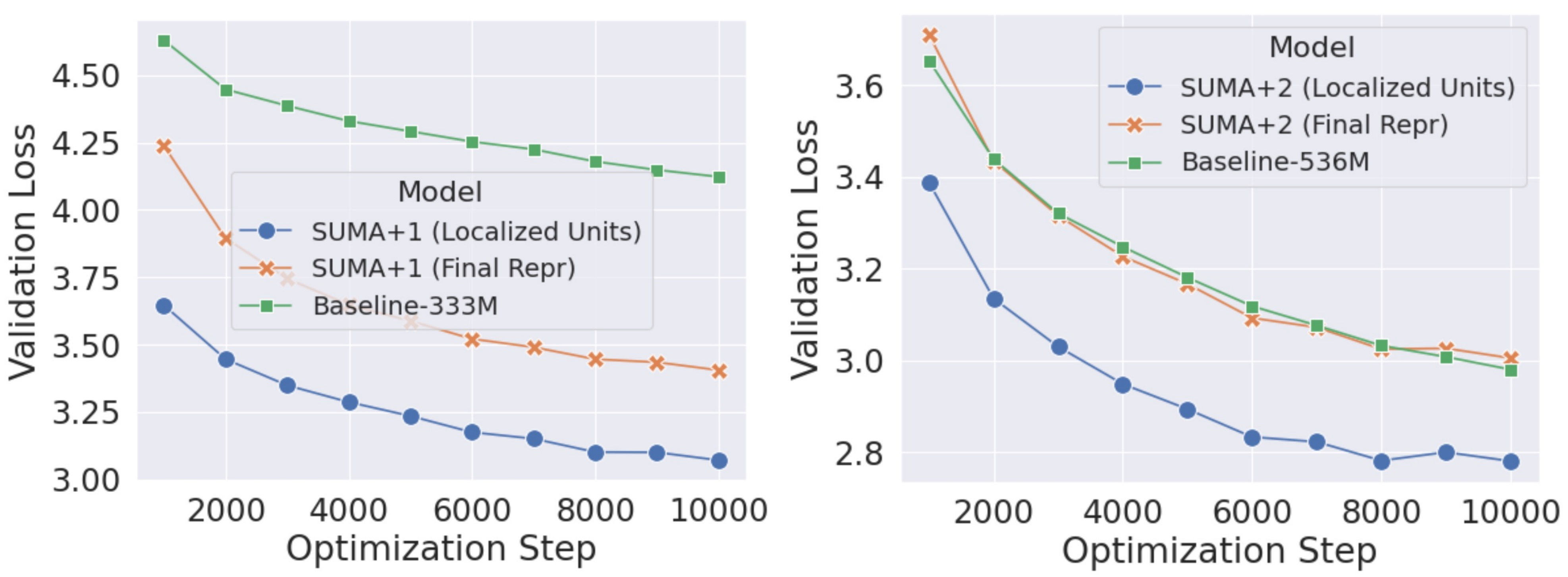
What are the key architectural components underlying the surprising brain alignment of untrained LLMs?

NeuroAI Lab

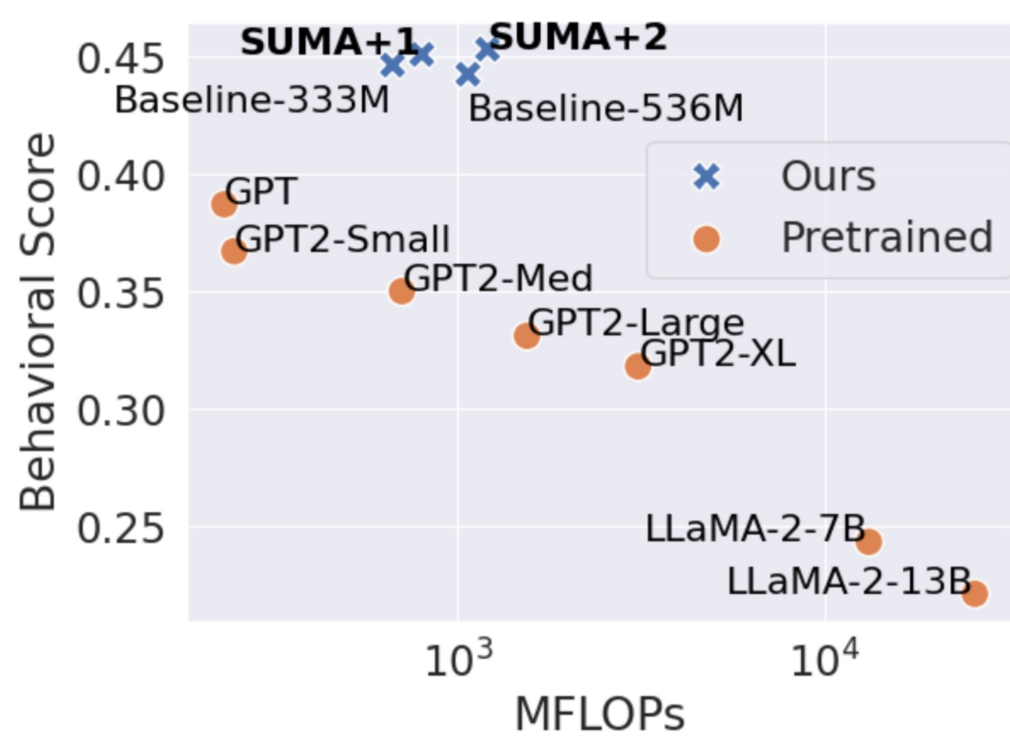
*Equal Supervision

CCN

5 Language Units Improves Language Modeling and Sample Efficiency



6 SoTA on Human Reading Time Alignment



7 Language Units Brain Alignment Scores in Pretrained and Untrained Models

Brain-Score

Model (MFLOPs)	Pereira2018	Blank2014	Fed2016	Tuckute2024	Wehbe2014	Average
GPT2-Small (170)	0.38/0.16	0.10/0.05	0.27/0.27	0.29/0.21	0.11/0.05	0.23/0.15
GPT2-Med (604)	0.38/0.16	0.10/0.04	0.29/0.26	0.37/0.19	0.11/0.05	0.25/0.14
GPT2-Large (1,420)	0.39/0.16	0.09/0.05	0.30/0.25	0.32/0.21	0.08/0.04	0.23/0.14
GPT2-XL (2,950)	0.34/0.15	0.04/0.04	0.27/0.25	0.34/0.23	0.04/0.04	0.21/0.15
LLaMA-2-7B (12,950)	0.32/0.32	0.01/0.24	0.22/0.34	0.34/0.13	0.02/0.15	0.18/0.24
LLaMA-2-13B (25,380)	0.41/0.28	0.04/0.14	0.26/0.32	0.34/0.17	0.06/0.09	0.22/0.20
SUMA (268)	- / 0.43	- / 0.44	- / 0.34	- / 0.19	- / 0.21	- / **0.32**