ICML
International Conference
On Machine Learning

Linguistic Properties and Model Scale in Brain Encoding: From Small to Compressed Language Models

Subba Reddy Oota Vijay Rowtula Satya Sai Srinath Khushbu Pahwa
Anant Khandelwal Manish Gupta Tanmoy Chakraborty Bapi S. Raju

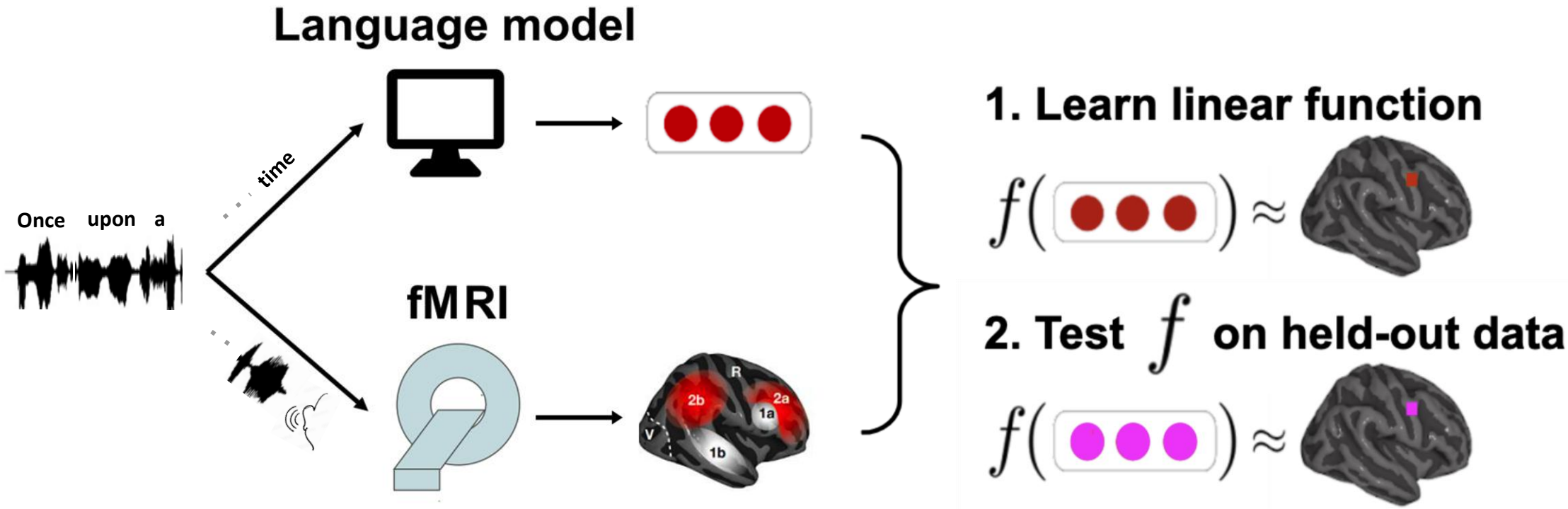
Technische Universität Berlin

GE HealthCare

Microsoft

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



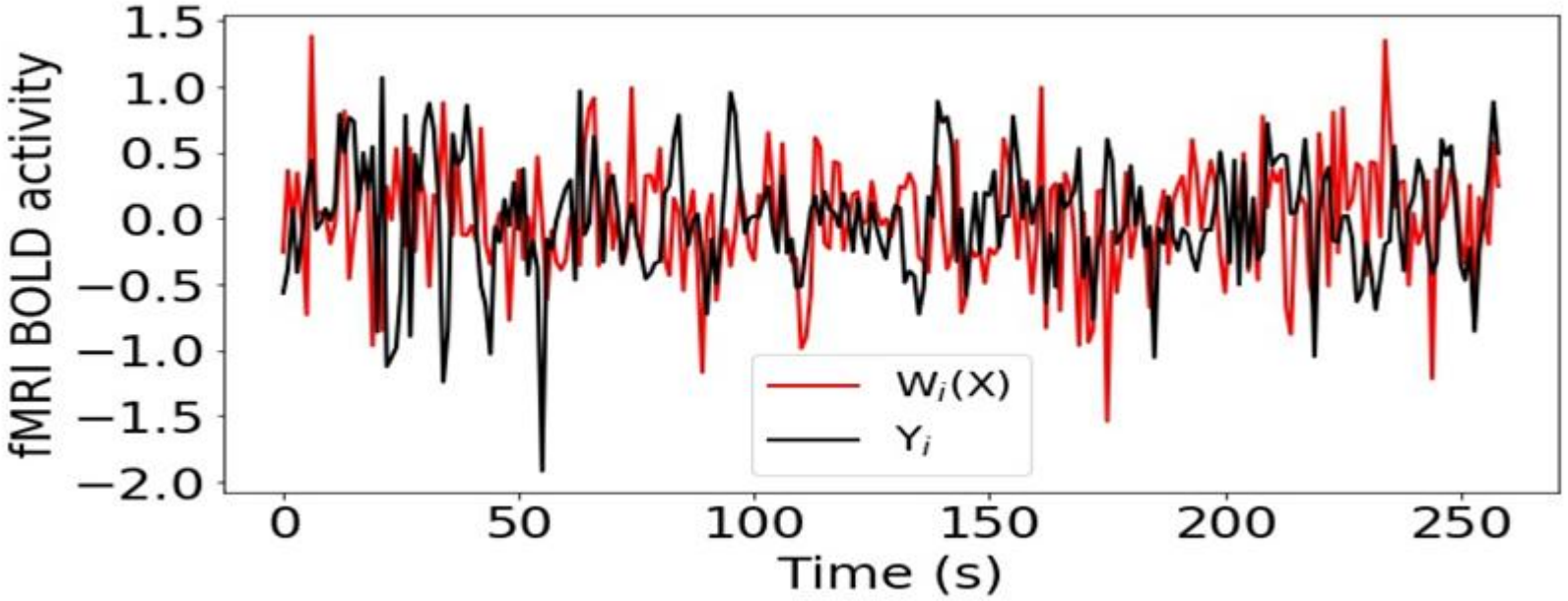
Brain alignment of a LM ⇒ how similar its representations are to a human brains

Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

Toneva and Wehbe 2019,
Caucheteux et al. 2020,
Toneva et al. 2020

Jain et al. 2020,
Schrimpf et al. 2021,
Goldstein et al. 2022
...

Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree

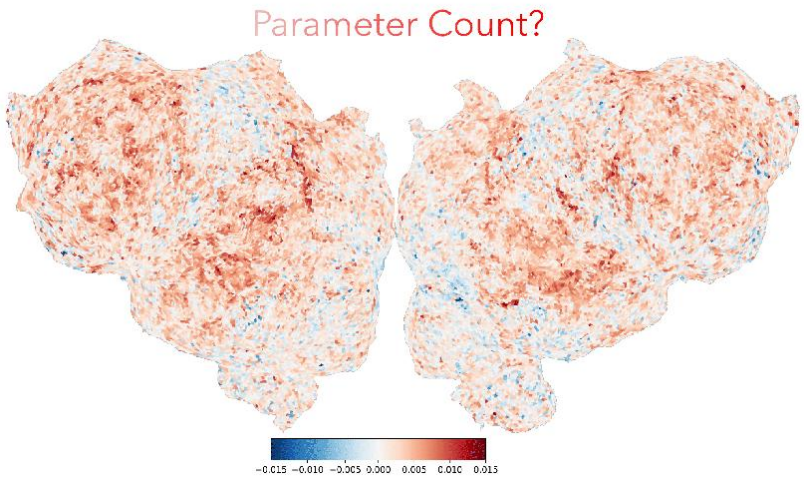


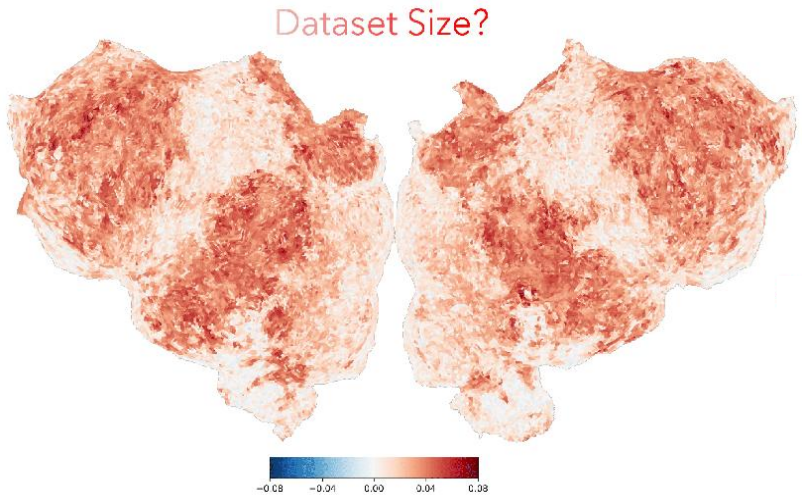
brain alignment$_i$ = Pearson corr(true v_i, pred v_i)

Brain alignment of a LM ⇒ does better brain alignment demand scaling of language models?

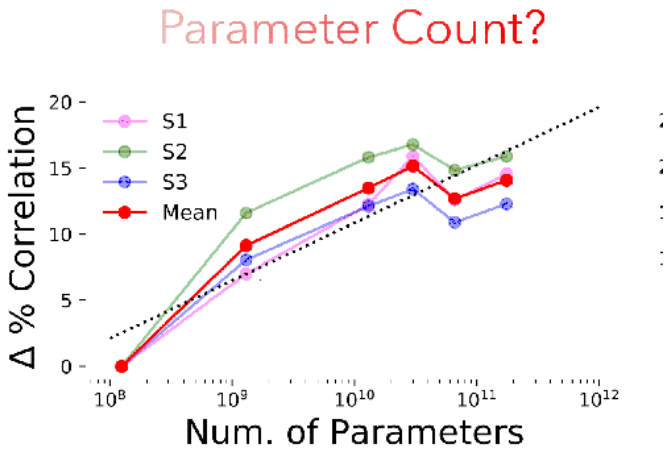
Jain and Huth. Incorporating context into language encoding models for fMRI. (NeurIPS 2018)
Toneva and Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). (NeurIPS 2019)

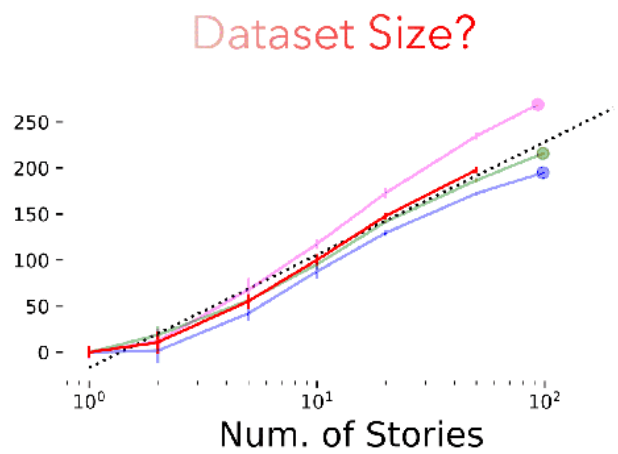
LMs predict brain activity, and that alignment improves as models scale, suggesting neural scaling laws





Antonello et al. 2023 NeurIPS





Larger models, better brain alignment, but harder to interpret and costly. How small can we go?

How small can we go? Three questions about scale, compression, and the brain

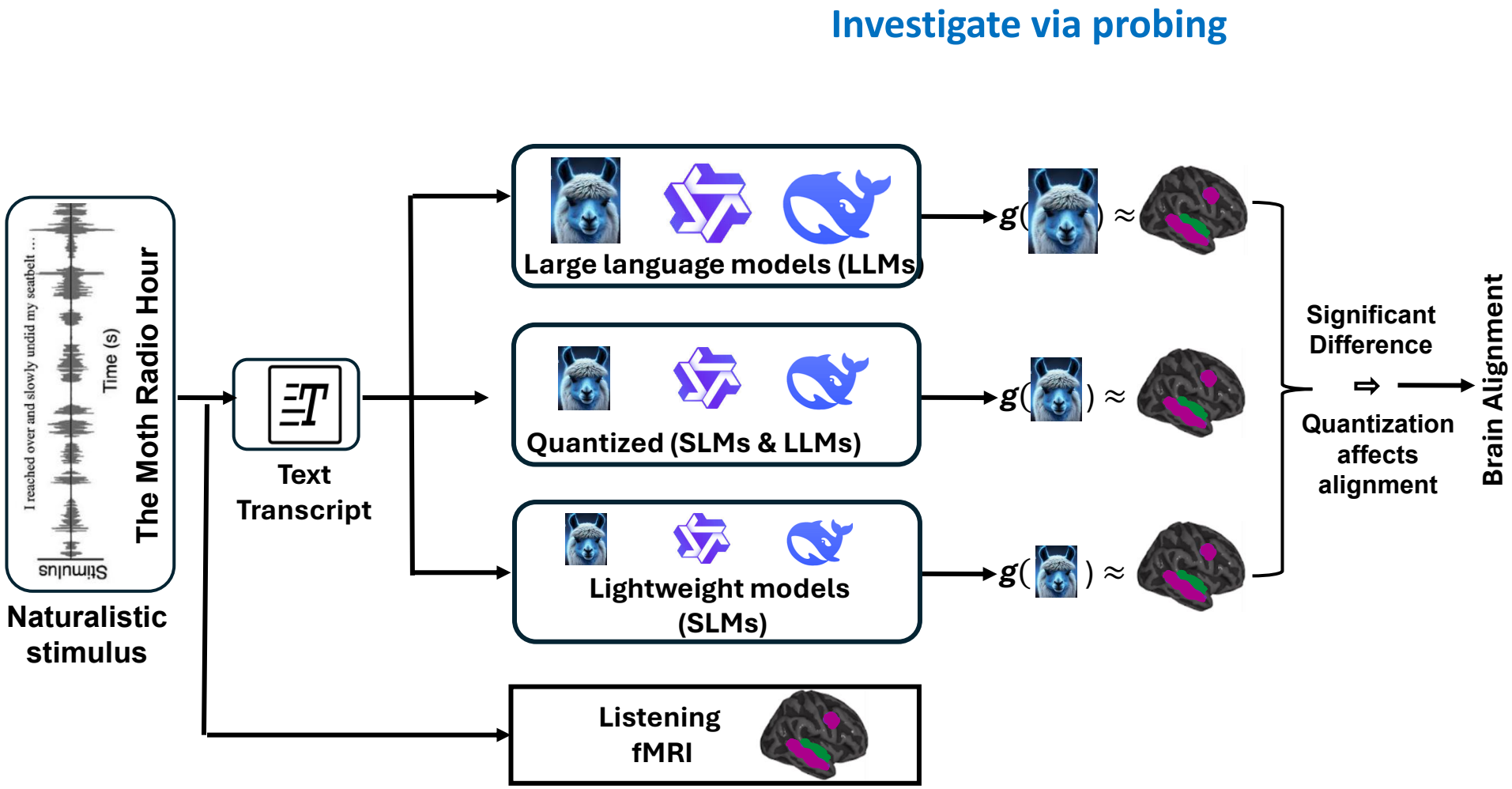
1. What's the **minimal model capacity** that matches **large LLMs** in brain alignment, for both encoding and decoding?

2. How do **compression methods (quantization, pruning)** affect brain alignment in small and large models?

3. Which **linguistic properties** survive or degrade across small/large/compressed models, and do those changes impact brain alignment?

Use model size and precision as controlled "knobs" to probe what representations the brain needs.

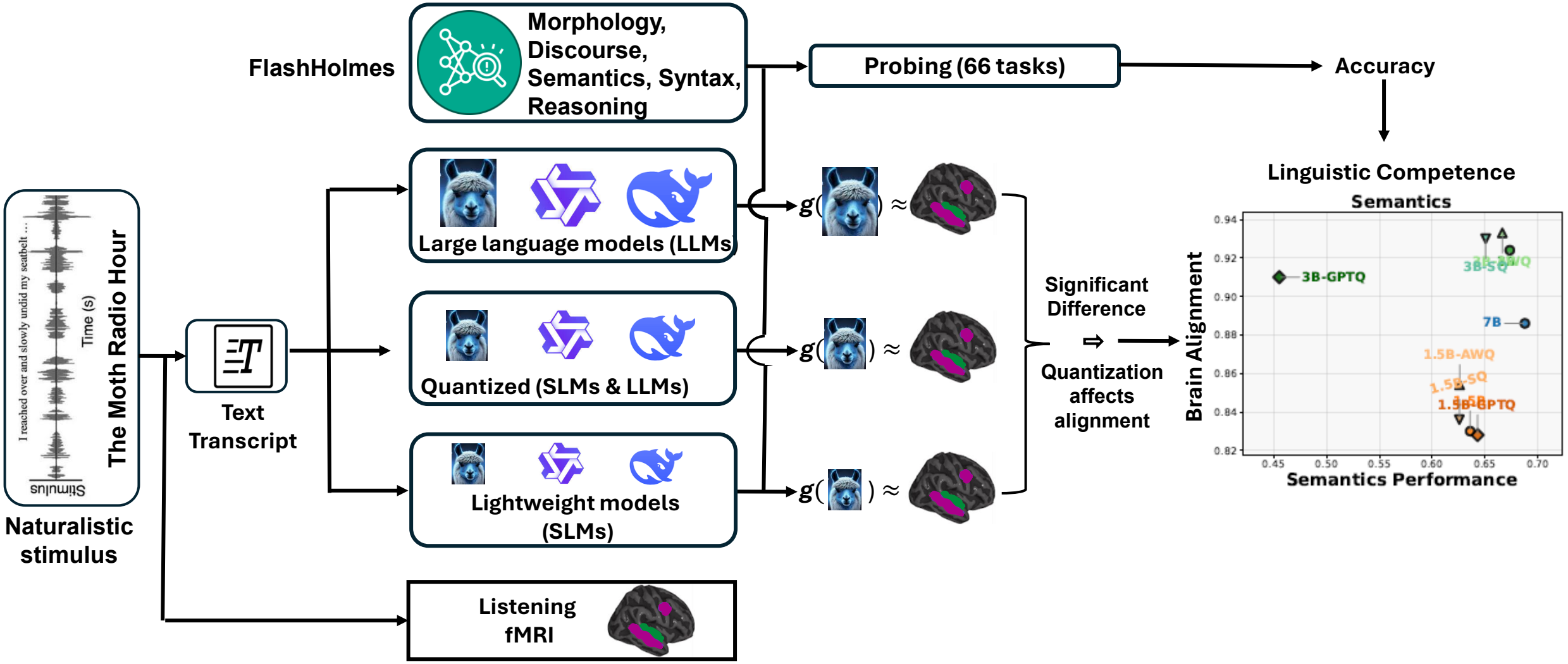
Two knobs: model scale and numerical precision



Brain alignment shifts with scale and compression - does linguistic competence shift the same way?

Does linguistic competence drive brain–model alignment?

Investigate via an indirect approach

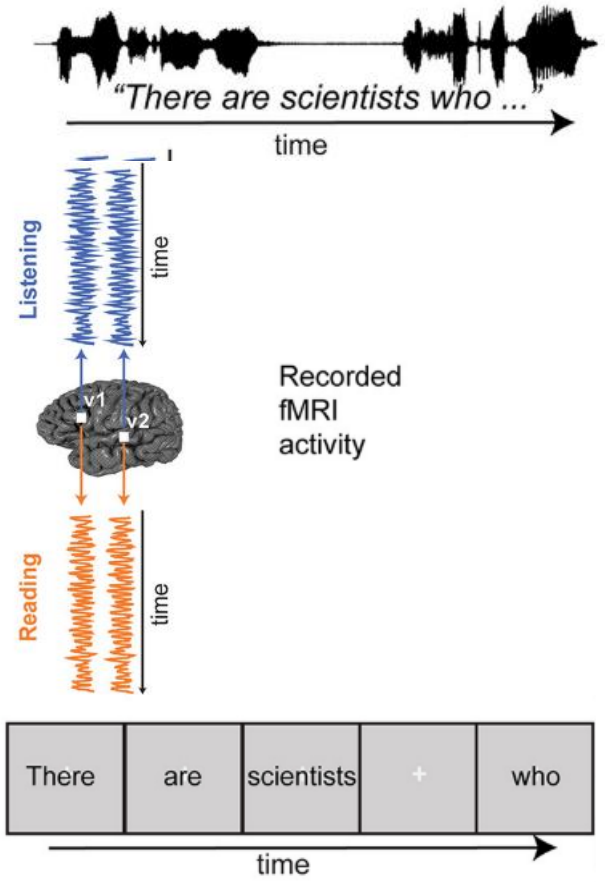


Datasets & Model

- Brain: fMRI recordings from Subset-Moth-Radio-Hour [Deniz et al. 2019]
  - Reading & Listening to the same short stories
  - N=9

- 3 language model families
  - Qwen-2.5 (1.5B, 3B, 7B, 14B)
  - LLaMA-3.2 (1B, 3B, 8B, 14B)
  - DeepSeek-R1 (1.5B, 3B, 7B, 14B)

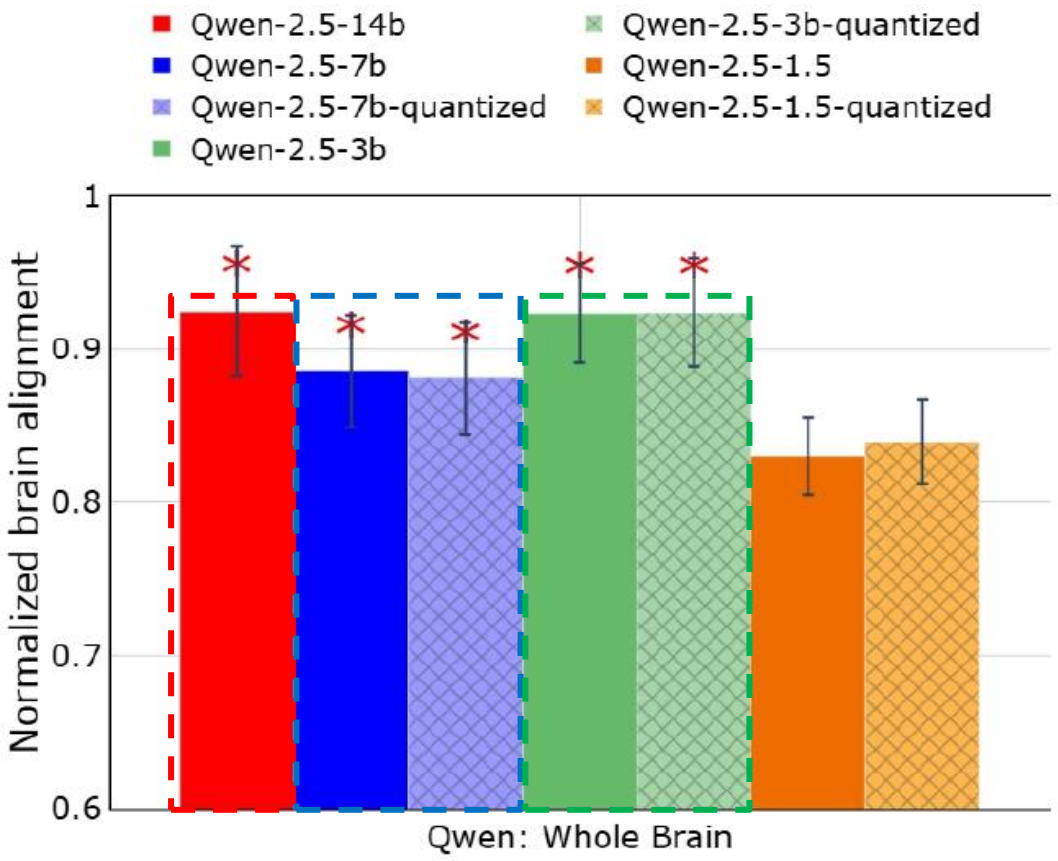
- 2 compression methods
  - Quantization (AWQ, GPTQ, SmoothQuant)
  - Unstructured Pruning (10%, 25%, 50%)

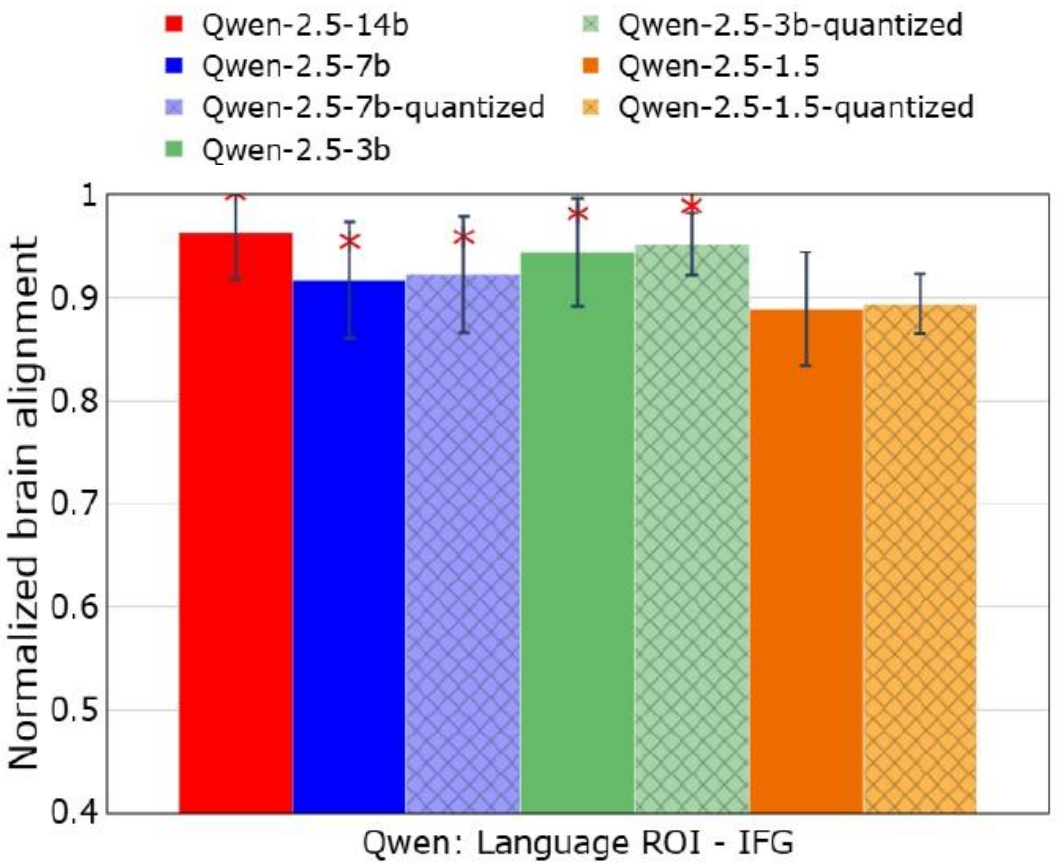


To quantify model predictions, we have an estimate of the explainable variance and use that to measure normalize brain alignment.

What is the minimal model capacity required to achieve brain alignment comparable to larger LLMs?

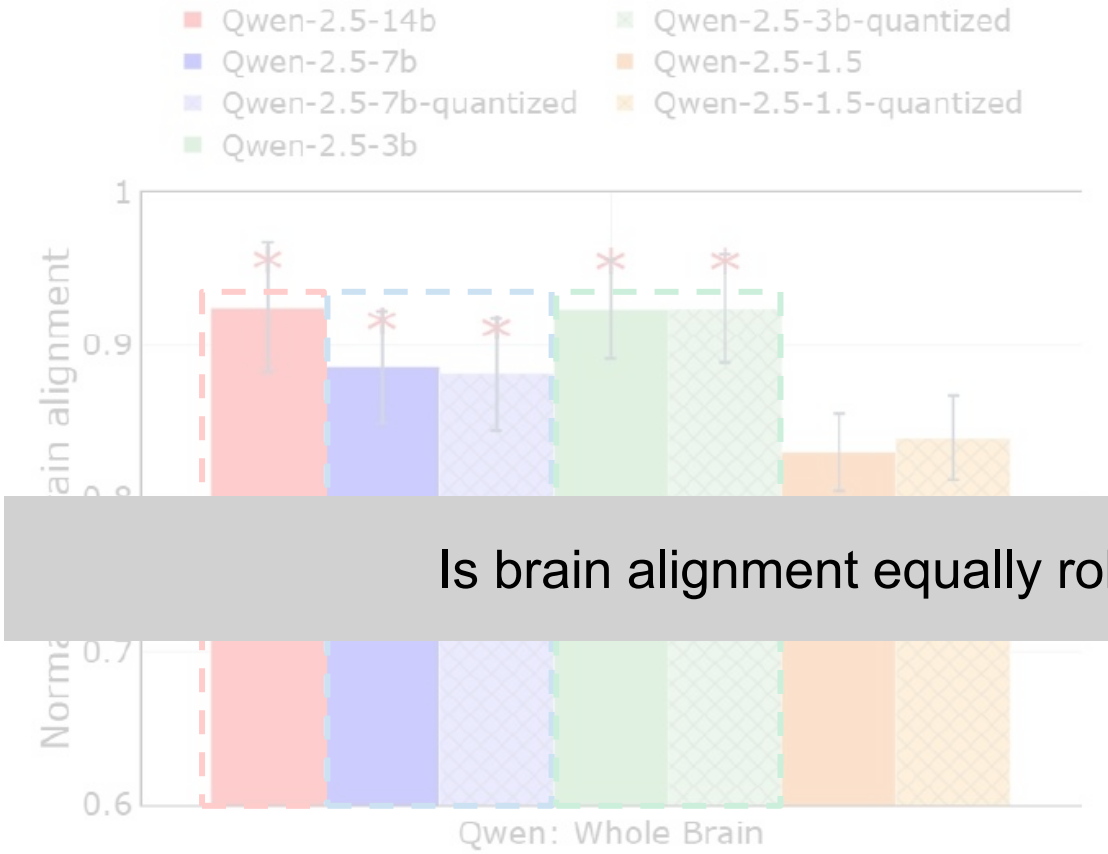
Result-1: SLMs vs. LLMs vs. Quantization variants & brain alignment

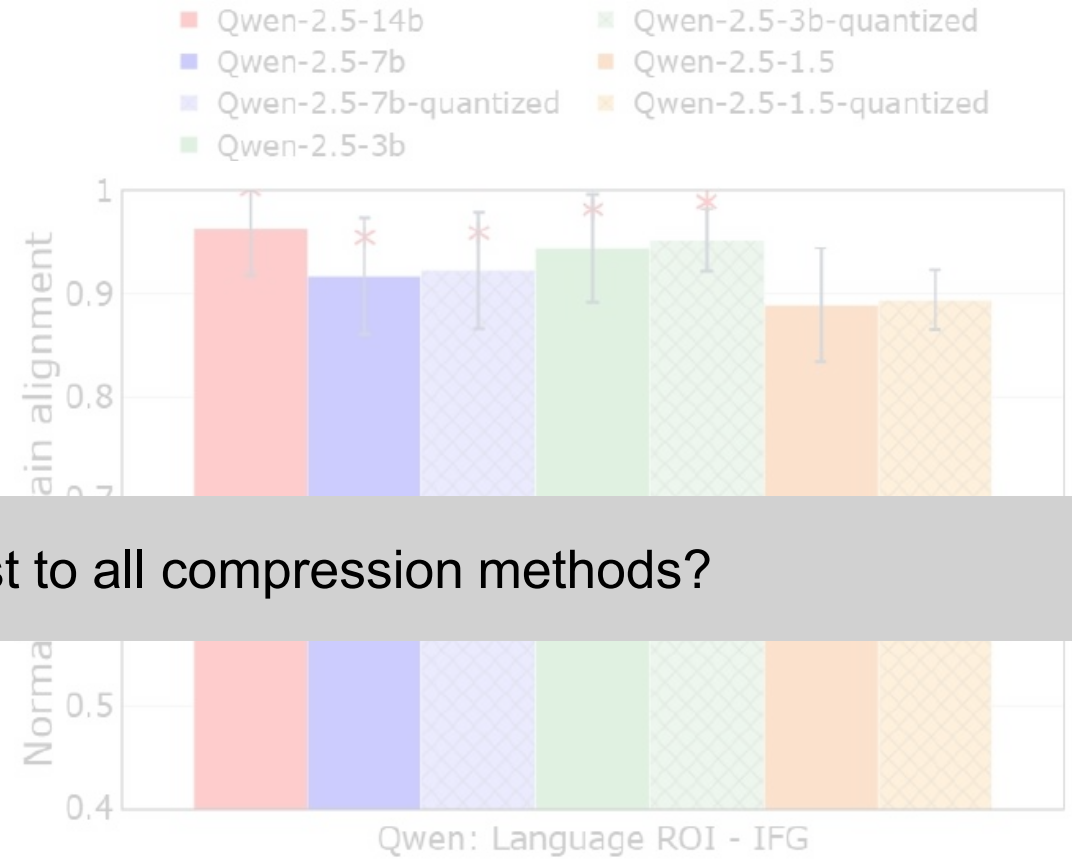




- **3B SLMs** match brain alignment with **7B-14B LLMs,** but **1.5B SLMs** consistently underperform
- Compression analyses show that brain alignment is largely robust to post-training efficiency methods.

Result-1: SLMs vs. LLMs vs. Quantization variants & brain alignment



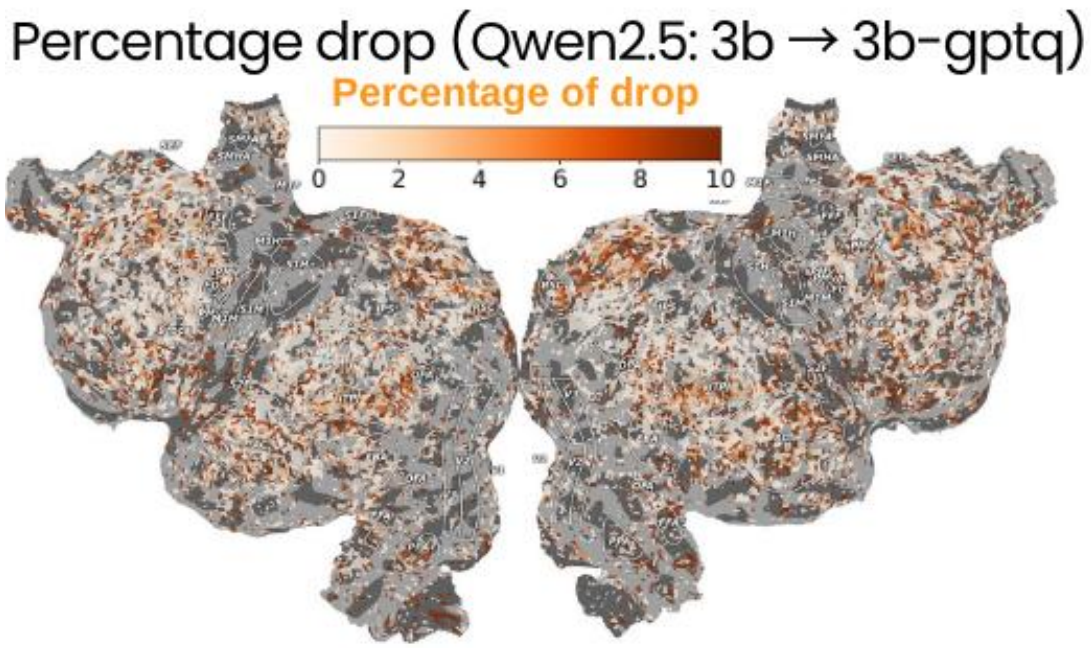


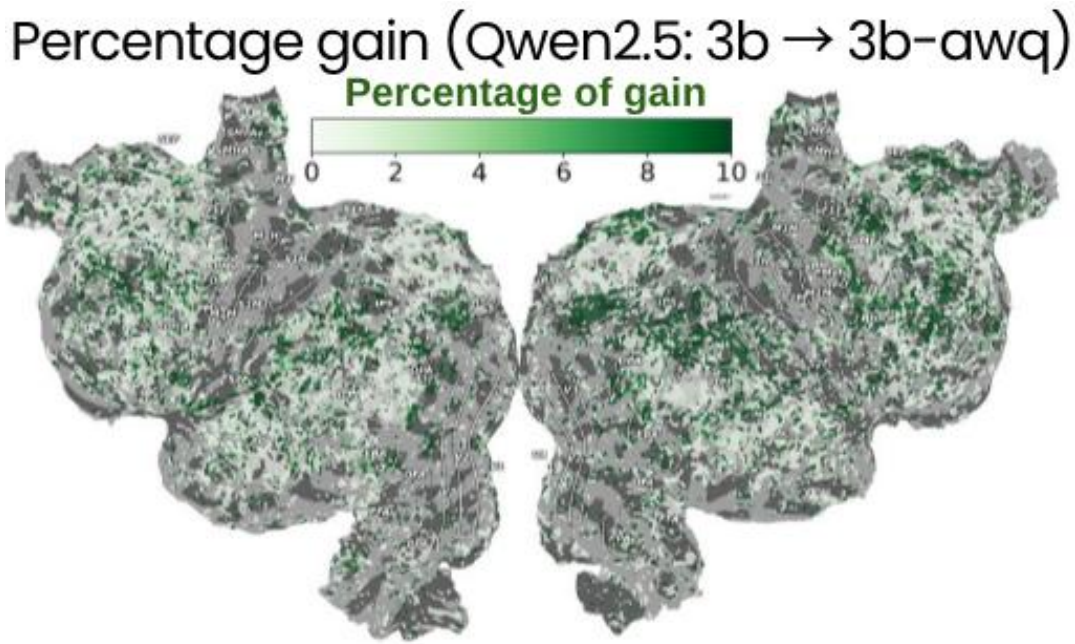
Is brain alignment equally robust to all compression methods?

- **3B SLMs** match brain alignment with **7B-14B LLMs**, but **1.5B SLMs** consistently underperform
- Compression analyses show that brain alignment is largely robust to post-training efficiency methods.

How do compression methods (quantization and pruning) affect brain alignment for both SLMs and LLMs?

Result-2: Brain alignment is robust to compression, except GPTQ





Model variant	Method	Sparsity	Normalized Brain Alignment
Qwen-2.5-3B	FP16 (baseline)	0%	0.924 ±0.033
Qwen-2.5-3B-AWQ	Quantization (AWQ)	0%	0.933 ± 0.035
Qwen-2.5-3B-GPTQ	Quantization (GPTQ)	0%	0.910 ± 0.037
Qwen-2.5-3B-Smooth	Quantization (SmoothQuant)	0%	0.930 ± 0.035
Qwen-2.5-3B-0.1	Pruning	10%	0.910 ± 0.032
Qwen-2.5-3B-0.25	Pruning	25%	0.908 ± 0.033
Qwen-2.5-3B-0.5	Pruning	50%	0.907 ± 0.043

Variant	Mean ± SEM	95% CI	Notes
FP16 (baseline)	0.830 ± 0.025	[0.781, 0.879]	full precision
AWQ	0.854 ± 0.031	[0.793, 0.915]	INT4 quantization
GPTQ	0.828 ± 0.026	[0.777, 0.879]	INT4 quantization
SmoothQuant	0.836 ± 0.025	[0.787, 0.885]	INT4 quantization
Prune 10%	0.847 ± 0.026	[0.796, 0.898]	unstructured pruning
Prune 25%	0.824 ± 0.025	[0.775, 0.873]	unstructured pruning
Prune 50%	0.608 ± 0.053	[0.504, 0.712]	unstructured pruning

Moderate pruning (10–25%) preserves alignment at both scales (1.5-3B), but aggressive 50% pruning collapses the 1.5B model

Result-2: Brain alignment is robust to compression, except GPTQ

Does linguistic competence vary across scale and precision?

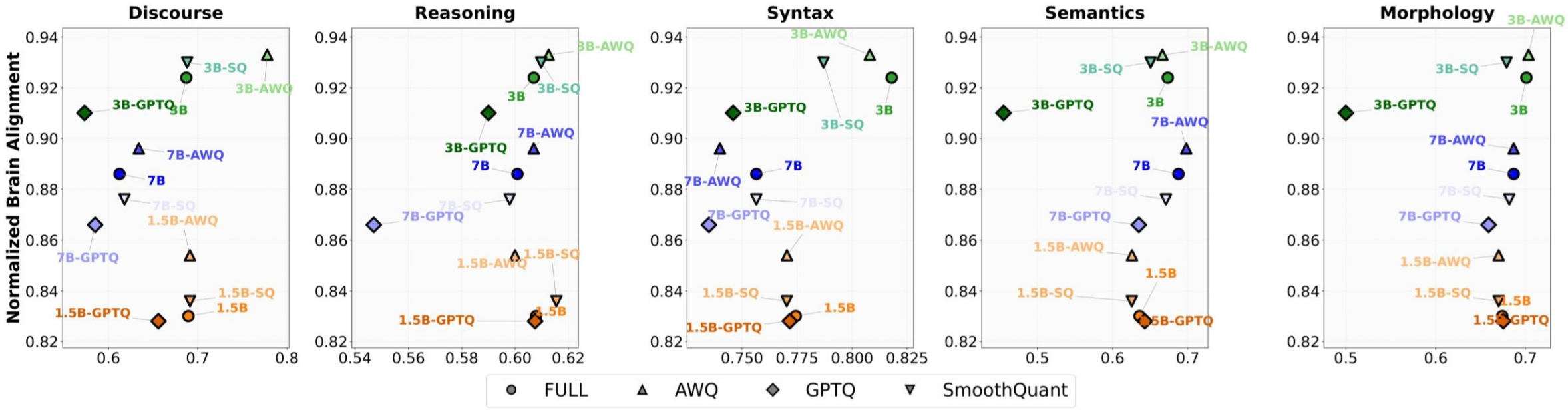
Model variant	Method	Sparsity	Normalized Brain Alignment
Qwen-2.5-3B	FP16 (baseline)	0%	0.924 ±0.033
Qwen-2.5-3B-AWQ	Quantization (AWQ)	0%	0.933 ± 0.035
Qwen-2.5-3B-GPTQ	Quantization (GPTQ)	0%	0.910 ± 0.037
Qwen-2.5-3B-Smooth	Quantization (SmoothQuant)	0%	0.930 ± 0.035
Qwen-2.5-3B-0.1	Pruning	10%	0.910 ± 0.032
Qwen-2.5-3B-0.25	Pruning	25%	0.908 ± 0.033
Qwen-2.5-3B-0.5	Pruning	50%	0.907 ± 0.043

Variant	Mean ± SEM	95% CI	Notes
FP16 (baseline)	0.830 ± 0.025	[0.781, 0.879]	full precision
AWQ	0.854 ± 0.031	[0.793, 0.915]	INT4 quantization
GPTQ	0.828 ± 0.026	[0.777, 0.879]	INT4 quantization
SmoothQuant	0.836 ± 0.025	[0.787, 0.885]	INT4 quantization
Prune 10%	0.847 ± 0.026	[0.796, 0.898]	unstructured pruning
Prune 25%	0.824 ± 0.025	[0.775, 0.873]	unstructured pruning
Prune 50%	0.608 ± 0.053	[0.504, 0.712]	unstructured pruning

Moderate pruning (10–25%) preserves alignment at both scales (1.5-3B), but aggressive 50% pruning collapses the 1.5B model

Which linguistic properties are preserved or degraded across small, large, and compressed models, and do these changes correlate with brain alignment?

Result-3: Linguistic competence and brain alignment dissociate



- **3B** and **7B** models cluster at the top of brain alignment, competence and alignment come apart
- AWQ and SmoothQuant preserve both competence and alignment; GPTQ reduces both
- Tiny SLMs (1.5B) can keep the linguistic properties yet lose brain-relevant representations

Conclusions for neuro-AI research field

1. **Brain alignment saturates early (~3B):** **3B SLMs** match **7B**–14B LLMs, while **1B–1.5B** models reduce alignment, especially in semantic regions.

2. **Alignment is robust to compression—except GPTQ:** AWQ and SmoothQuant preserve alignment; GPTQ consistently degrades it.

3. **Linguistic competence and brain alignment dissociate:** Compression can degrade specific linguistic properties without impacting alignment, while **1B–1.5B** models keep linguistic properties yet fail to capture brain-relevant representations.

Linguistic Properties and Model Scale in Brain Encoding: From Small to Compressed Language Models (ICML-2026)

Subba Reddy Oota

Vijay Rowtula

Satya Sai Srinath

Khushbu Pahwa

Anant Khandelwal

Manish Gupta

Tanmoy Chakraborty

Bapi S. Raju