LMs align with brain regions that represent concepts across modalities

Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, Evelina Fedorenko

McGOVERN INSTITUTE

EvLab

Concepts in LMs

Do language models capture cross-modal conceptual meaning?

No
(Bender & Koller, 2020)

theoretical debate ⟷
(Mitchell & Krakauer, 2023)

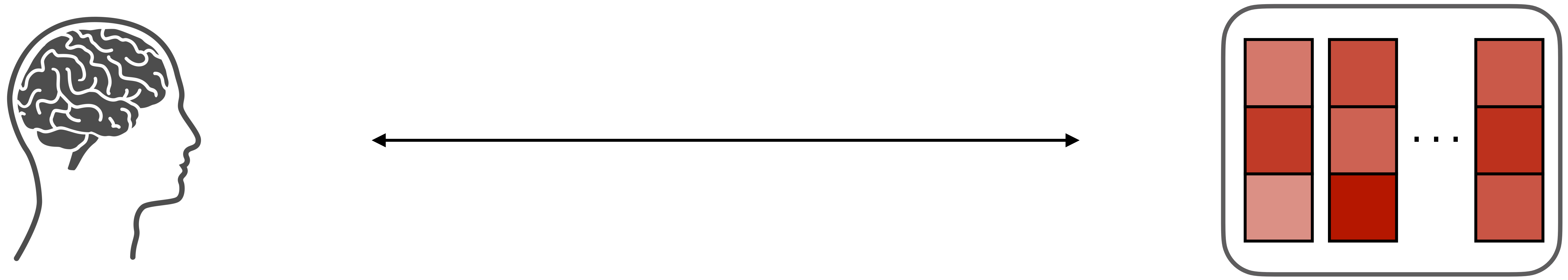
Yes
(Piantadosi & Hill, 2022)

- LMs learn visual concepts from text only (Abdou et al., 2021)
- Models trained on different modalities converge on isomorphic representations (Merullo et al., 2023; Maniparambil & Akshulakov et al., 2024; Li et al., 2024; etc.)
- Platonic representation hypothesis (Huh, Cheung, Wang & Isola, 2024)

LM—brain alignment

- Growing evidence of LM-brain similarity (Oota et al., 2024; Sucholutsky & Muttenthaler et al., 2024; Tuckute et al., 2024)
- Universality of representations (Hosseini et al., 2024; Chen & Bonner, 2024)

LM—brain alignment



Brain representations

LM representations

Representations taken from established networks of brain regions (language network, visual cortex, etc.)

Where in the brain is cross-modal conceptual meaning processed?

Concepts in the brain

$\underbrace{\text{Bird, Broken, Emotion, Science, …}}_{\textit{180 concepts}}$

Pereira et al. "Toward a universal decoder of linguistic meaning from brain activation" (2018)

Concepts in the brain

Bird, Broken, Emotion, Science, …

Sentence paradigm:

The **bird** flew around the cage.

Picture paradigm:

bird

Word cloud paradigm:

winged

nest

bird

beak

mating

flock

Pereira et al. "Toward a universal decoder of linguistic meaning from brain activation" (2018)

Concepts in the brain

Bird, Broken, Emotion, Science, …

Sentence paradigm:

The **broken** mirror went missing.

Picture paradigm:

broken

Word cloud paradigm:

mess bad failing **broken** crooked ruined

Pereira et al. "Toward a universal decoder of linguistic meaning from brain activation" (2018)

Concepts in the brain

17 participants

+

fMRI

=

Brain responses to 4-6 sentences, pictures, and word clouds per concept

Sentence paradigm:

The **broken** mirror went missing.

Picture paradigm:

broken

Word cloud paradigm:

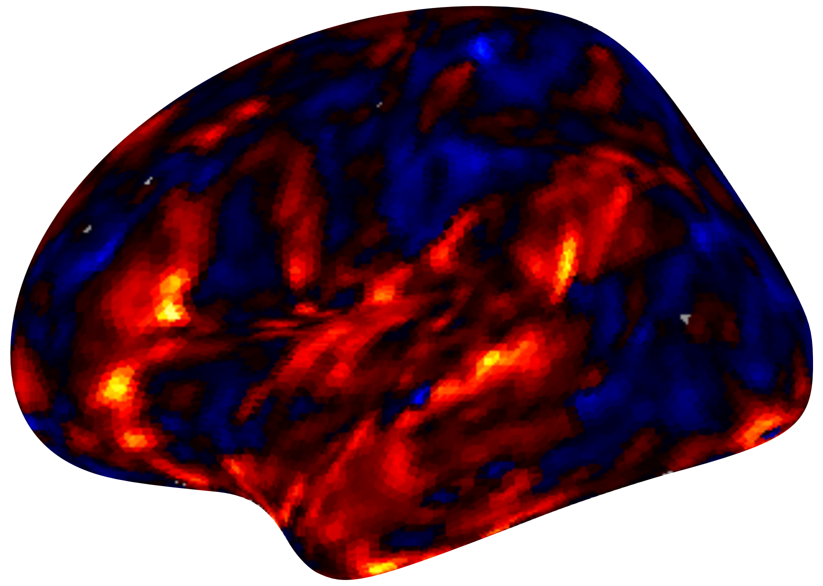
mess bad failing **broken** crooked ruined

Pereira et al. "Toward a universal decoder of linguistic meaning from brain activation" (2018)

fMRI data

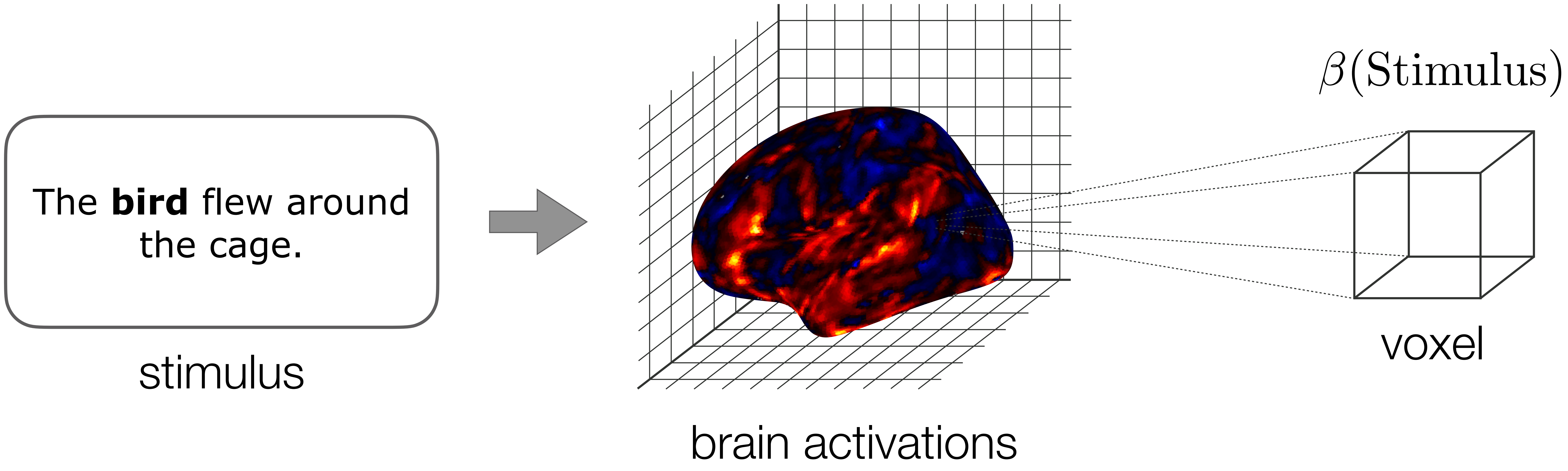
The **bird** flew around the cage.

stimulus



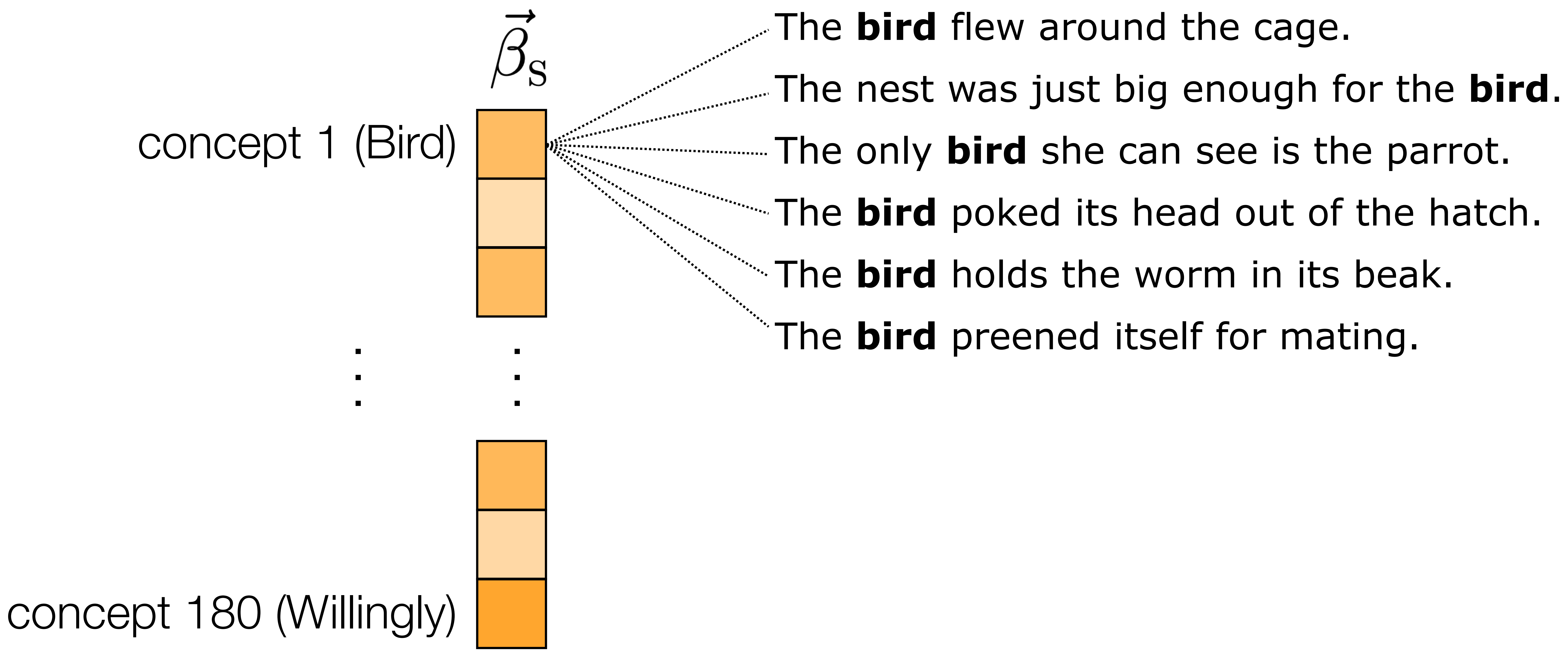
brain activations

fMRI data


The **bird** flew around the cage.
stimulus
brain activations
$\beta(\mathrm{Stimulus})$
voxel

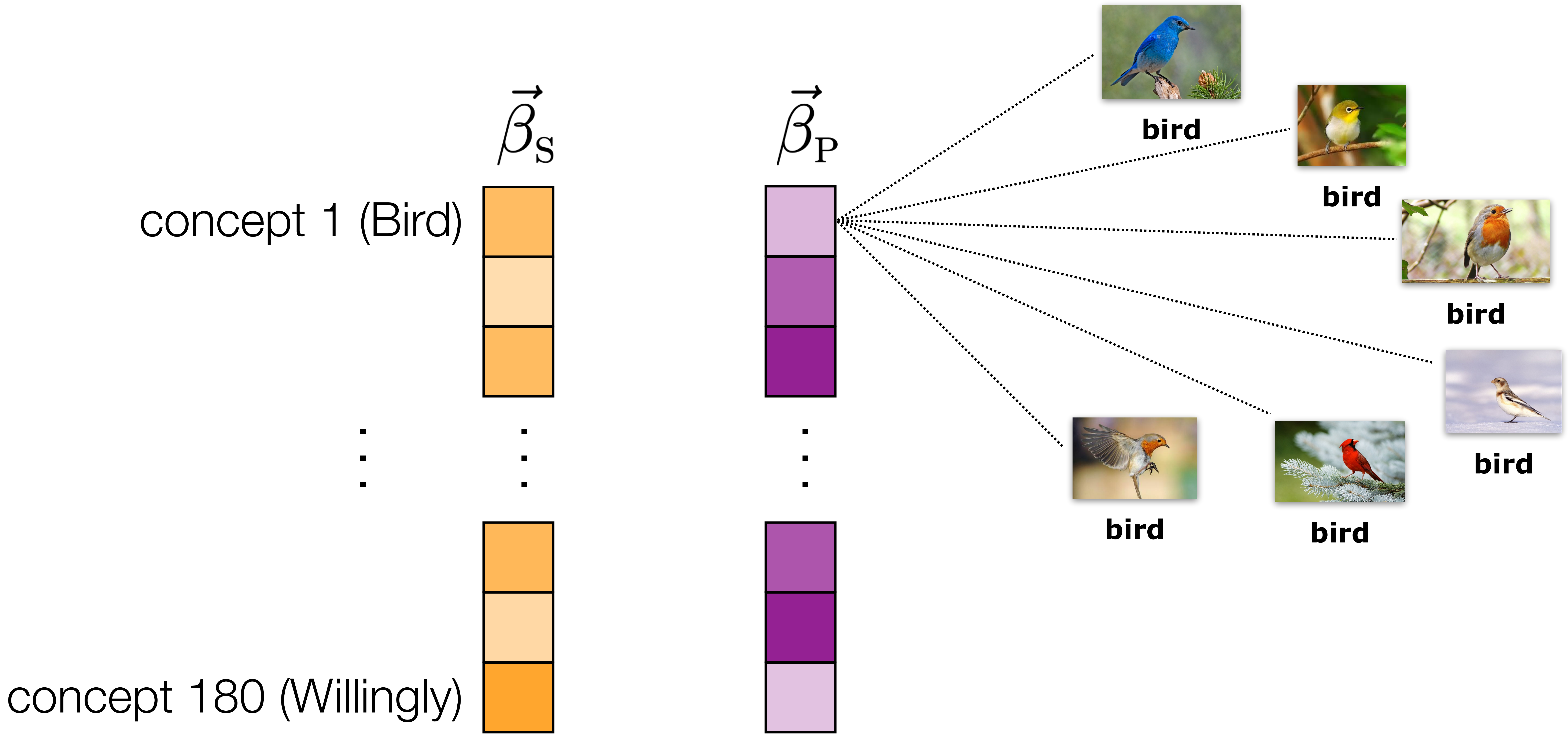
Semantic consistency

- For every voxel:



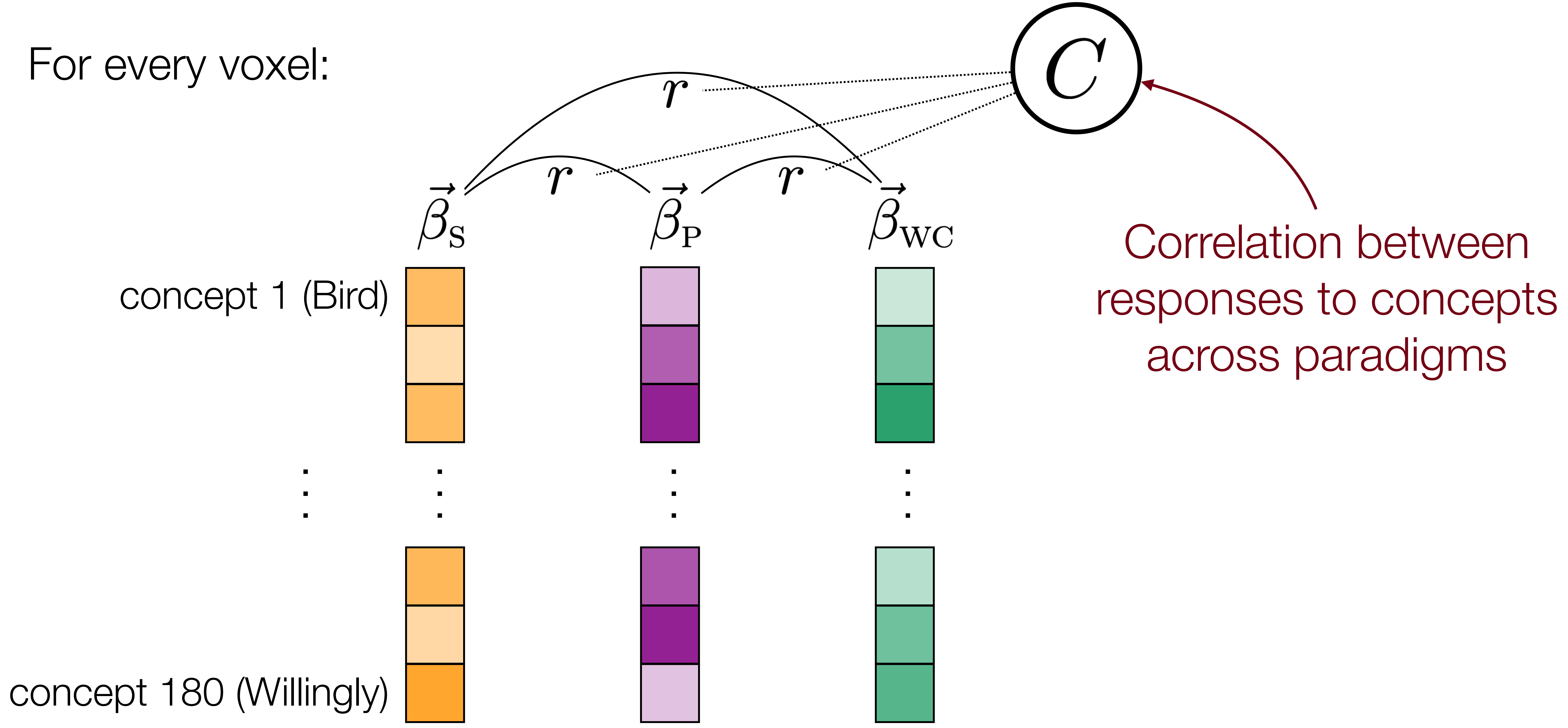
Semantic consistency

- For every voxel:



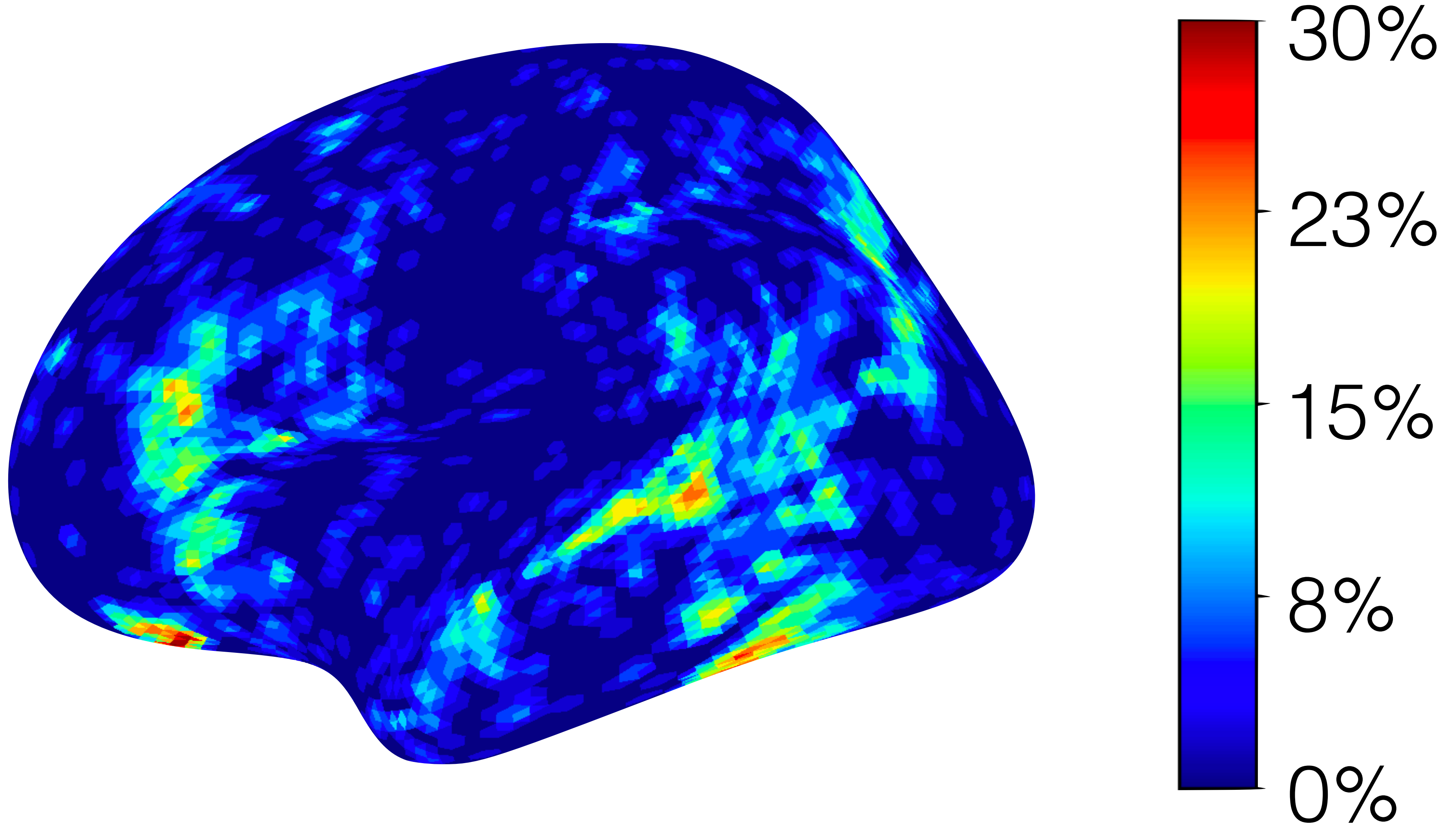
Semantic consistency

- For every voxel:



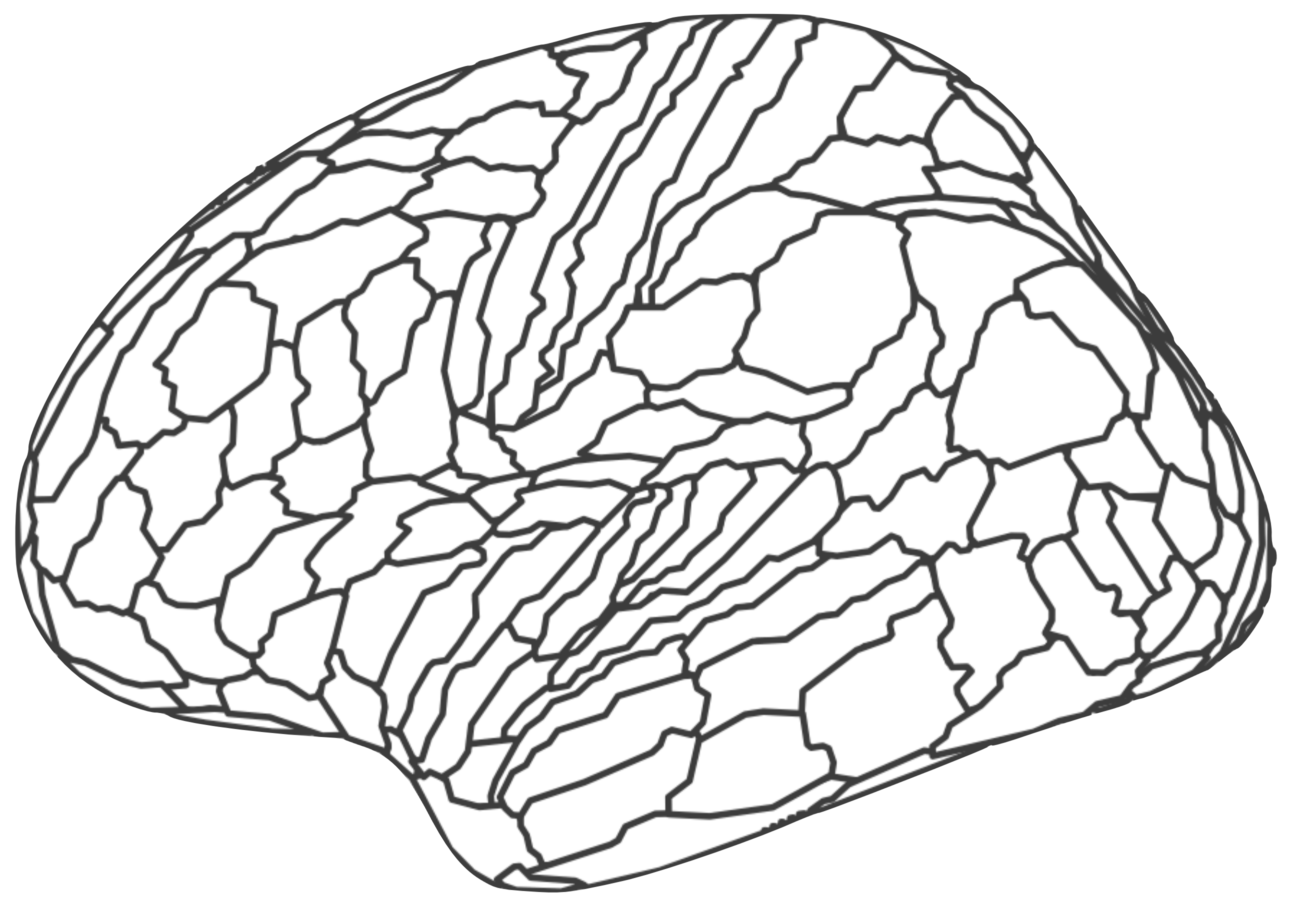
Consistent brain regions

Distribution of voxels with statistically significant semantic consistency



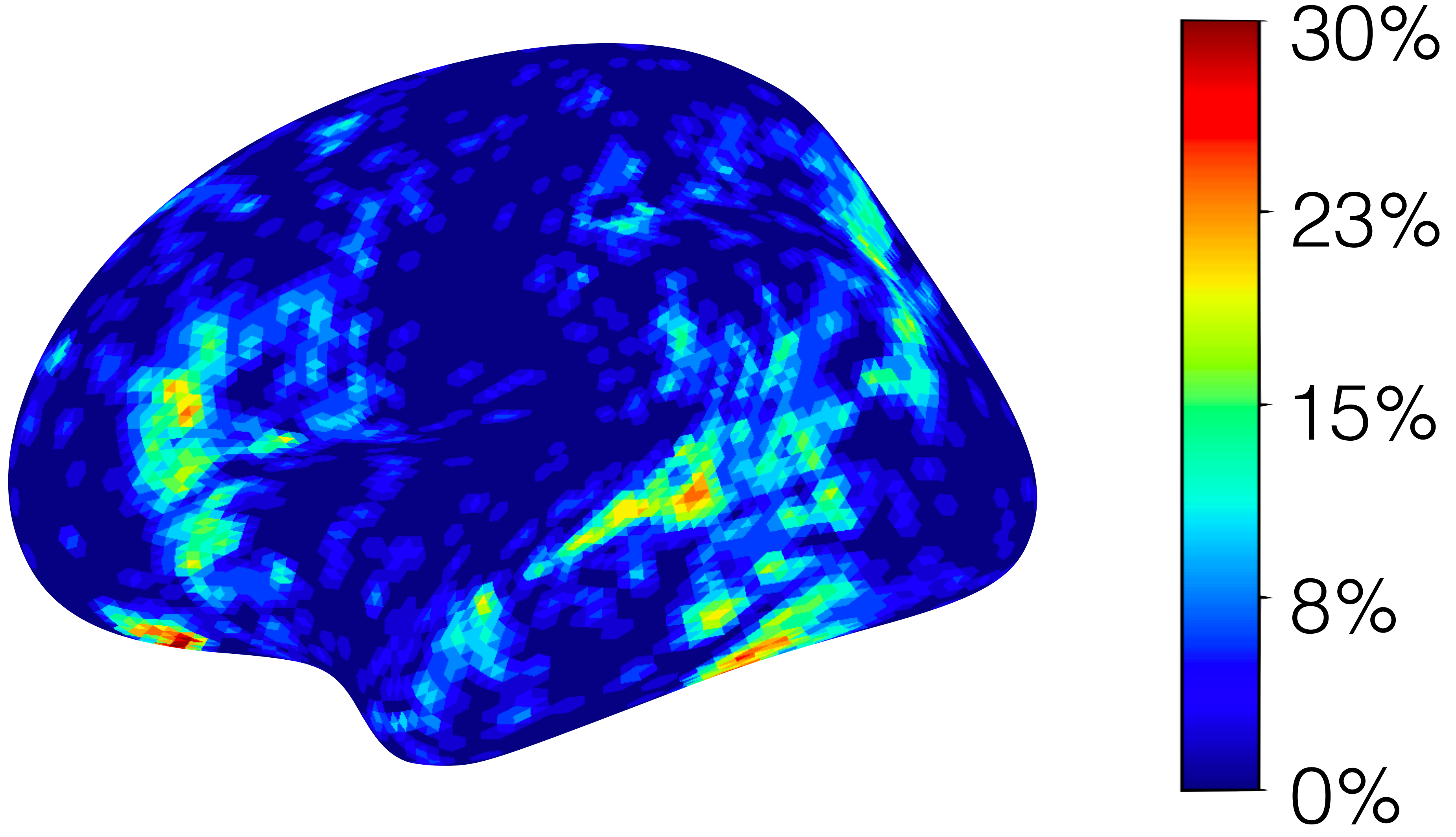
+

Brain cortex segmentation
(Glasser et al., 2016)

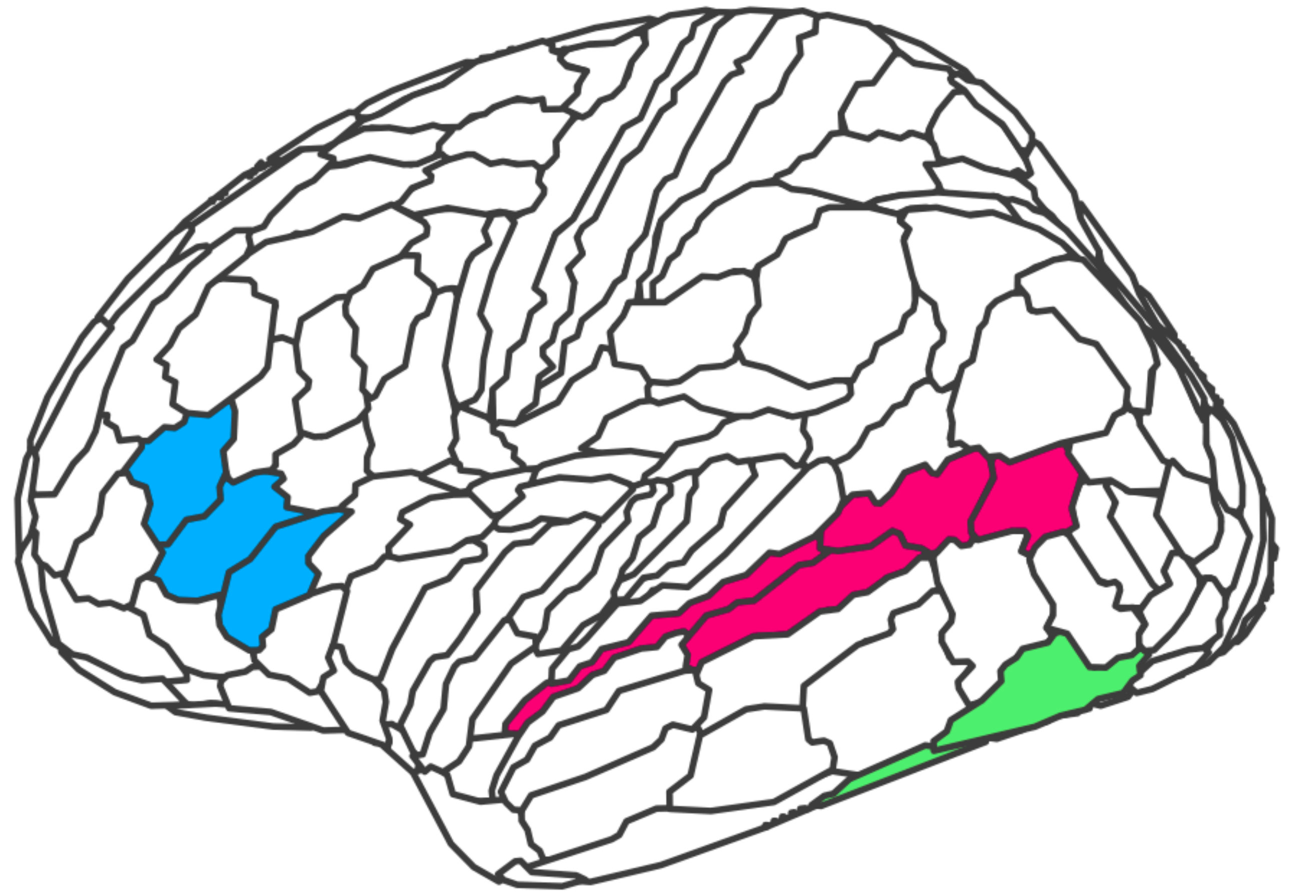


Consistent brain regions

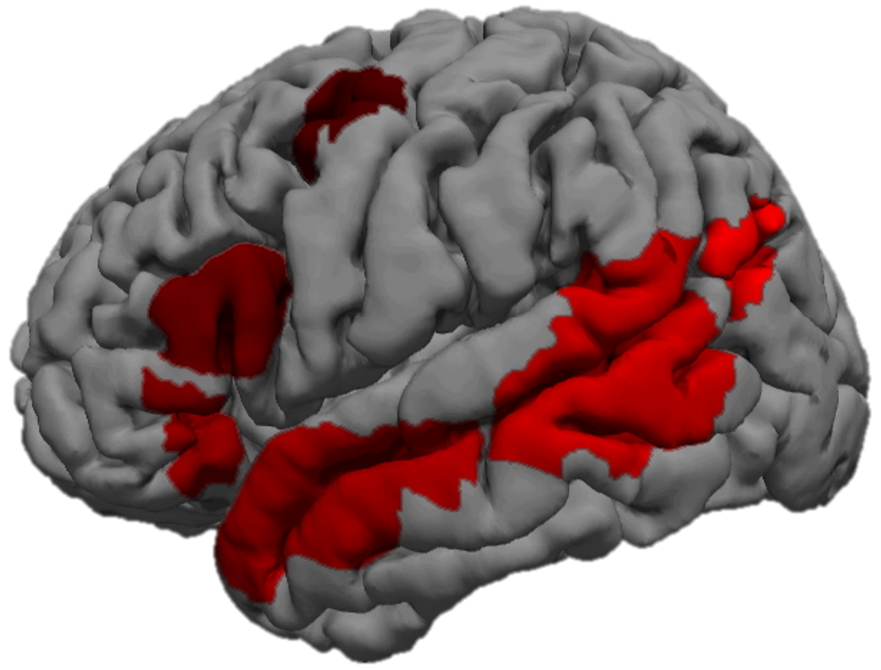
Distribution of voxels with statistically significant semantic consistency



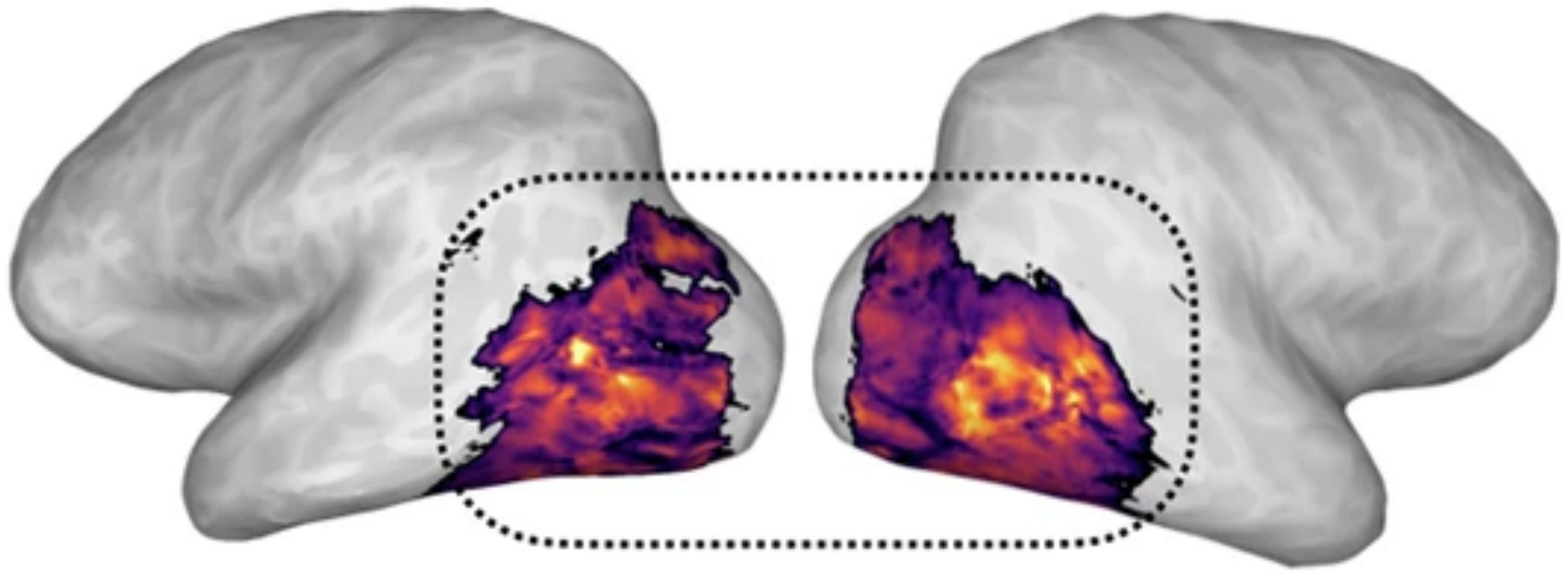
Regions of interest (ROI)



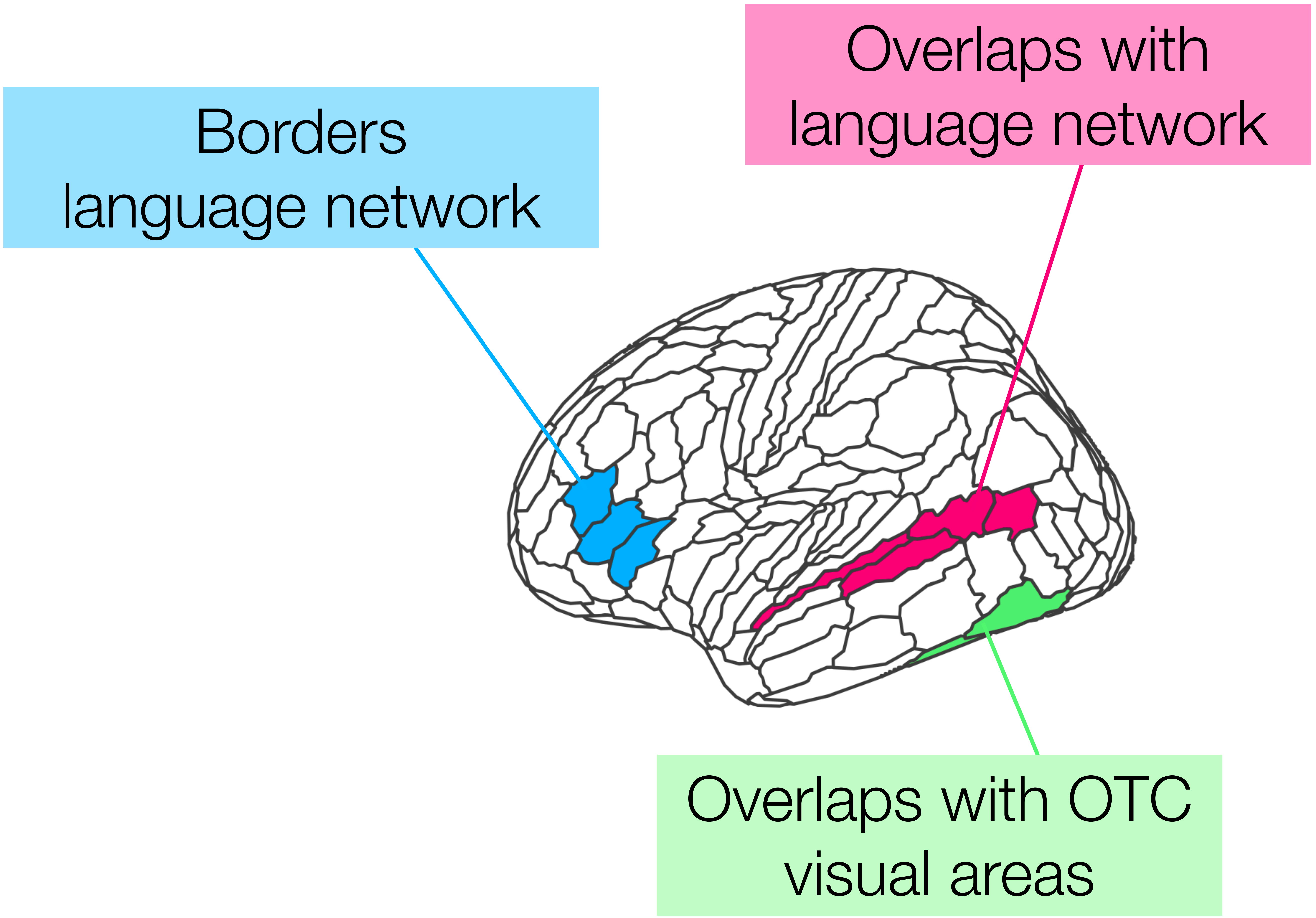
Consistent brain regions



Language network
(Lipkin et al., 2022)

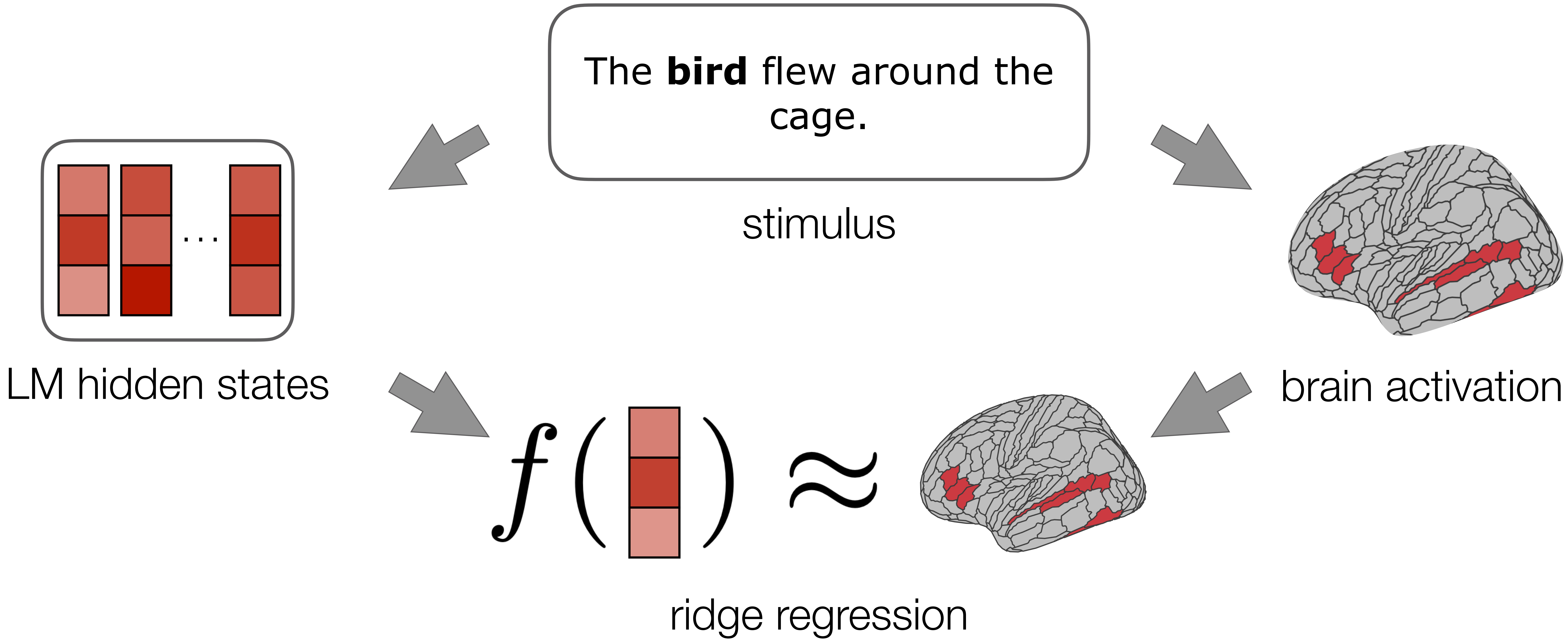


Occipitotemporal cortex
(Conwell, 2024)

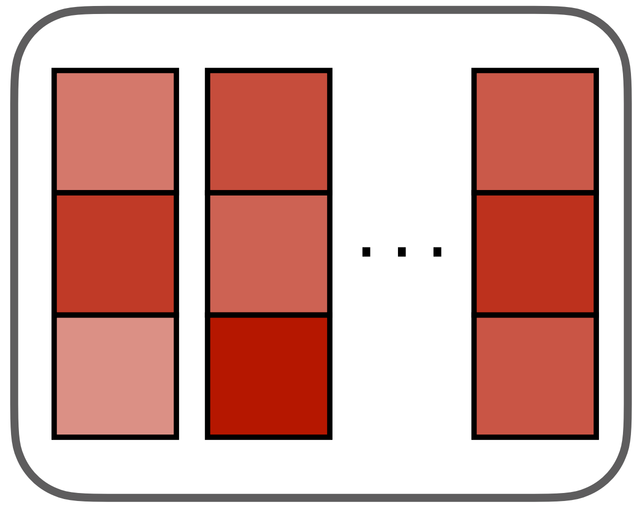


Brain encoding

Using LM representations to predict brain activations



Brain encoding



LM hidden states

- 15 transformer LMs:
  - Model size: GPT-2 (S vs. M vs. L vs. XL); Qwen2.5 (1.5B vs. 3B vs. 7B)
  - Instruction tuning: Qwen2.5 vs. Qwen2.5-Instruct
  - Multimodality: Qwen2.5-Instruct vs. Qwen2.5-VL-Instruct; Vicuna vs. LLaVA
- Cross-validation to choose best layer and token pooling method for each model

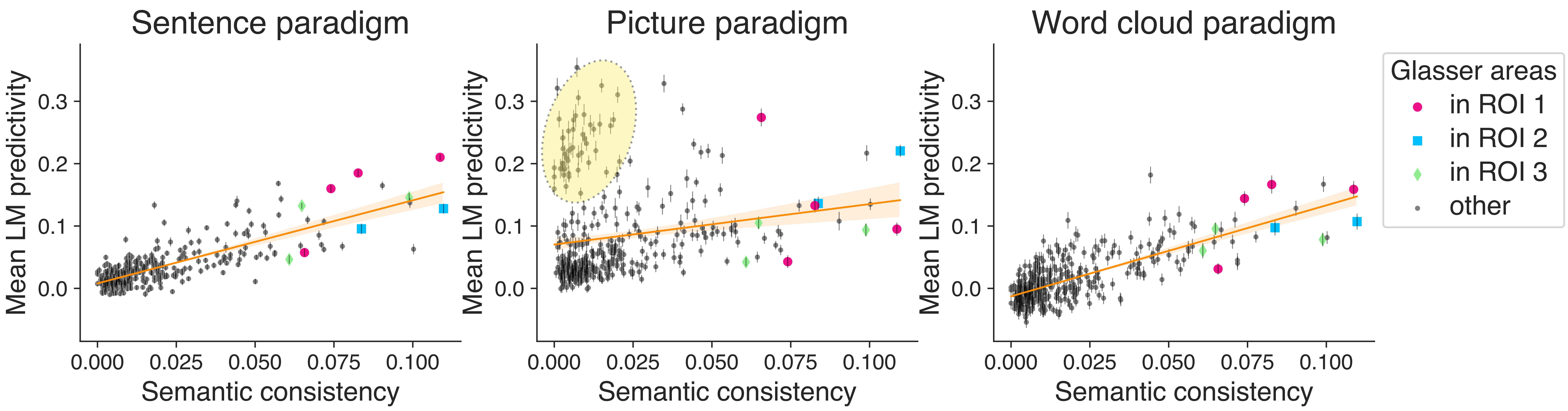
Brain encoding: whole brain

Across the brain, areas with higher semantic consistency are better predicted by LMs

$r(\vec{\beta}_{pred}, \vec{\beta}^{*})$

Brain encoding: whole brain

Across the brain, areas with higher semantic consistency are better predicted by LMs

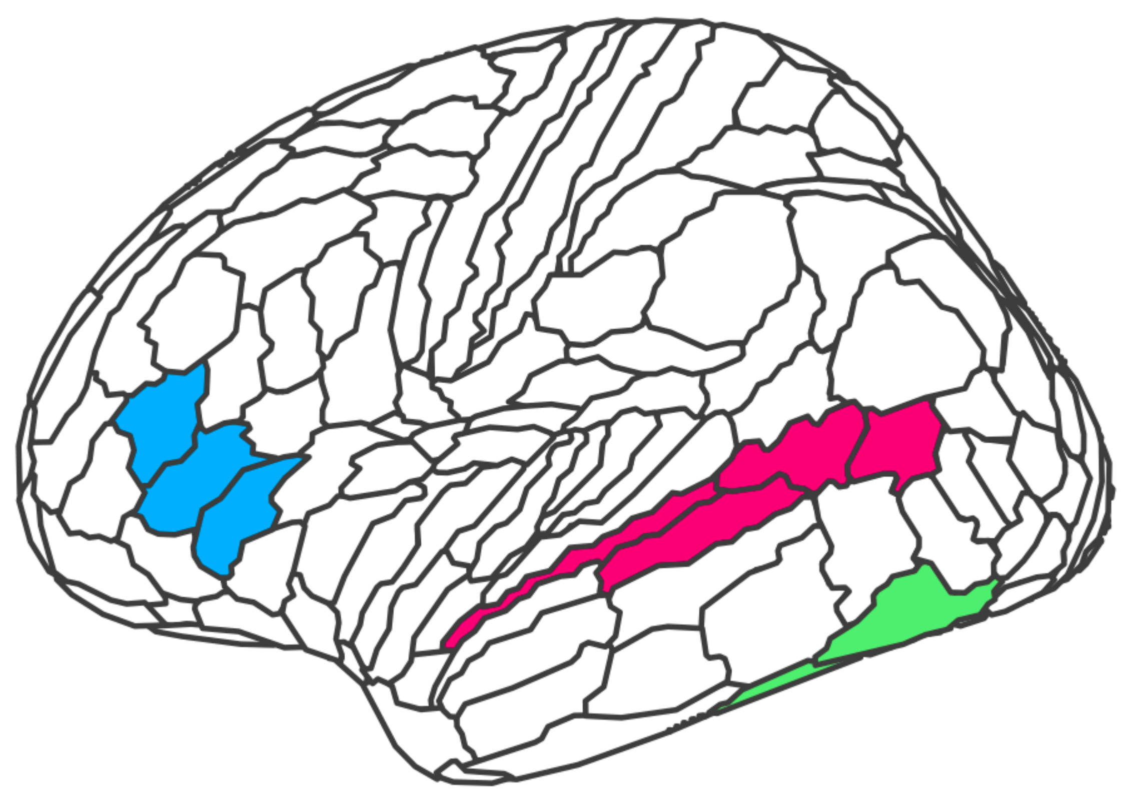


Brain encoding: our ROIs

What drives LM encoding performance —
semantic consistency or response to language?

LAST MONTH EVERYONE … > REDENTION ZOOD CRE …

Selectivity for sentences > non-words (Fedorenko et al., 2010)



Brain encoding: our ROIs

What drives LM encoding performance —
semantic consistency or response to language?

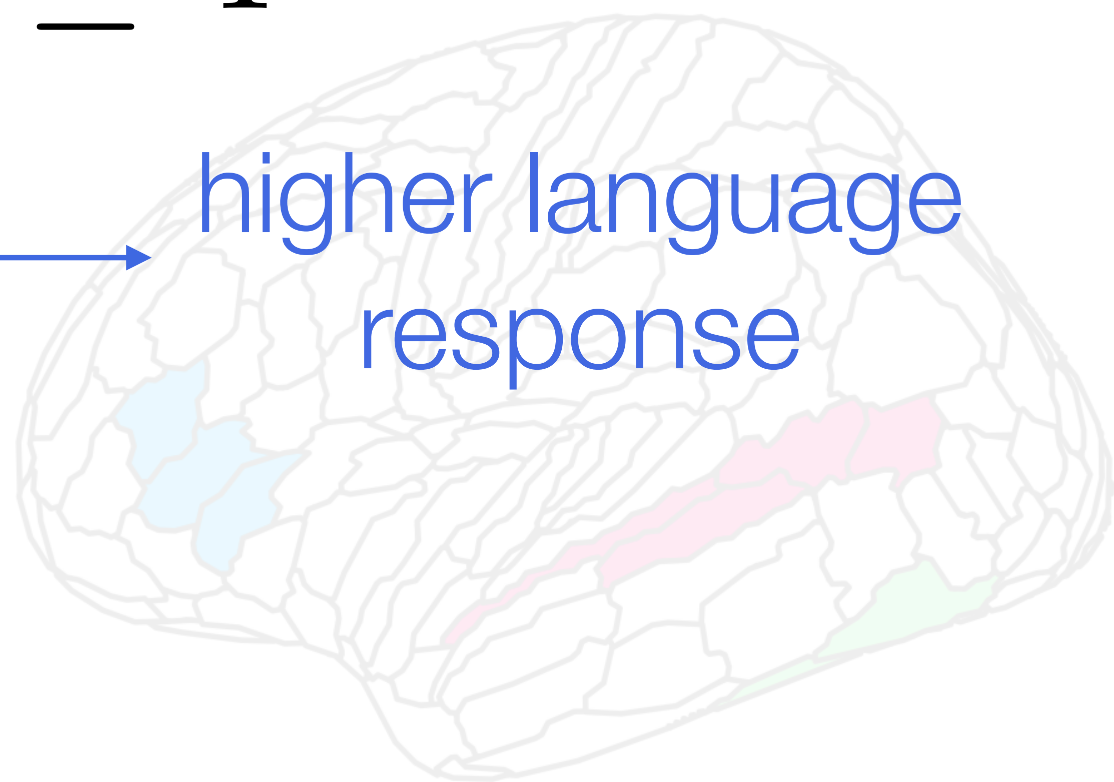
1. Split voxels in each ROI into quartiles by either metric: $v \in (b_C, b_L)$

$$1 \leq b_C \leq 4$$

lower semantic consistency ⟷ higher semantic consistency

$$1 \leq b_L \leq 4$$

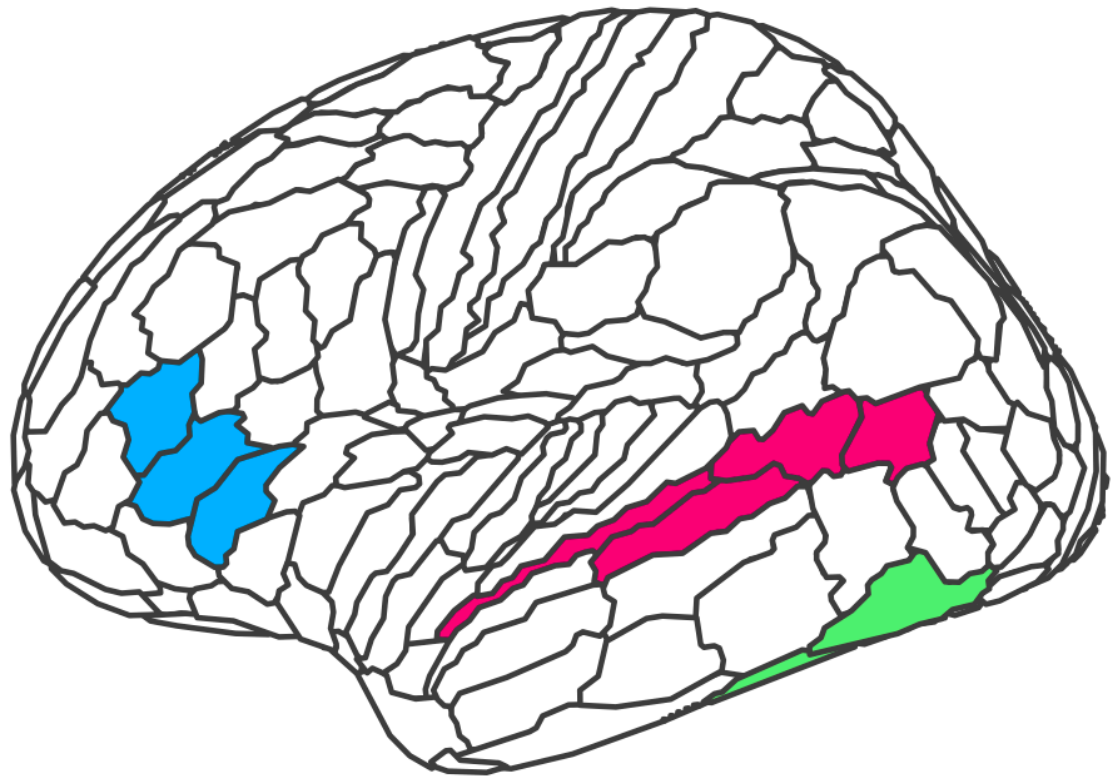
lower language response ⟷ higher language response



Brain encoding: our ROIs

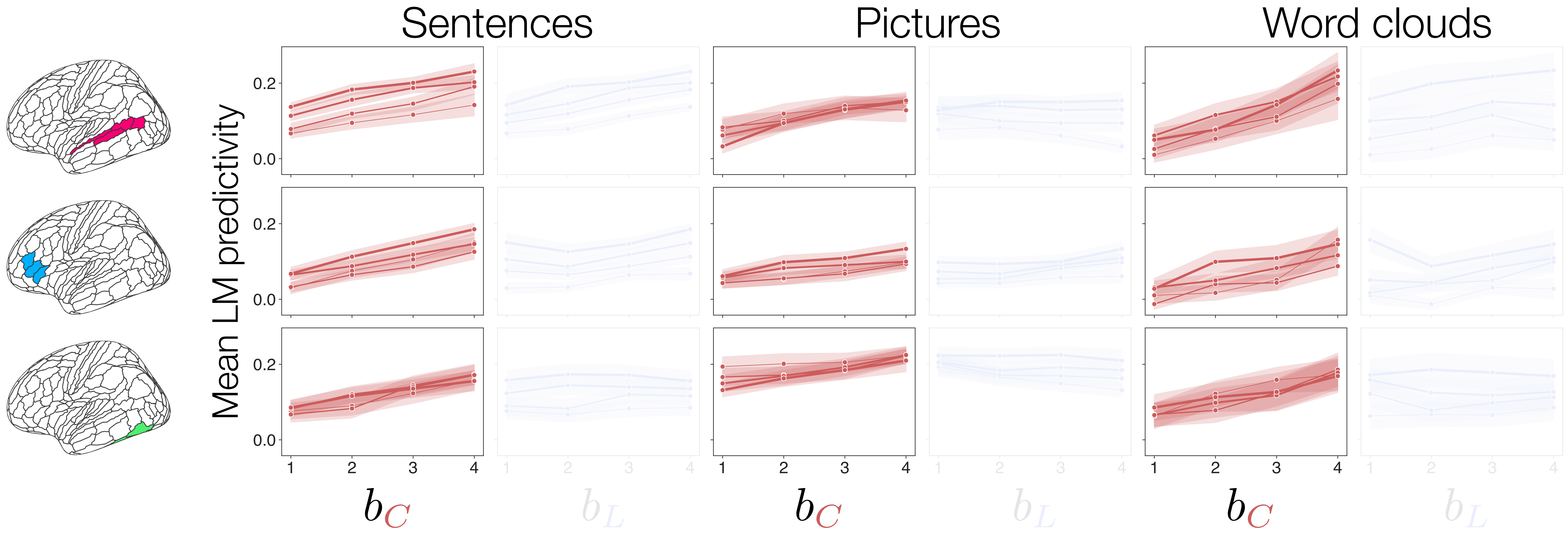
What drives LM encoding performance —
semantic consistency or response to language?

1. Split voxels in each ROI into quartiles by either metric: $v \in (b_C, b_L)$
2. Hold one fixed and vary the other: $(b_C, \cdot), (\cdot, b_L)$



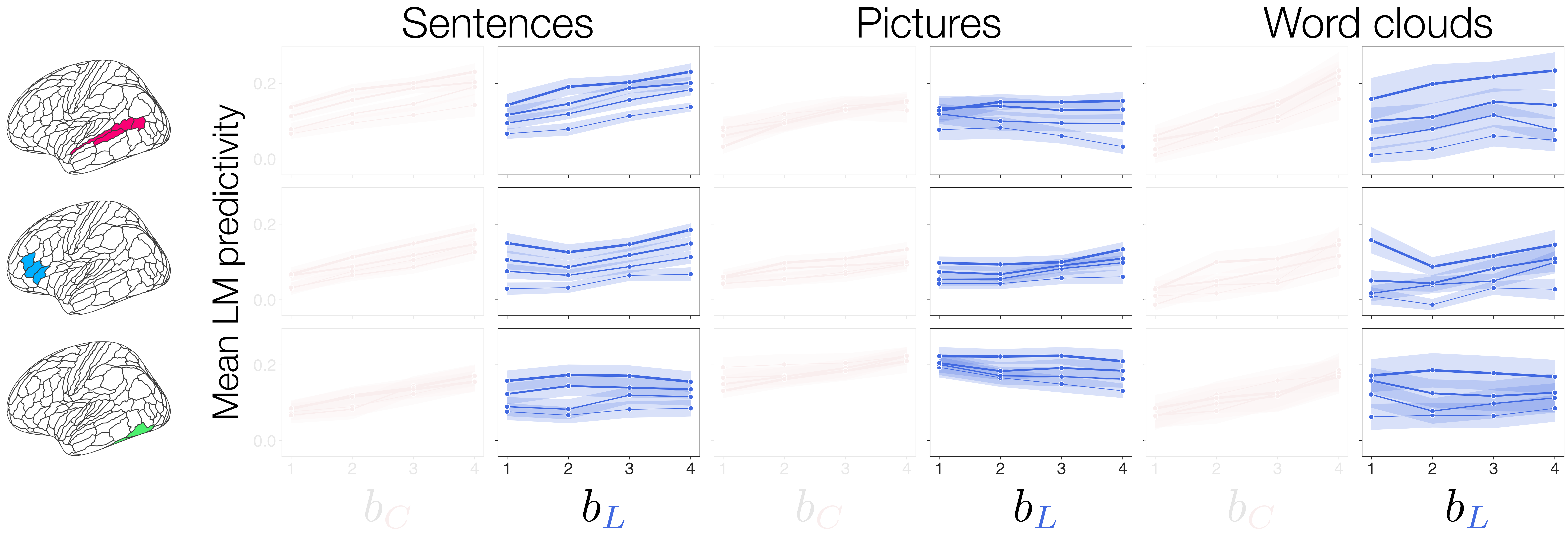
Brain encoding: our ROIs

Strong correlation with semantic consistency



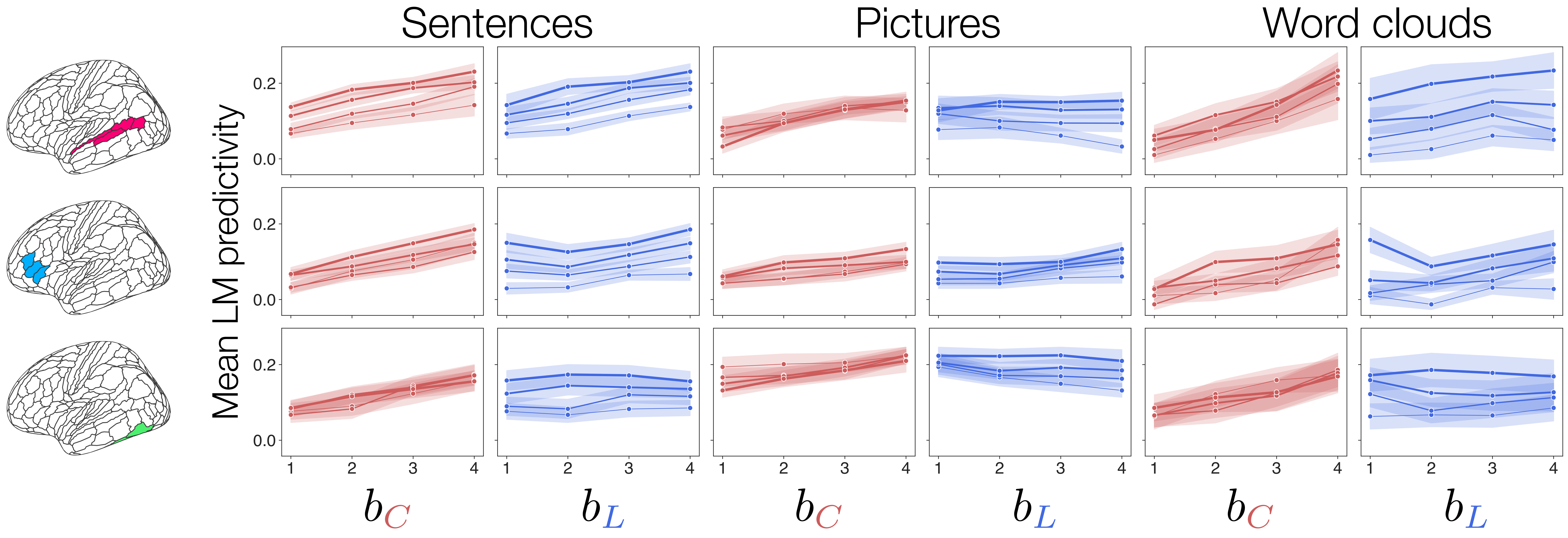
Brain encoding: our ROIs

Strong correlation with semantic consistency



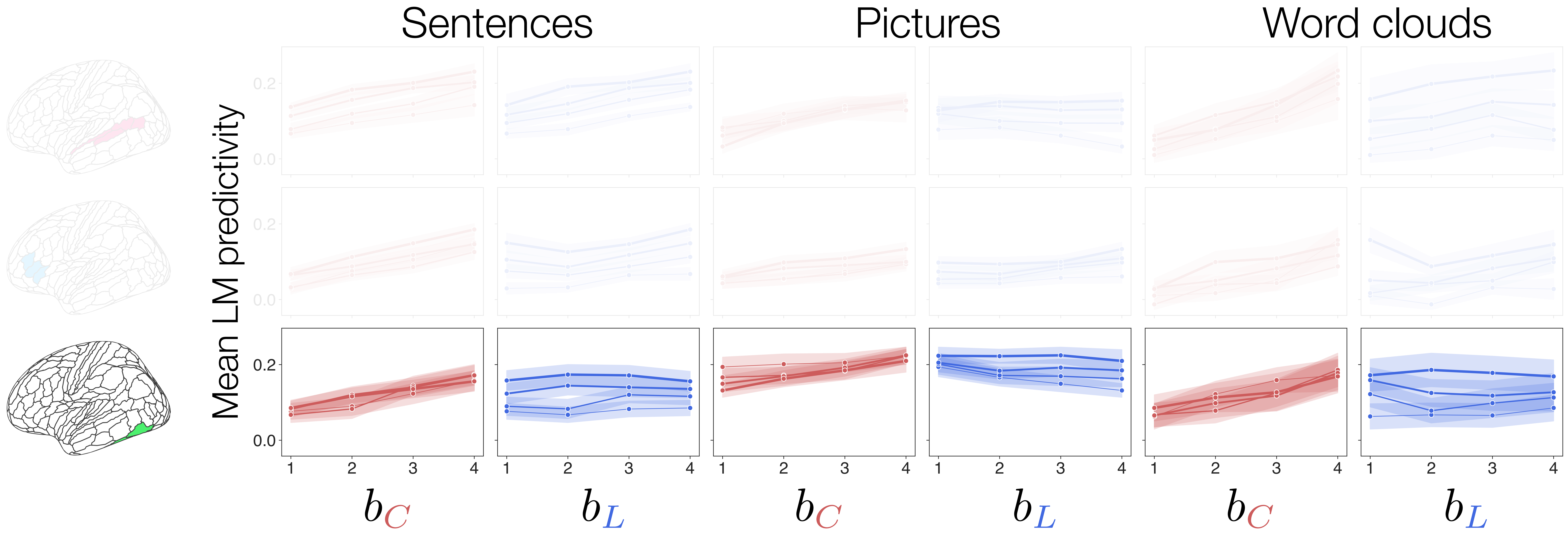
Brain encoding: our ROIs

Strong correlation with semantic consistency

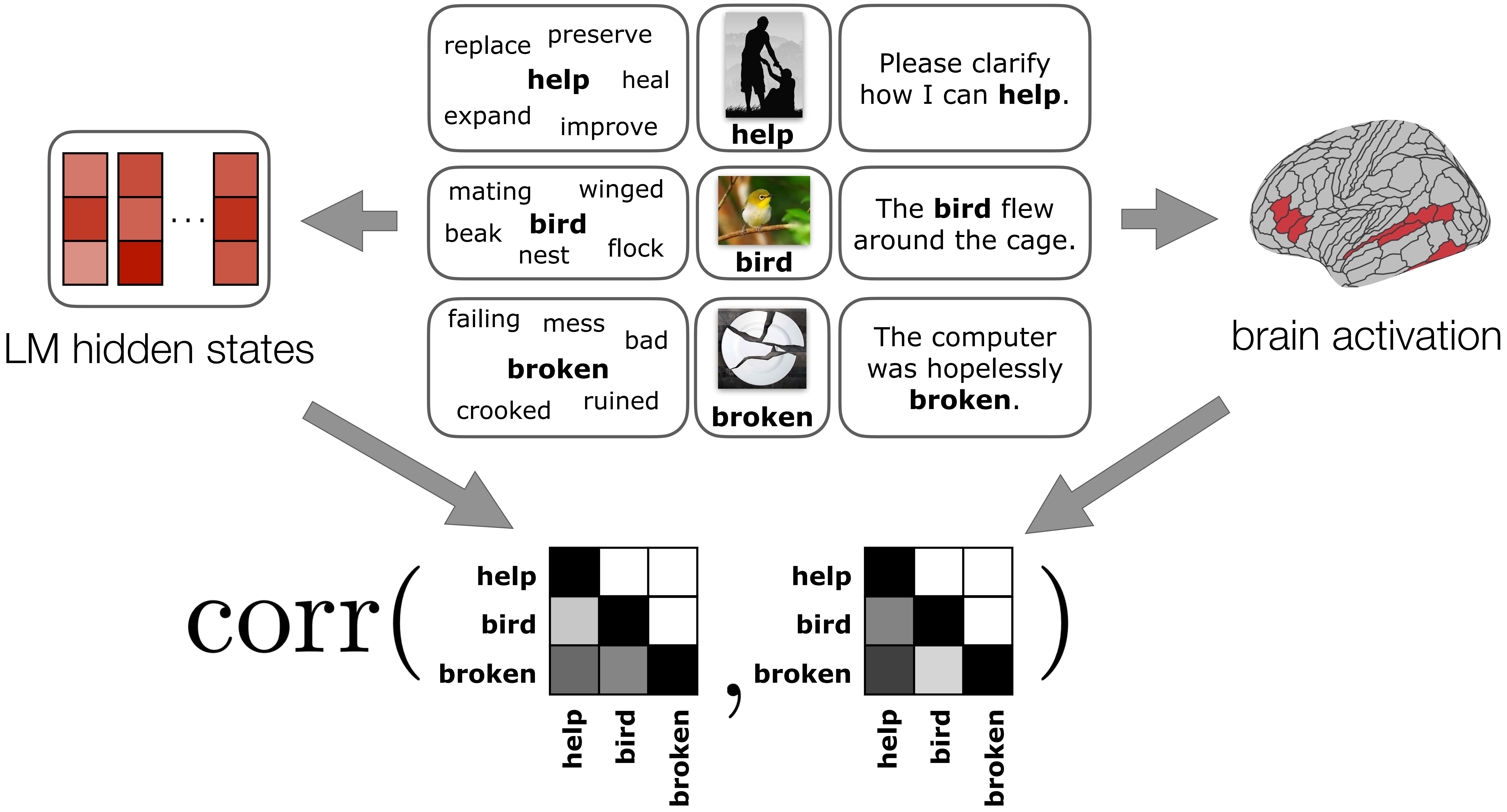


Brain encoding: our ROIs

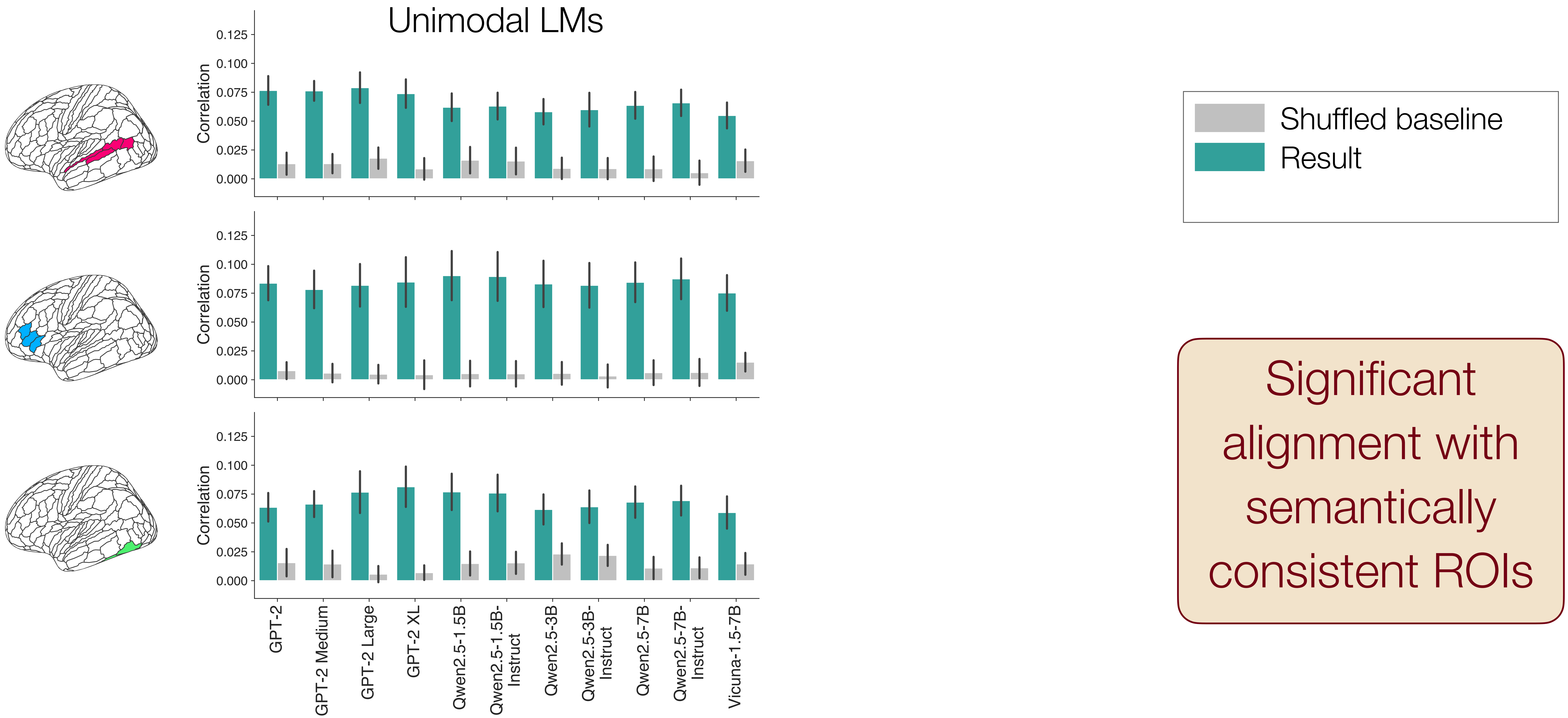
Strong correlation with semantic consistency, even when decoupled from language



Representational similarity

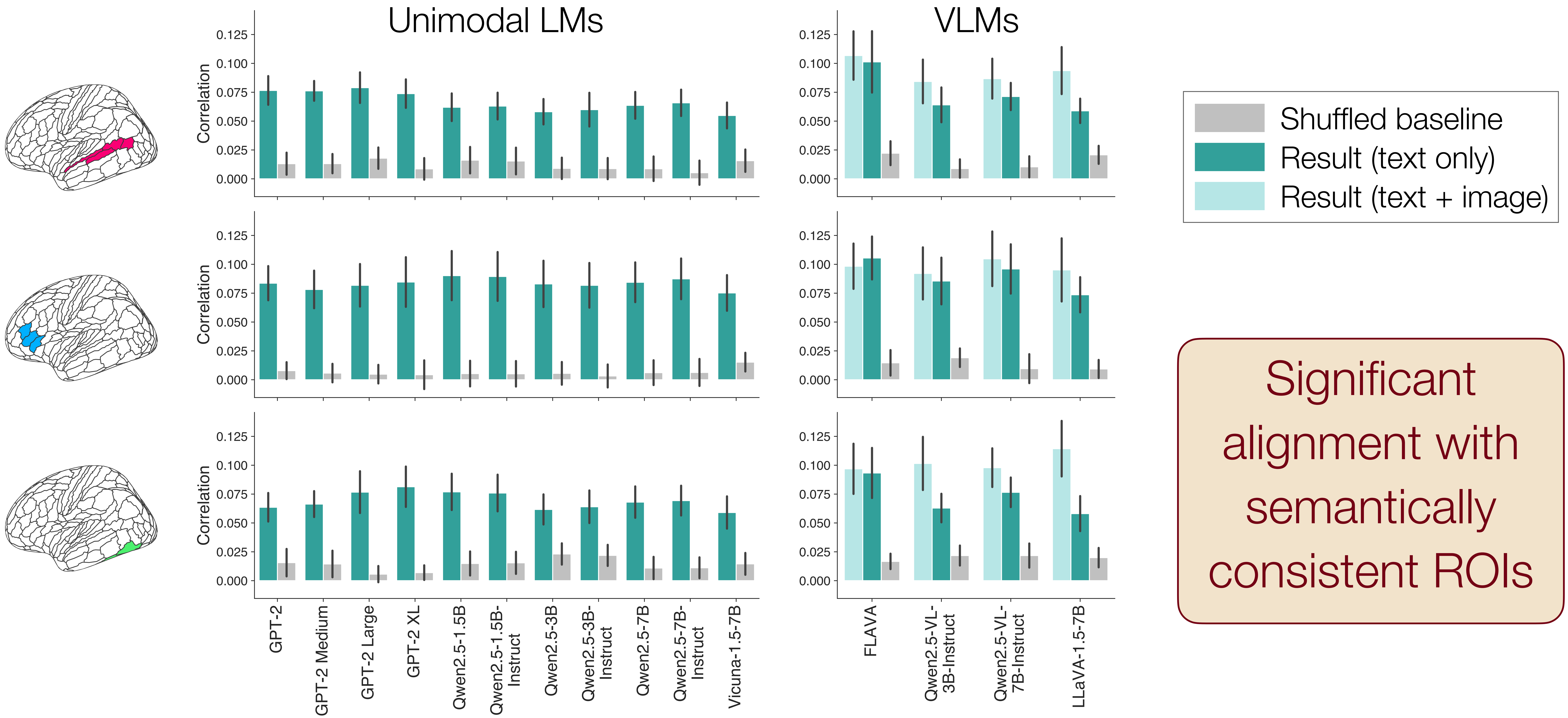


Representational similarity



Significant alignment with semantically consistent ROIs

Representational similarity



Significant alignment with semantically consistent ROIs

Takeaways

- New fMRI-based measure of semantic consistency in the brain, used to find brain regions that represent concepts consistently across modalities
- Two LM-brain alignment analyses: brain encoding (predicting brain activations from LM representations) and RSA (comparing representational geometries)
- LM encoding performance is correlated with semantic consistency, even in regions with low response to language
- Significant representational similarity between LMs and semantically consistent brain regions, further increasing when both images and text are used

Evidence for LMs' ability to capture cross-modal conceptual meaning

Thank you!

Chat with me: Poster Session 2
maria.ryskina@vectorinstitute.ai

Paper:

Code + data:




github.com/ryskina/concepts-brain-llms

McGOVERN INSTITUTE

EvLab