Blackbox meets Blackbox:

Representational Similarity and Stability Analysis of Neural Language Models and Brains

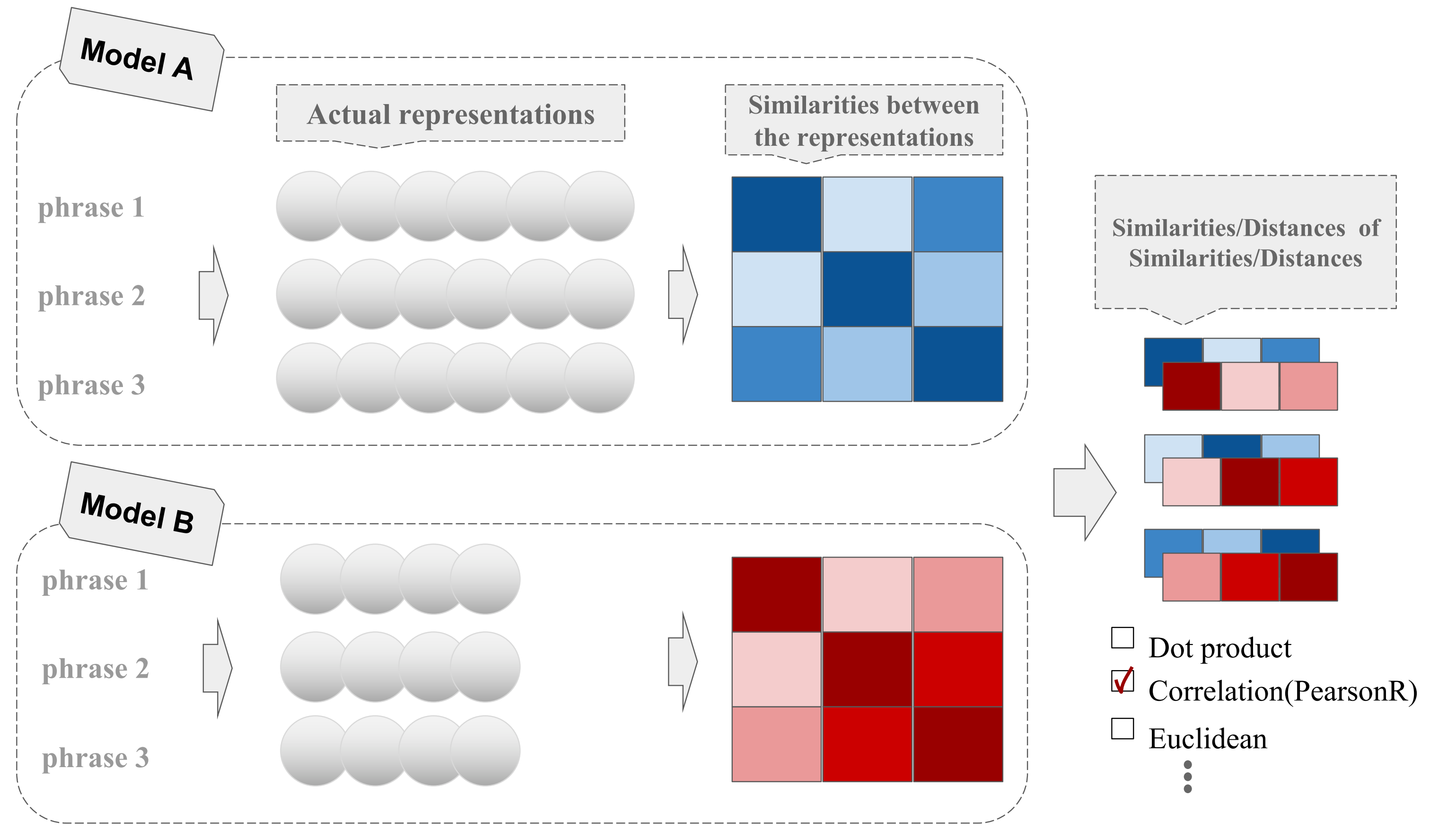
Samira Abnar, Lisa Beinborn, Rochelle choenni, Willem Zuidema

University of Amsterdam, Institute for Logic, Language and Computation

@samiraabnar

LANGUAGE in INTERACTION

RSA and ReStA


Model A
Actual representations
Similarities between the representations
phrase 1
phrase 2
phrase 3
Model B
phrase 1
phrase 2
phrase 3
Similarities/Distances of Similarities/Distances

☐ Dot product
☑ Correlation(PearsonR)
☐ Euclidean

RSA is used, specially in cognitive neuroscience, to compare different representational spaces.

We can use RSA also to study a single model:

- We can compare different components of the model,
- We can compare the representations obtained from the same component of the same model for the same input under different conditions.

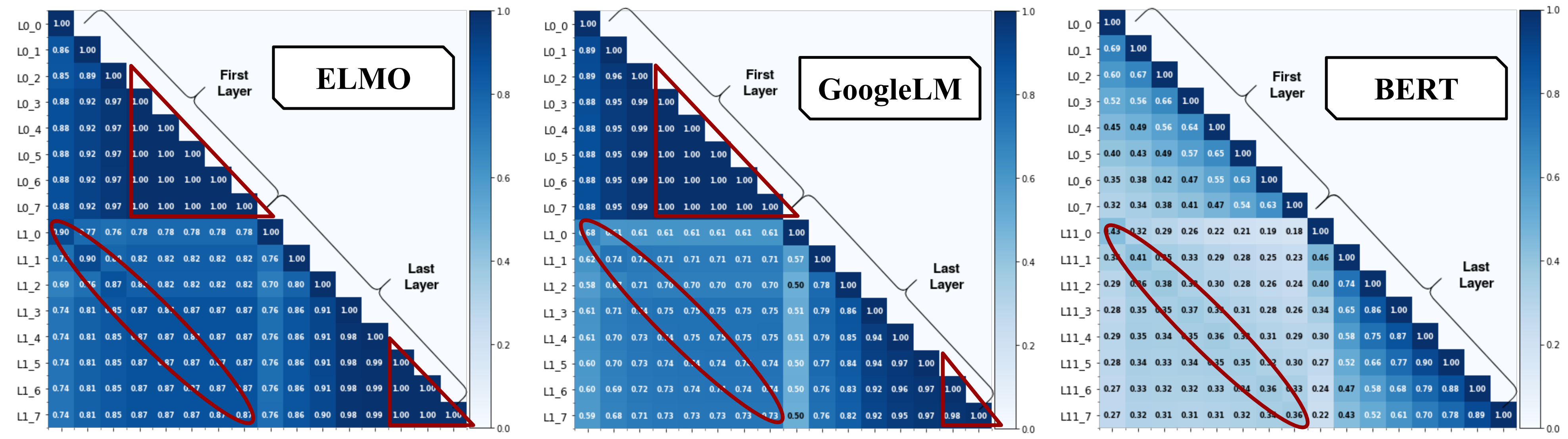
When we use RSA to see how a changes in a single condition affects the representational space of a model, we call it **ReStA.**

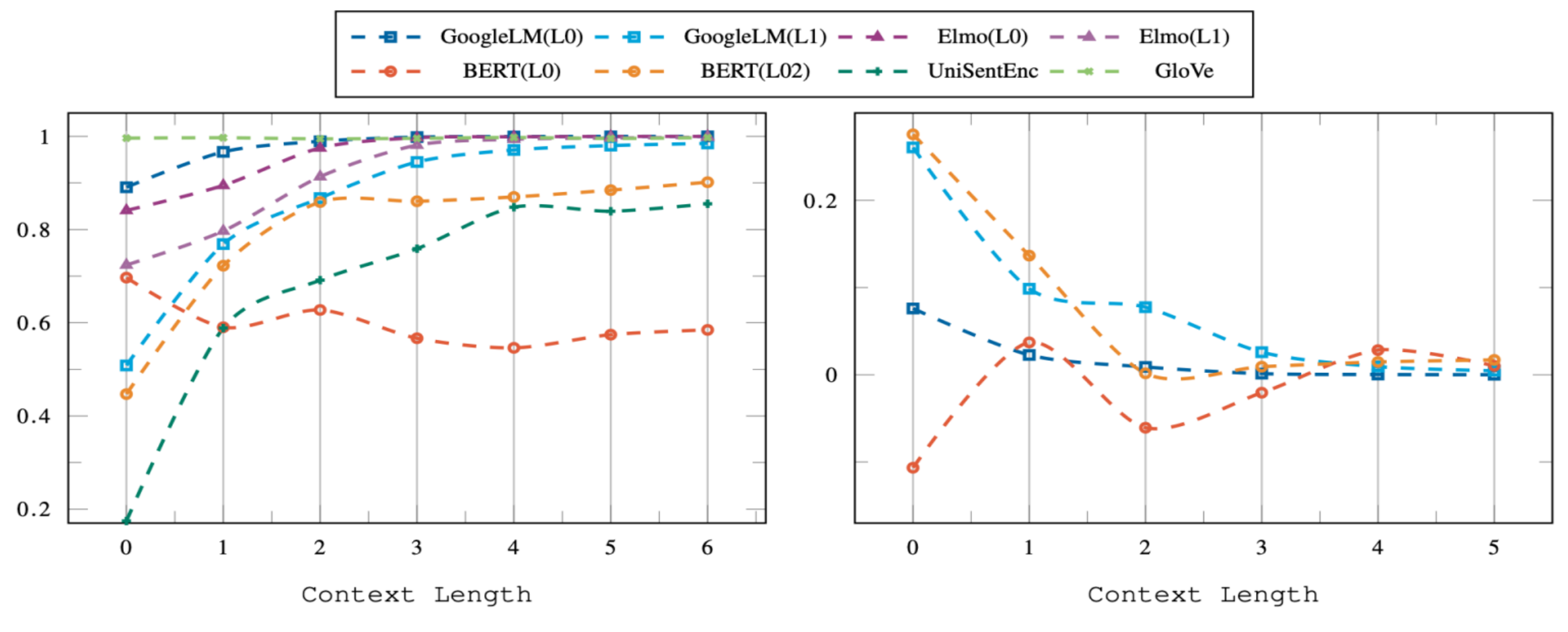
- For example*: How stable are the representations obtained from a language model when the amount of* ***prior context*** *increases.*

- Invariant to linear transformations: scaling, rotation, etc.
- Can compare inherently different representational spaces.
- One interpretation concern: No need to make explicit assumptions on how we expect the model to work.

Language Models meet Language Models

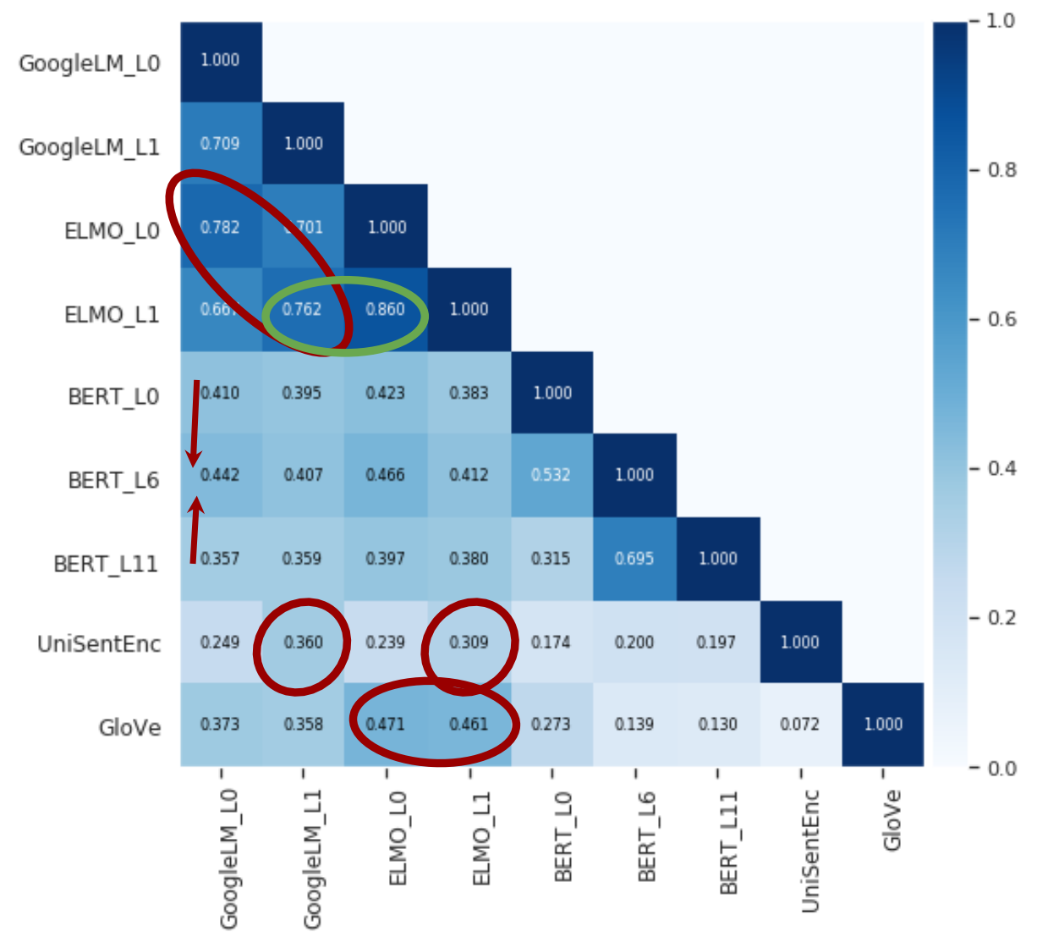
```
1. markings around its eyes.

2. it had the same markings around its eyes.

3. He was sure it was the same one; it had
   the same markings around its eyes.

4. It was now sitting on his garden wall. He
   was sure it was the same one; it had the
   same markings around its eyes.

5. As he pulled into the driveway of number
   four, the first thing he saw -- and it
   didn't improve his mood -- was the tabby
   cat he'd spotted that morning. It was now
   sitting on his garden wall. He was sure
   it was the same one; it had the same
   markings around its eyes.
```

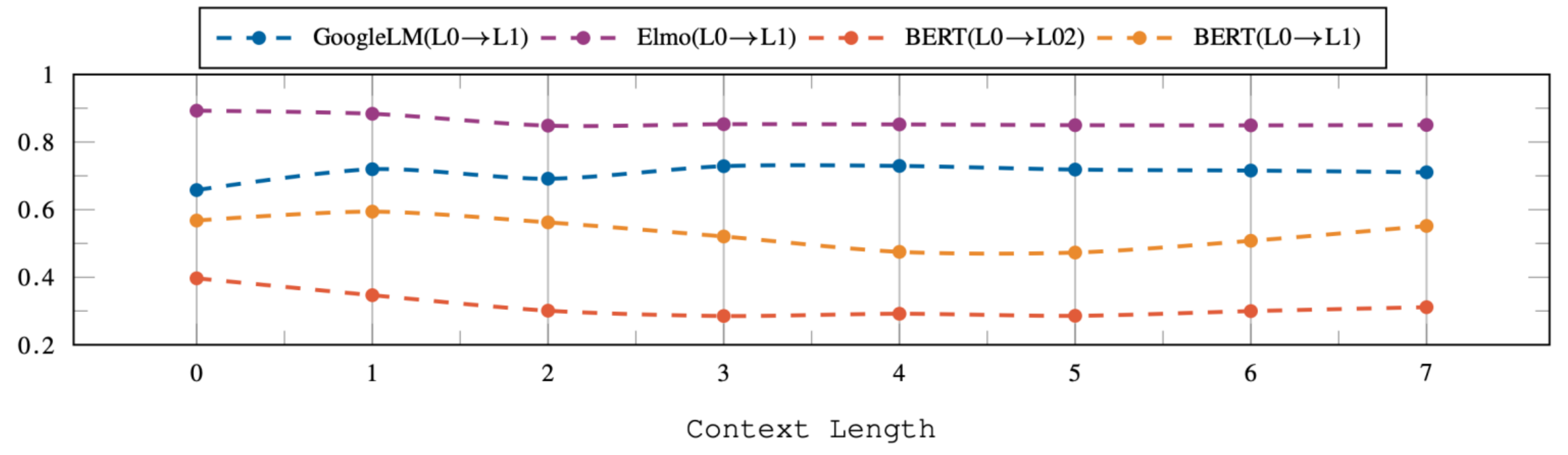




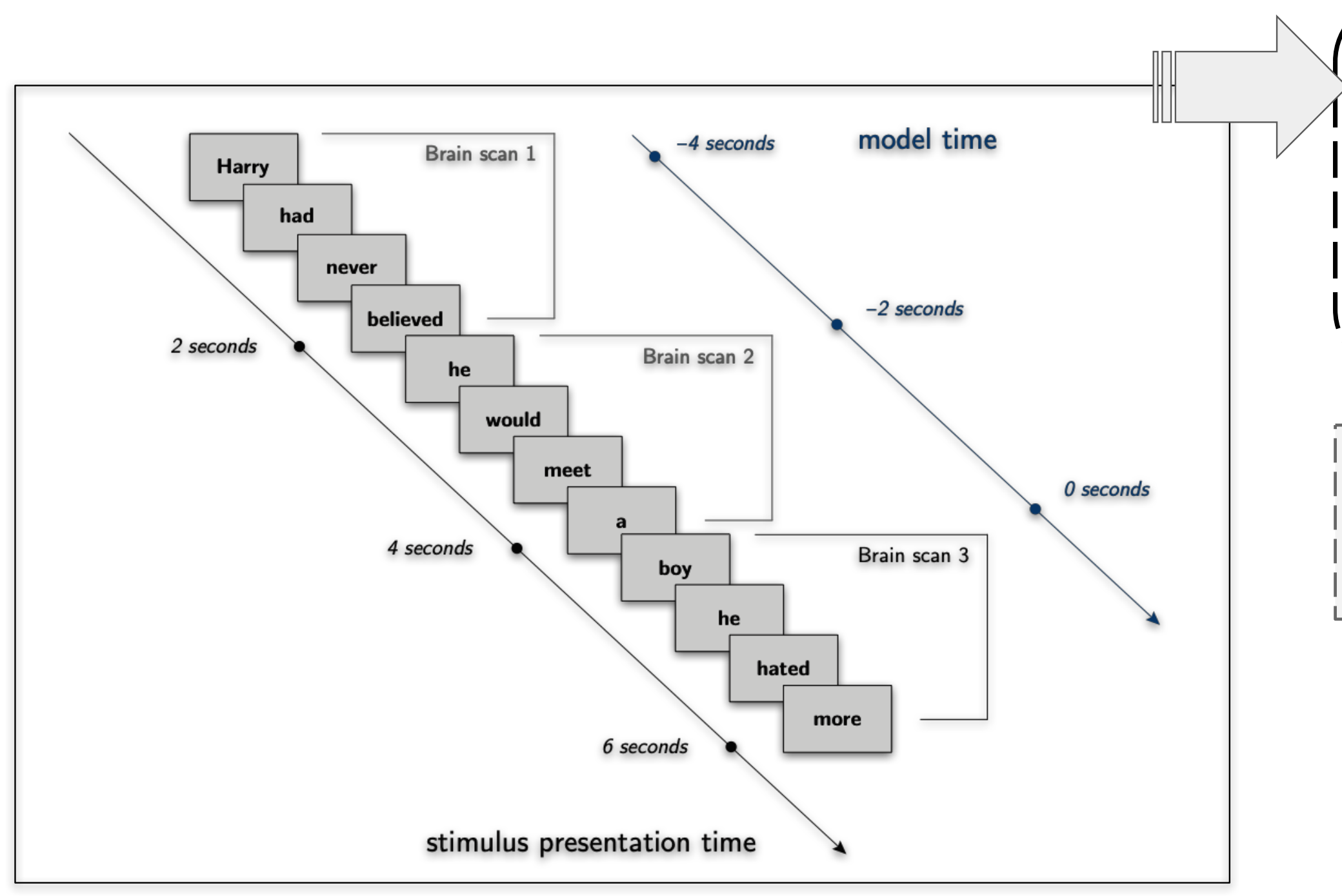
(a) Context Sensitivity ($RSA(L_k c_i, L_k c_{i+1})$) (b) Changes in Context Sensitivity ($\delta RSA(L_k c_i, L_k c_{i+1})$)

- We compare the representations of different layers given different context lengths for **GoogleLM, ELMO, BERT, Universal Sentence Encoder** and **GloVe.**
  - ○ different architectures, different objectives
- We study the *context sensitivity* and the *effect of depth* for each of these models.
- We compare the models with each other and with themselves.



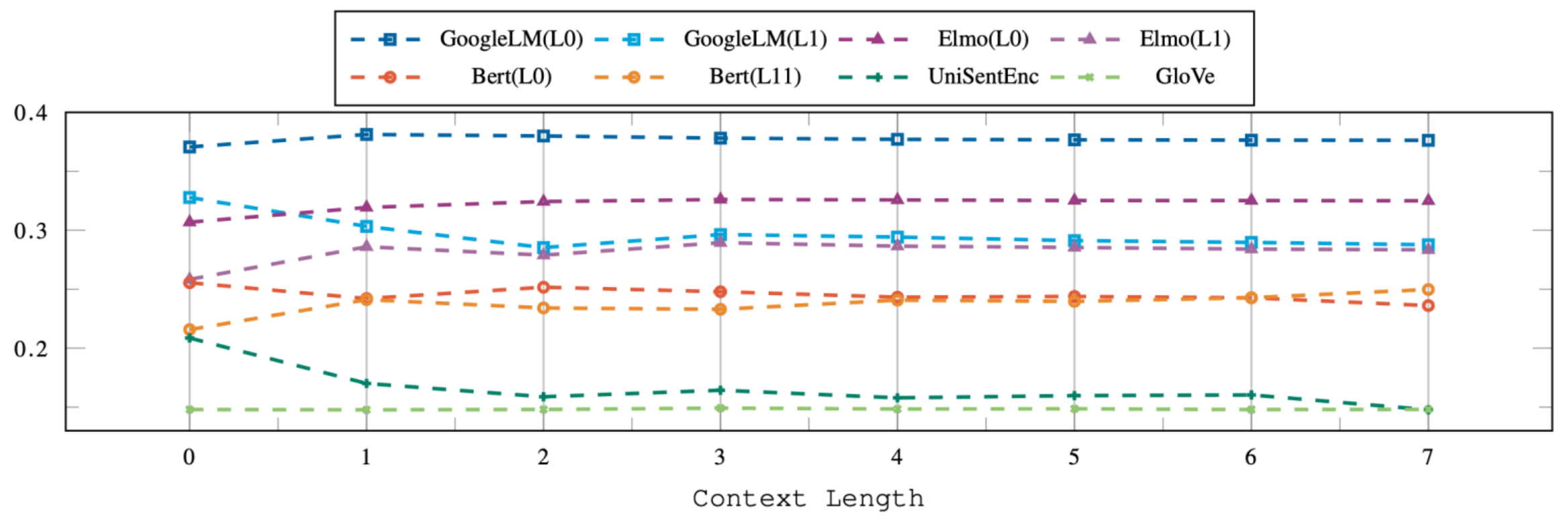


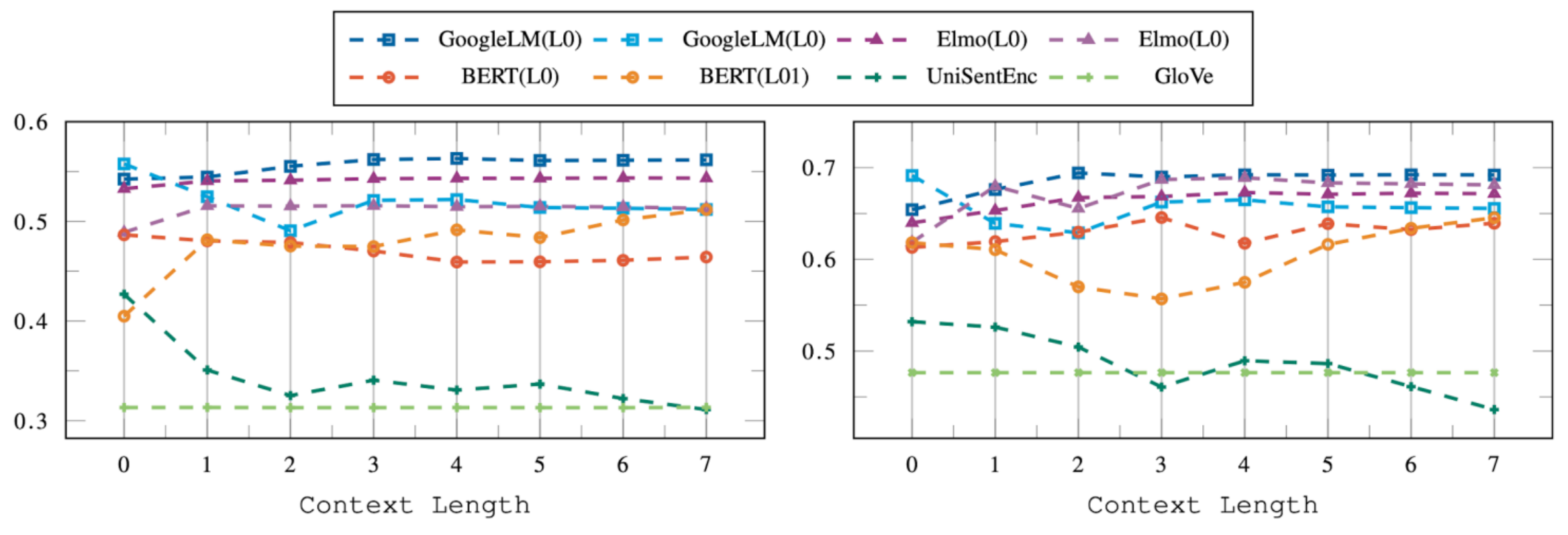
Language Models meet Brain



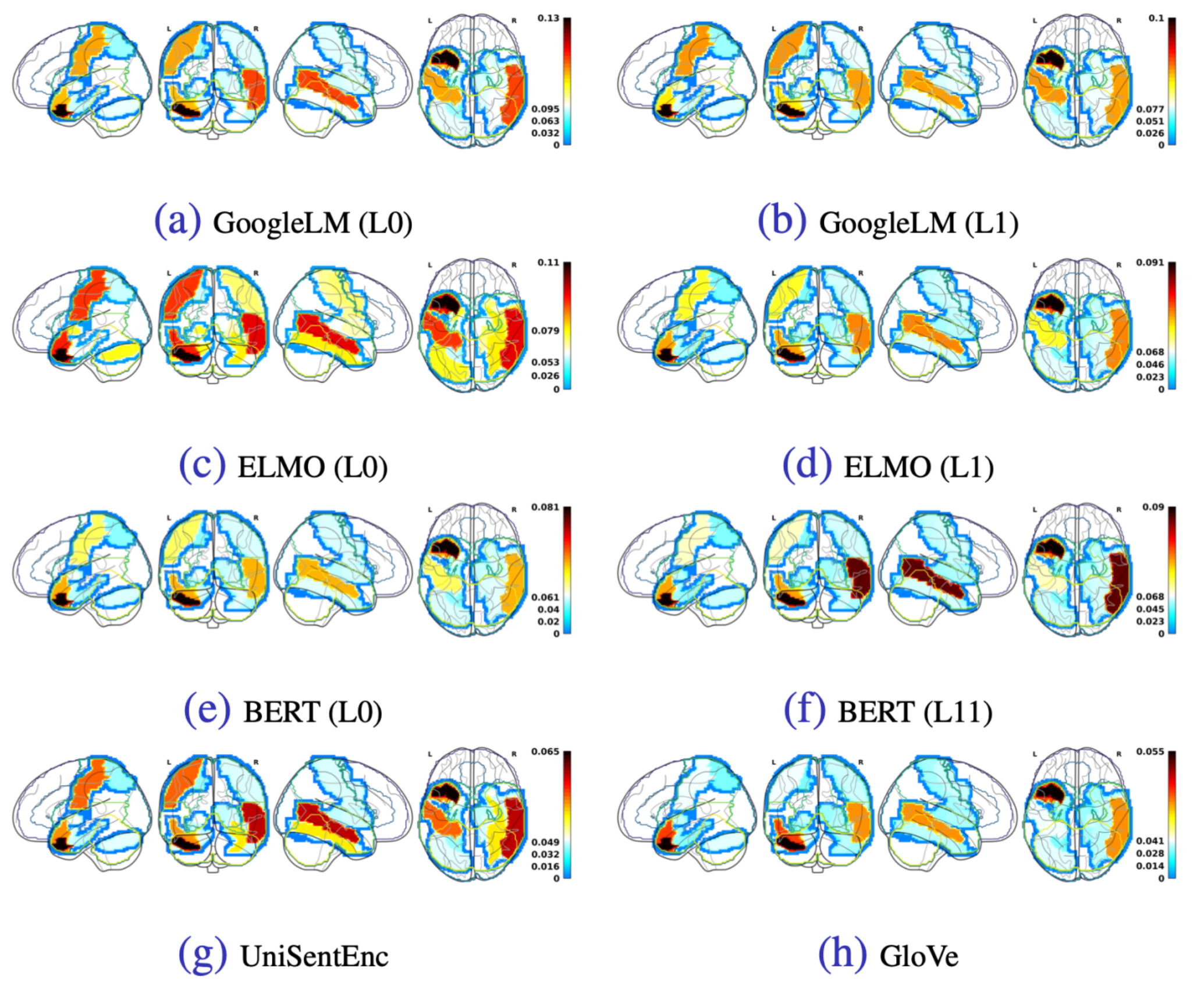
We use the dataset by ***Wehbe 2014*** which consists of the fMRI scans of 8 participants reading chapter 9 of *Harry Potter and the Sorcerer's stone.*

Similarity of the representations obtained from different layers of different models, given different amount of context with brain signals, averaged over all subjects.





(a) Complete sentences (b) Mentions of story character



(a) GoogleLM (L0) (b) GoogleLM (L1) (c) ELMO (L0) (d) ELMO (L1) (e) BERT (L0) (f) BERT (L11) (g) UniSentEnc (h) GloVe

RSA of representations learned at different layers of different models with representations at different regions of subject#4's brain which is chosen randomly. In order to emphasize the difference of the similarity of each model with different brain regions, the color bar is scaled independently for each model. The darkest region for all models is the Left Anterior Temporal Lobe.

Yes! We can learn some things by comparing black boxes.

We can measure the important of a parameter for a language encoding model, by measuring the stability/sensitivity of the representations with respect to that factor.

In the two layer LSTM based language models, the second layers are more sensitive to more context.

BERT continues to respond to more prior context, even if it has already seen a lot.

For language models such as ELMO, context makes a difference but sky is not the limit.

Based on our RSA metric, the lower layers are more similar to fMRI brain signals.

Model architecture plays an important role!

Increasing context length does not affect the similarity of the representations obtained from these models with brain signals.