# Training language models for deeper understanding improves brain alignment

Khai Loong Aw  Mariya Toneva

SMU SINGAPORE MANAGEMENT UNIVERSITY

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

## 1. Summary

### Q: Are NLP models truly learning a deeper understanding of language?

We compare the brain-NLP alignment of 4 models trained with language modeling ("base models") against 4 models trained for deeper understanding on the BookSum narrative summarization dataset ("booksum models").

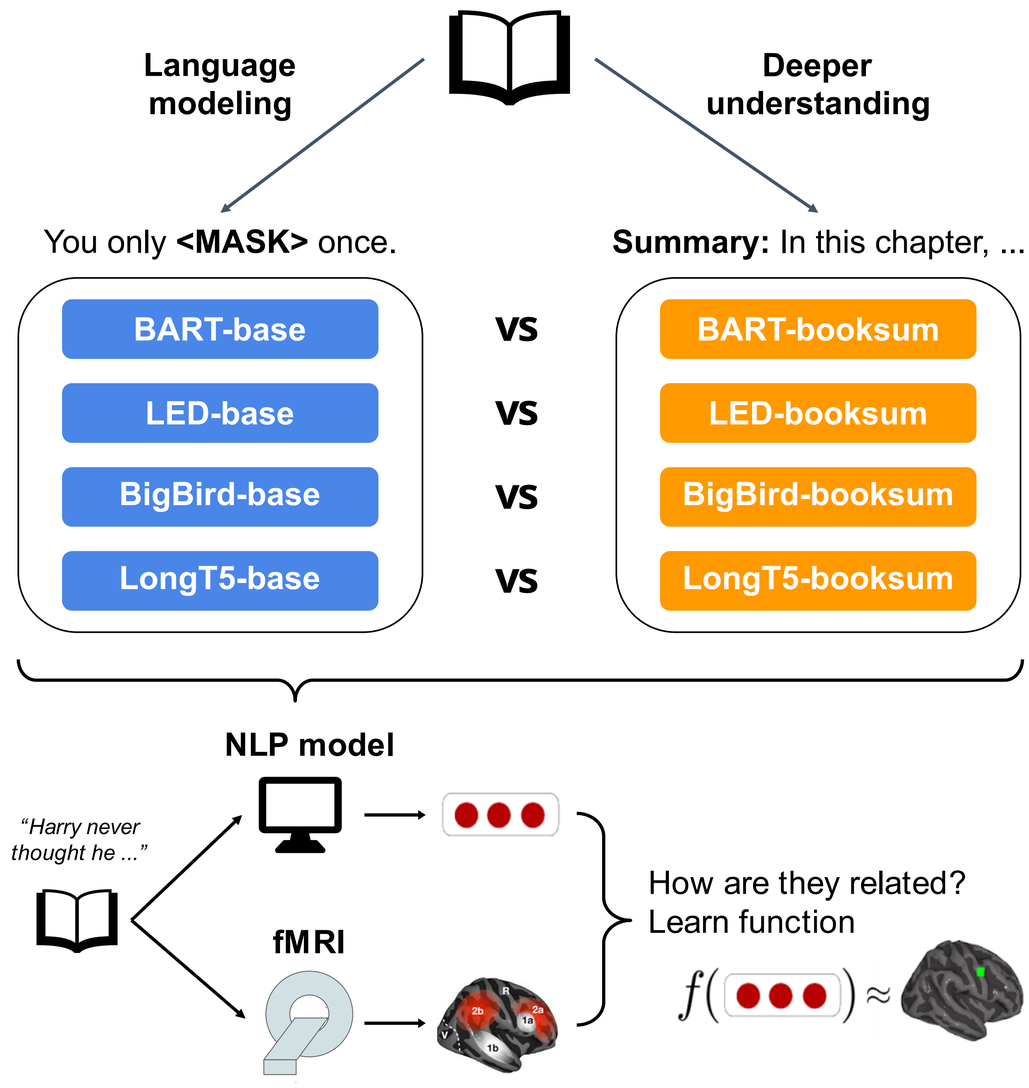
**Conclusions for brain-NLP field:**

- Training language models for deeper understanding improves brain alignment
- Understanding of characters and other discourse features is a significant factor in brain-NLP alignment

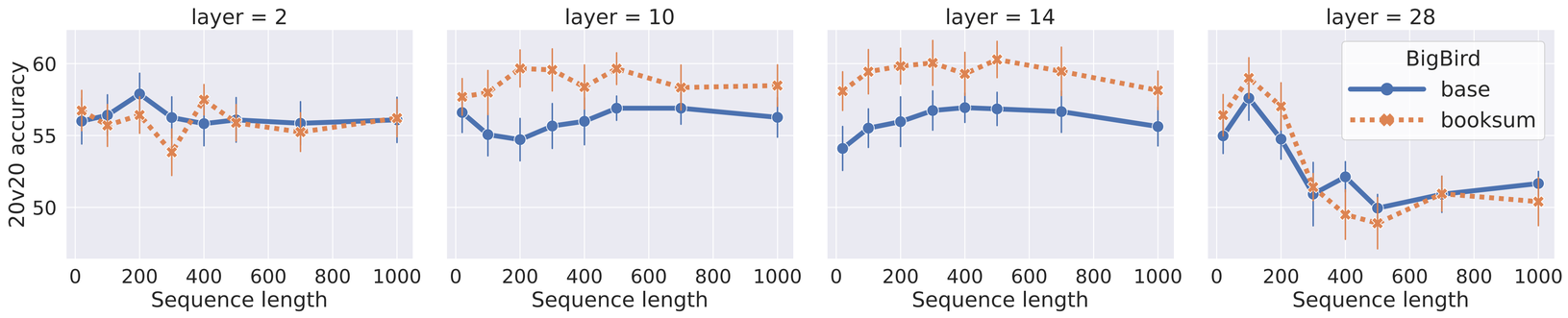
**Conclusions for NLP:**

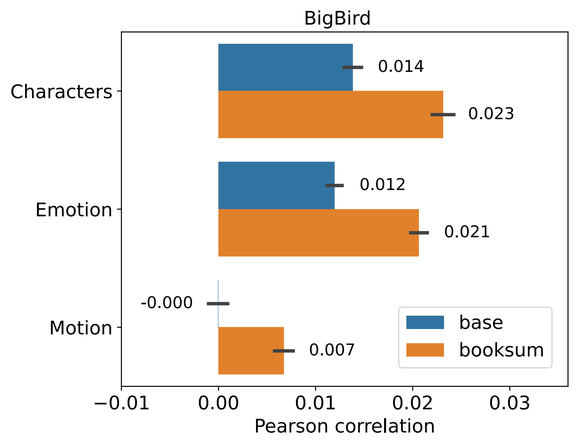
- Existing training methods for narrative understanding indeed develop deeper language understanding
- Language modeling (LM) achieves poorer representations and worse brain alignment — we should explore training strategies beyond LM

## 2. Approach




## 3. Results



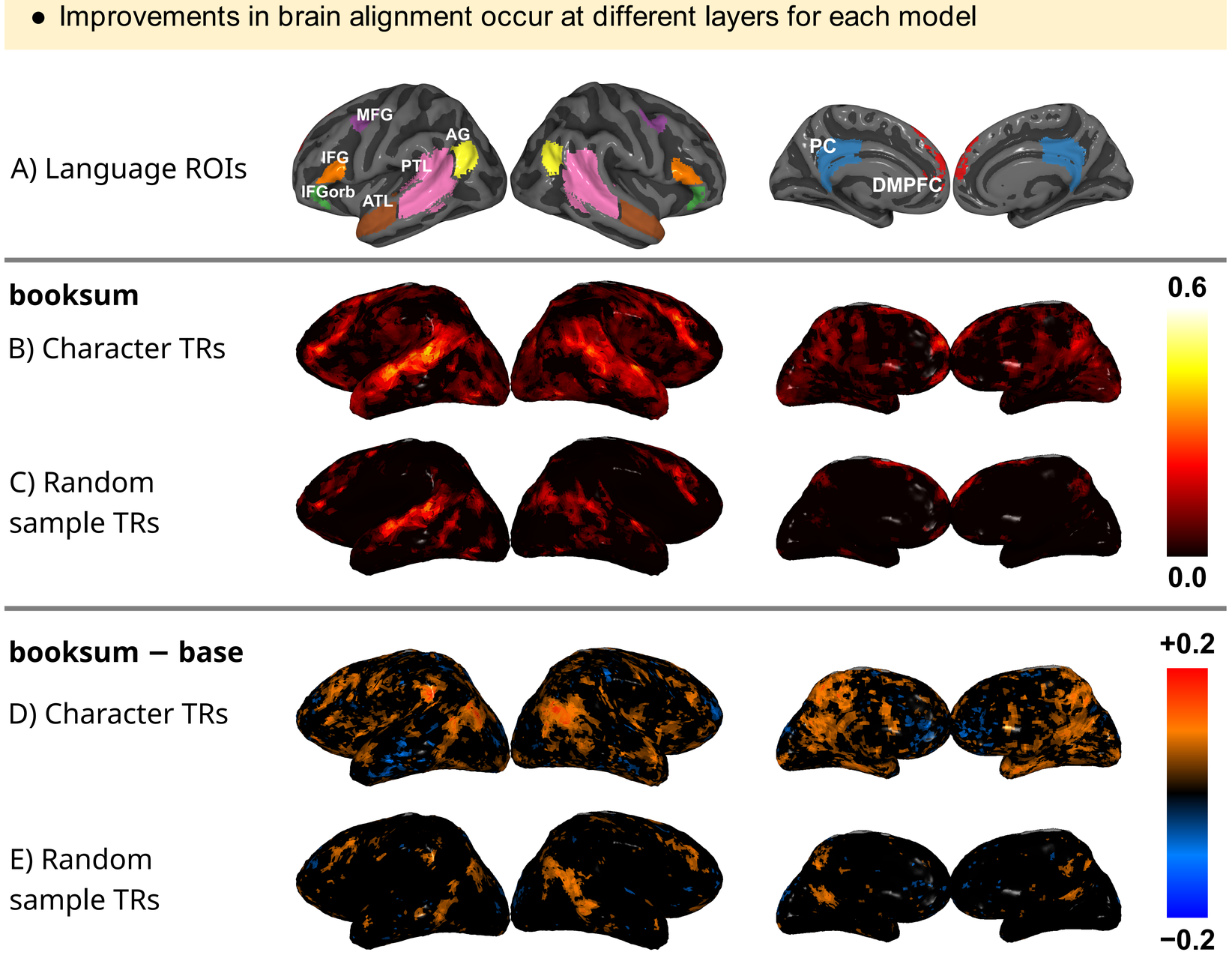




**Training language models for deeper understanding improves brain alignment**

- Brain alignment peaks around 500 words of context
- Improvements in brain alignment occur at different layers for each model

**NLP models learned richer representations across all tested discourse features**

- Characters has the greatest brain-NLP alignment, and also improved the most when trained for deeper understanding




**Comparing row B and C:** Greater brain alignment for fMRI intervals (TRs) that contain Characters than for random sample TRs, both inside and outside language regions

**Comparing row D and E:** Character TRs improve in more brain voxels (orange regions) than random sample TRs