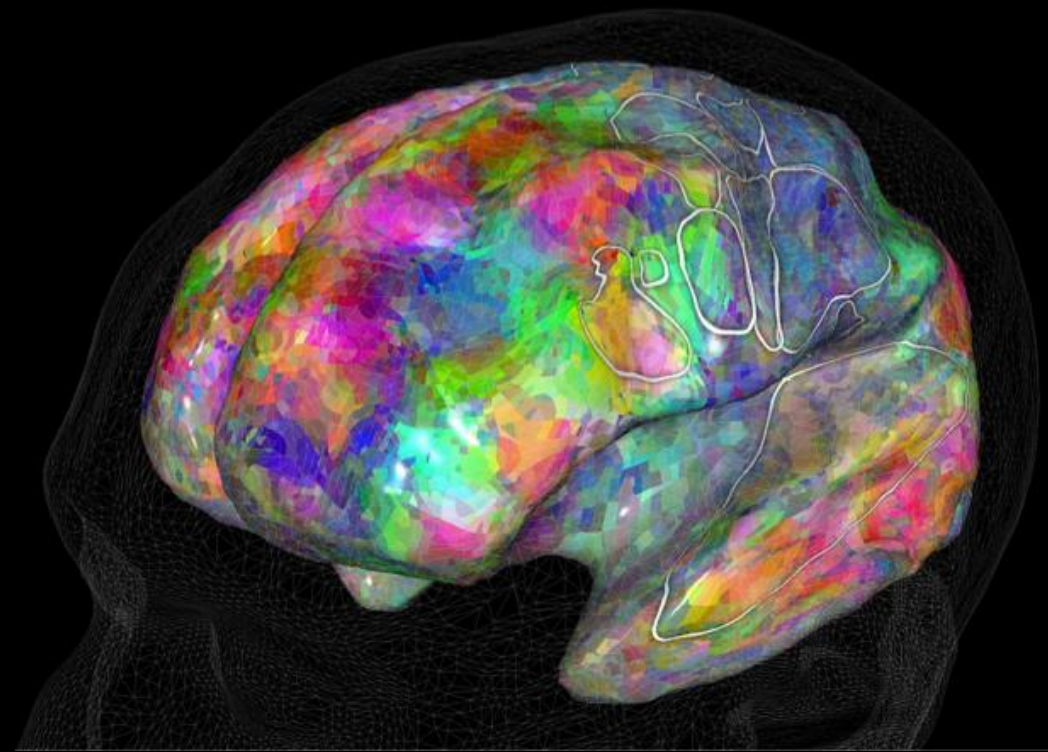
# **Crafting Interpretable Embeddings** by Asking LLMs Questions

Vinamra Benara*, Chandan Singh*,
Jack Morris, Richard Antonello,
Ion Stoica, Alexander G. Huth, Jianfeng Gao

Microsoft Research
UC Berkeley
UT Austin
Cornell University



# LLMs let us build explanation-based embeddings

Input text

*I started to walk in this beautiful neighborhood. It was 10:30 at night*

+

Answer yes/no to the following question:

Does the sentence include numerical information?
Is the sentence a command?
Is the sentence grammatically complex?
Does the sentence mention a specific location or place?
.....
Does the sentence include dialogue?
Does the input describe a sensory experience?
Does the sentence describe a relationship between people?
Does the input mention time?

Map yes/no → 1/0

LLM

Embedding

1
0
0
1
...
0
0
0
1

Predict fMRI response

Actual response

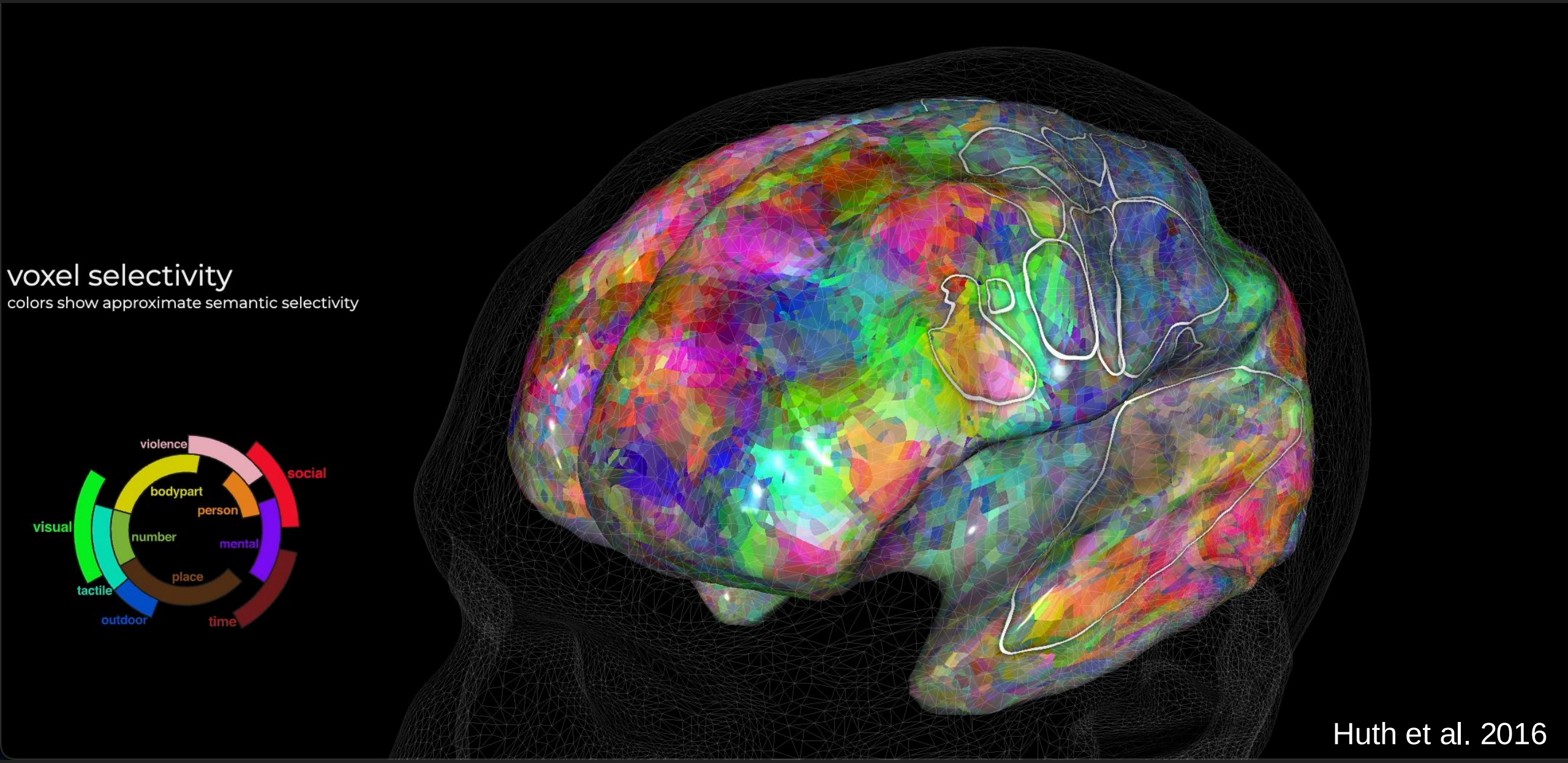
Predicted response

Retrieve

**Limitations:** computational cost, potentially inaccurate

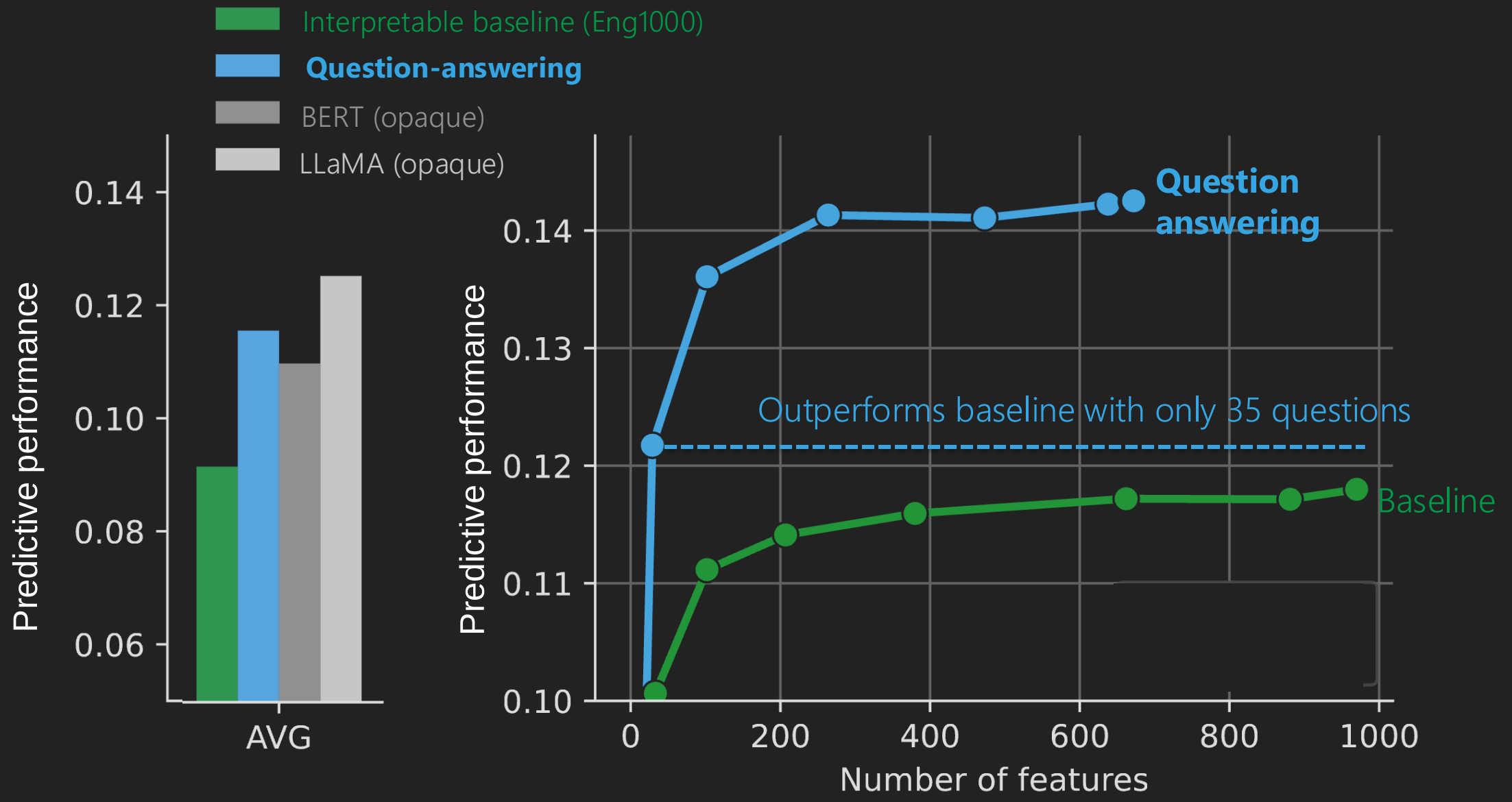
**Other works using yes-no questions:**
Style Embeddings (Patel et al. 2023); CHiLL (McInerney et al. 2023); Tree-Prompting (Morris et al. 2023); BC-LLM (Feng et al. 2024)

# Explaining semantic representations is a fundamental goal of language neuroscience

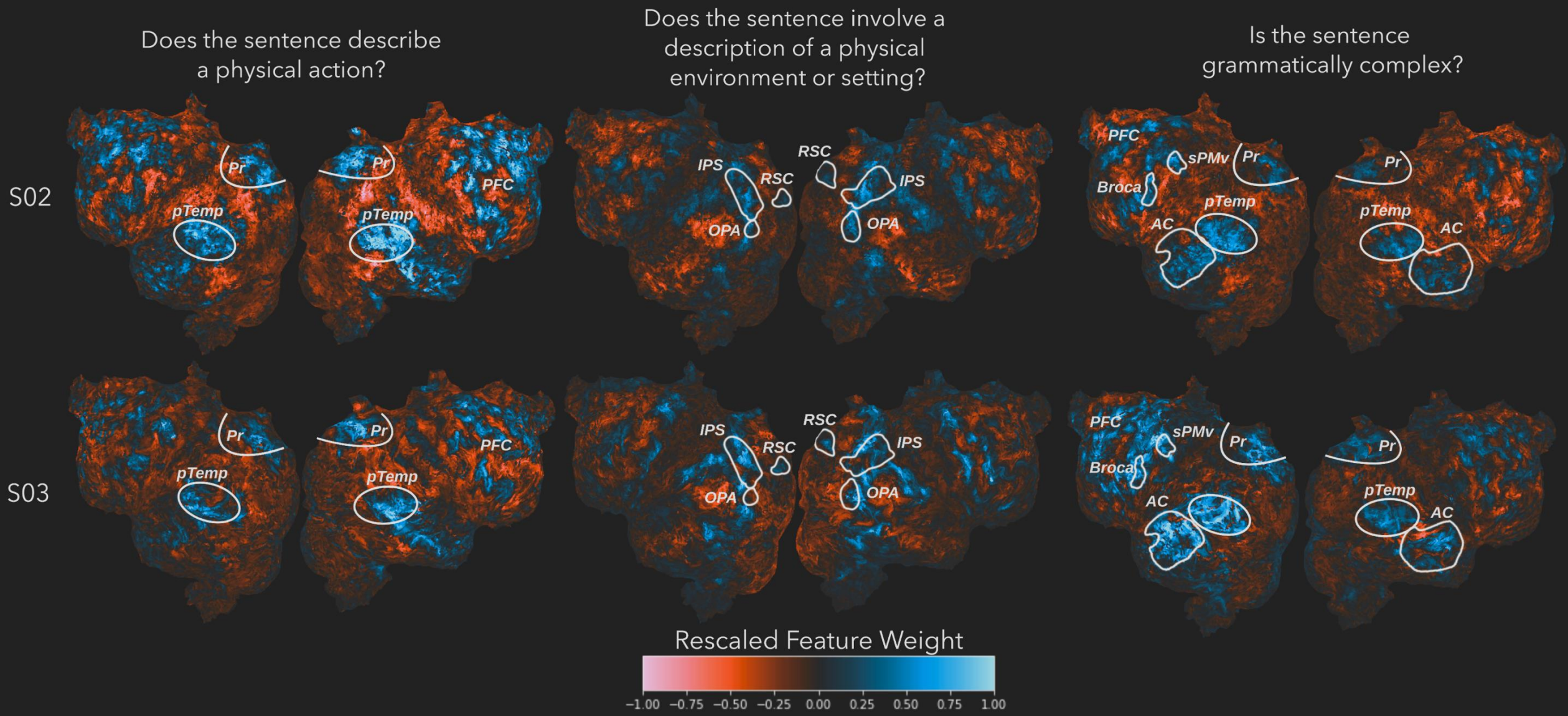



Huth et al. 2016

# A small set of questions predicts well



# Each question yields a cortex-level selectivity map



# Ongoing work: Causally test these explanations

Generative explanation-mediated validation (antonello*, singh* et al. 2024, arXiv)