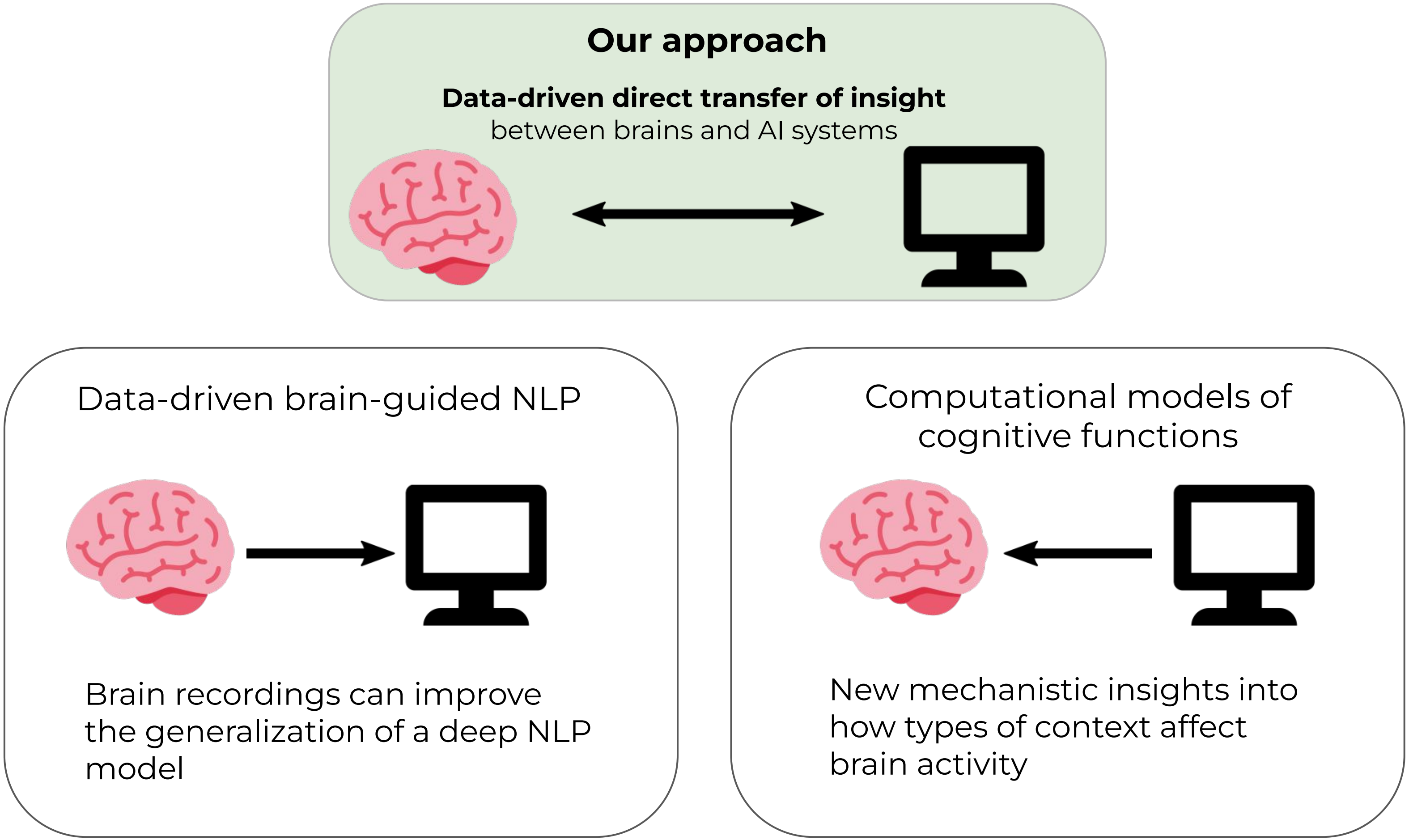
# Why do large language models align with human brains: insights, opportunities, and challenges

Mariya Toneva

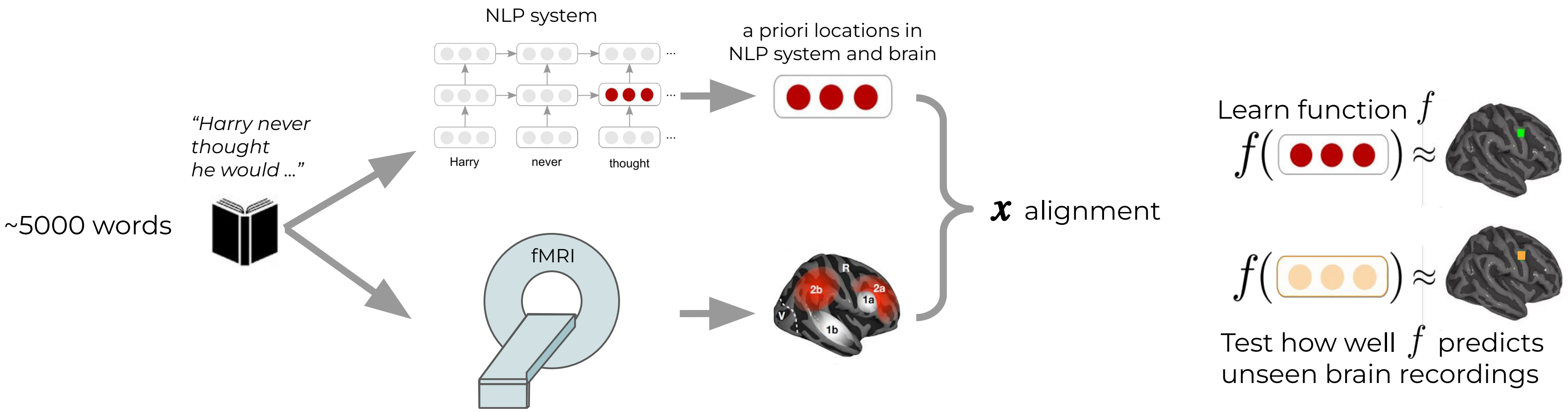
Nov 17, 2022

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

**Br**idging
**AI** and
**N**euroscience
Group


Our approach
Data-driven direct transfer of insight
between brains and AI systems
Data-driven brain-guided NLP
Brain recordings can improve the generalization of a deep NLP model
Computational models of cognitive functions
New mechanistic insights into how types of context affect brain activity

# (M)LMs already align with human brain recordings to an impressive degree

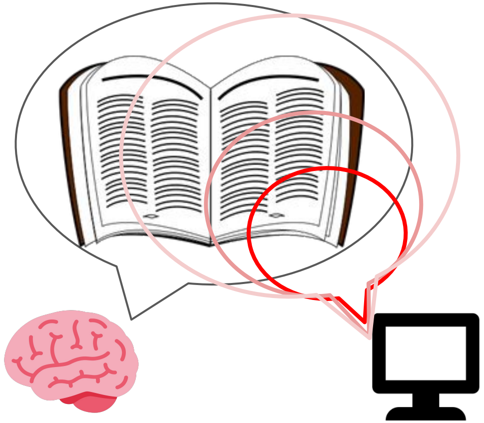



Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

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

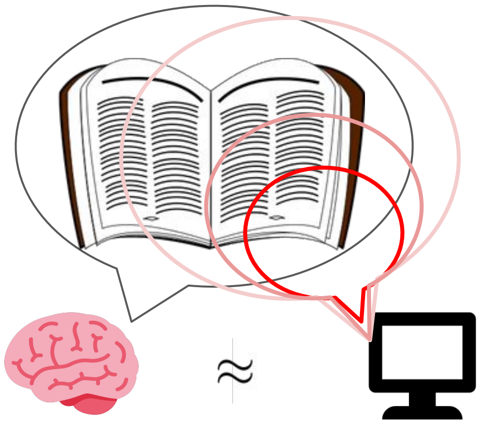
Jain et al. 2020,
Schrimpf et al. 2021,
Goldstein et al. 2022
...

# Challenge for neural networks for NLP: long-term dependencies

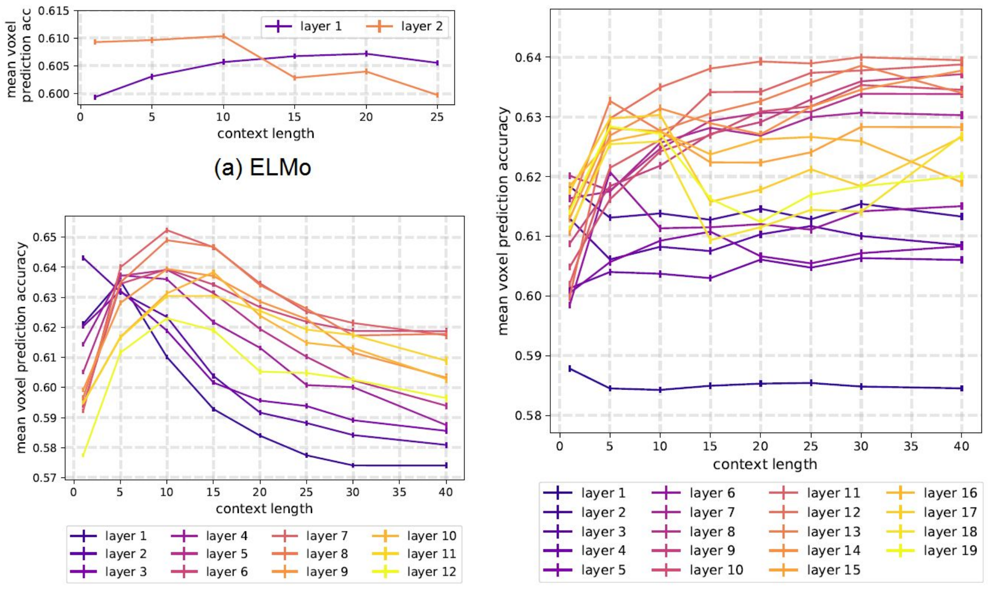


Goodfellow, Bengio, Courville 2016
Khandelwal et al. 2018
Dai et al. 2019

# Vary the context length and observe how alignment with fMRI recordings changes



# Vary the context length and observe how alignment with fMRI recordings changes




(a) ELMo

(b) BERT

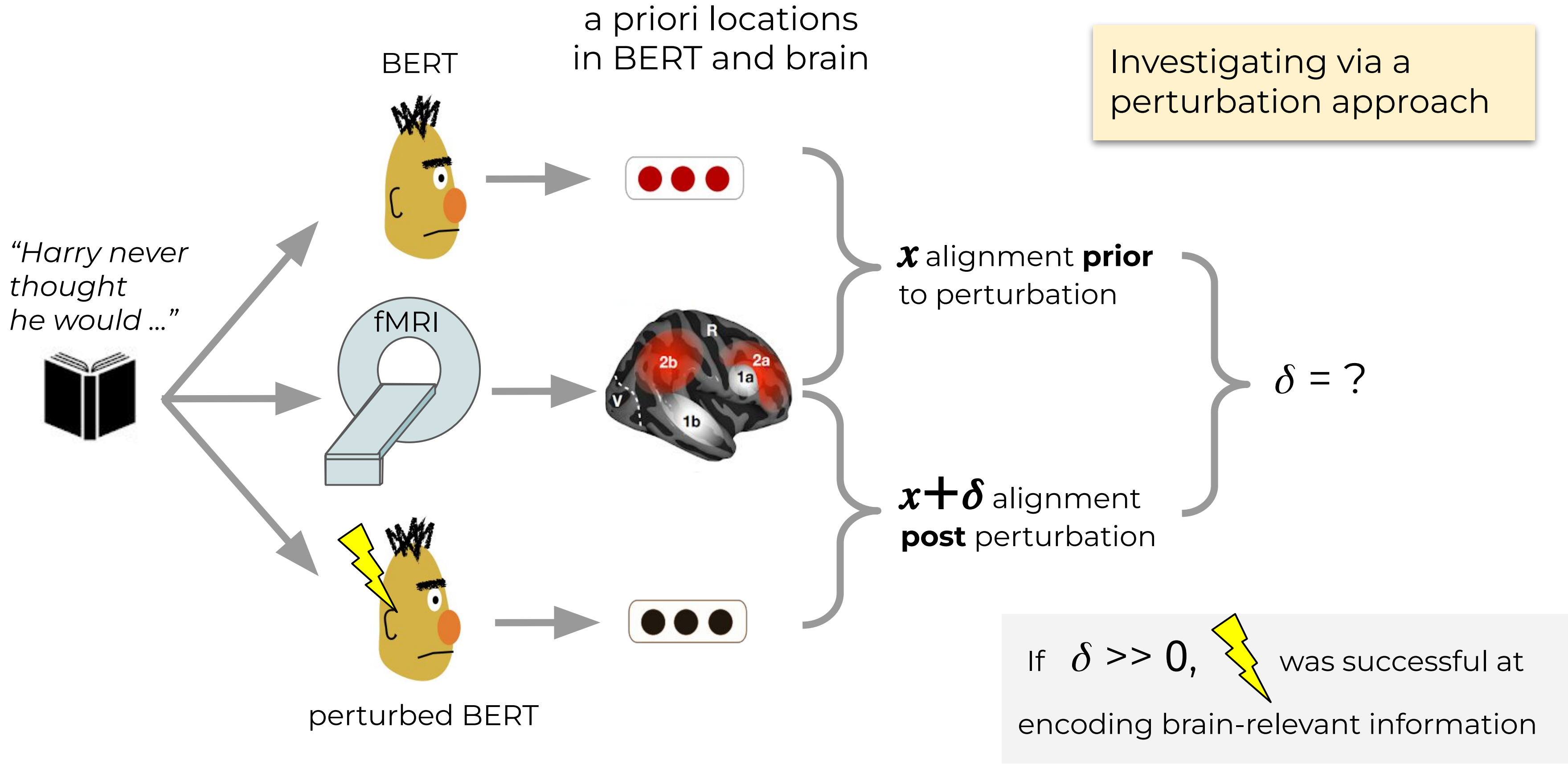
(c) T-XL

middle layers align best with the brain recordings

only Transformer-XL continues to increase alignment (up to ~50 words) as the context length is increased

Toneva and Wehbe, *NeurIPS* 2019

# What are the reasons for this alignment?



# Today: evidence from 3 perturbation case studies

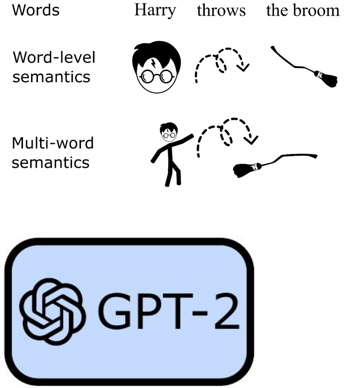
1. Alignment due to **more** than next-word prediction & word-level semantics
   [Merlin & Toneva, 2022 arXiv soon]
2. Joint processing of linguistic properties
   [Oota, Gupta, and Toneva 2022 arXiv soon]
3. Training to summarize narratives improves brain alignment [Aw & Toneva, 2022 In Submission]

# Today: evidence from 3 perturbation case studies

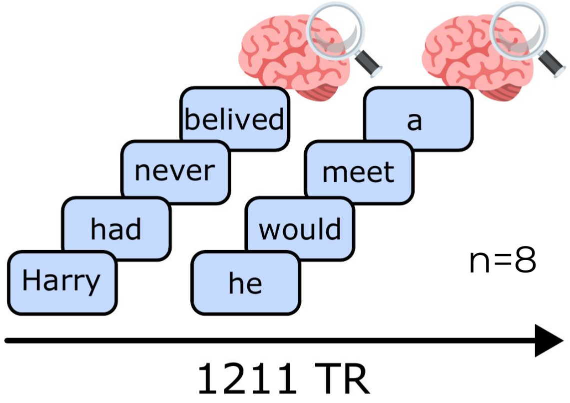
1. **Alignment due to more than next-word prediction & word-level semantics**
   [Merlin & Toneva, 2022 arXiv https://arxiv.org/abs/2212.00596]
2. Joint processing of linguistic properties
   [Oota, Gupta, and Toneva 2022 arXiv https://arxiv.org/abs/2212.08094]
3. Training to summarize narratives improves brain alignment [Aw & Toneva, ICLR 2023 https://arxiv.org/abs/2212.10898]

# Case study 1

- Next-word prediction performance correlates with brain alignment [Schrimpf et al. 2021, Goldstein et al. 2022]
- Necessary or simply sufficient?



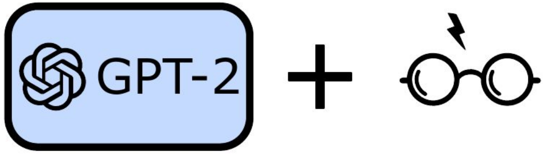

fMRI Harry Potter Dataset
[Wehbe et al. 2014]



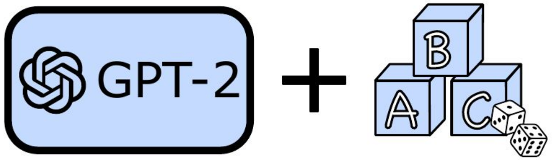

Gabriele Merlin

Language models and brain alignment: beyond word-level semantics and prediction, Merlin and Toneva 2022 arXiv [https://arxiv.org/abs/2212.00596](https://arxiv.org/abs/2212.00596)

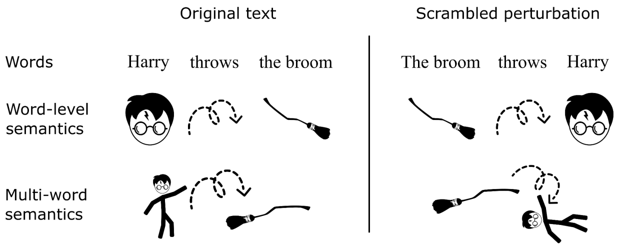
# Perturbations



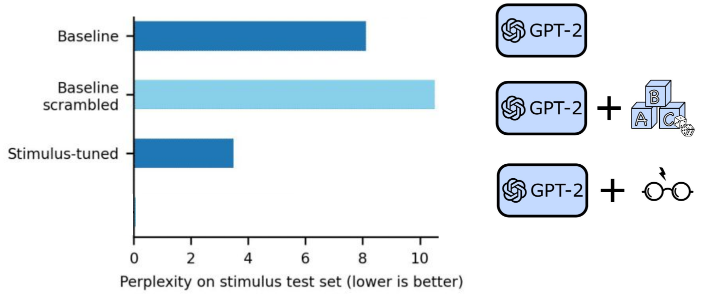

Perturbation 1: **stimulus-tuning**
- fine-tune with LM objective on Harry Potter stimulus
- expected to affect:
    - next-word prediction
    - word-level semantics
    - multi-word semantics




Perturbation 2: **scrambling**
- scramble input words
- expected to affect:
    - next-word prediction
    - multi-word semantics
- **not** expected to affect word-level semantics

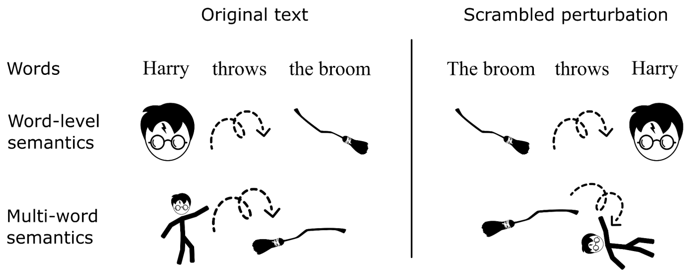


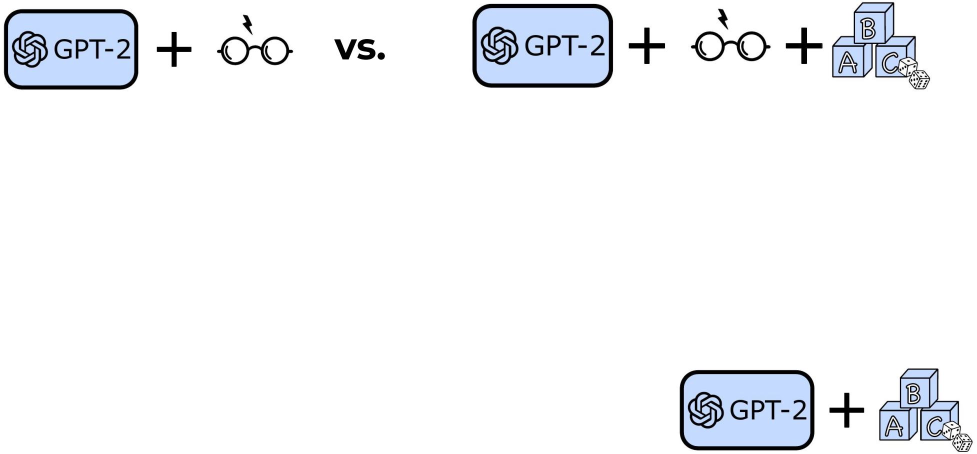
# Next-word prediction capabilities affected as expected




Scrambling decreases next-word prediction performance

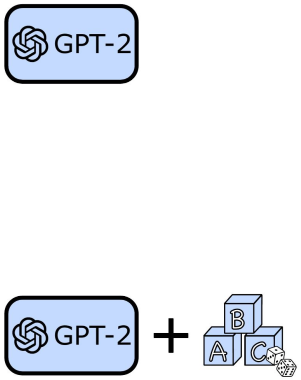
Stimulus-tuning increases next-word prediction performance

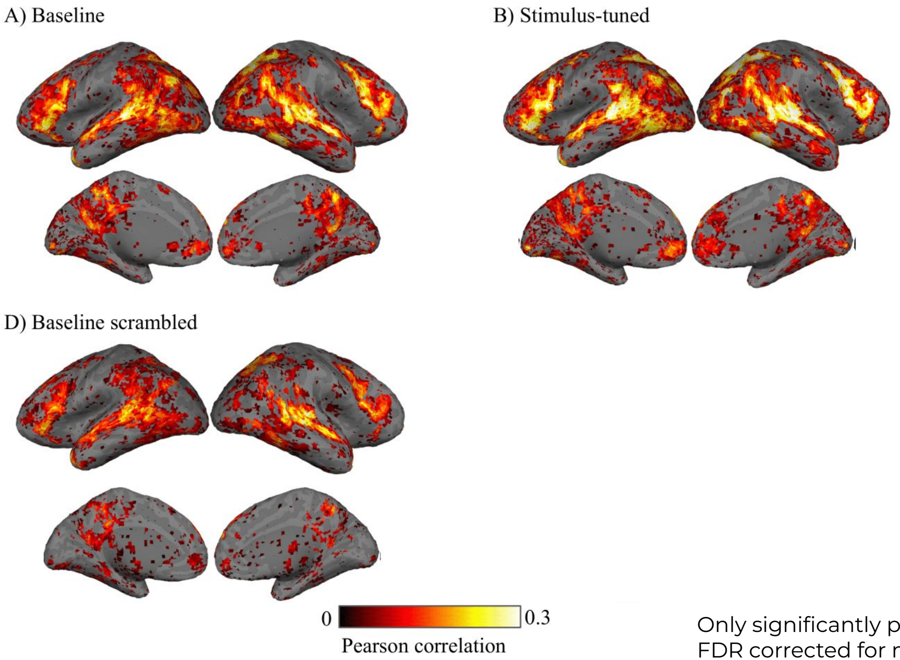


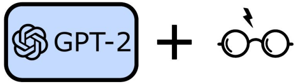


GPT-2
vs.
GPT-2
GPT-2

# Voxel-wise brain alignment

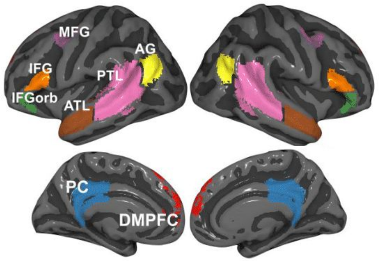


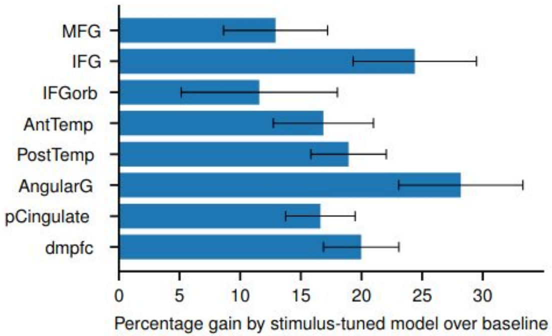




Only significantly predicted voxels displayed (permutation test,
FDR corrected for multiple comparisons

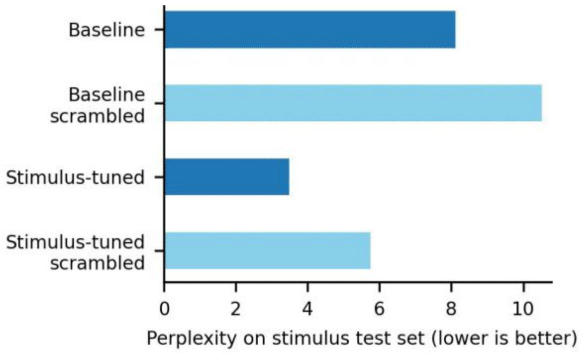
# Stimulus-tuning improves brain alignment across all language regions

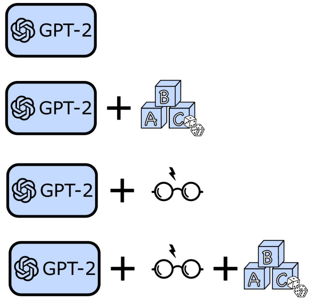



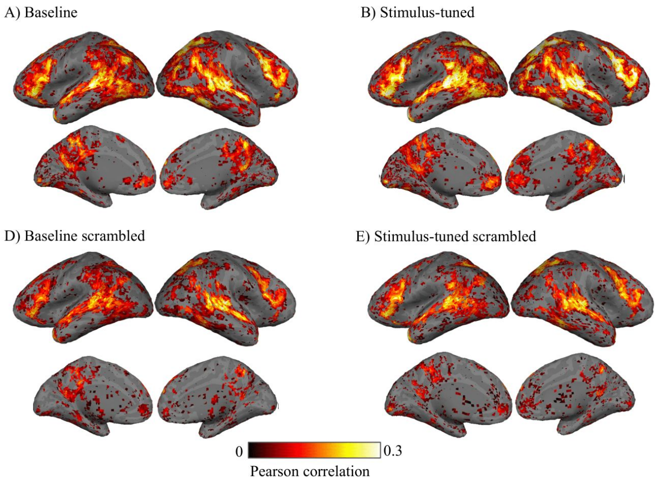




- Improvement could be due to:
  - next-word prediction
  - word-level semantics
  - multi-word semantics

# Combining perturbations reveals a divergence in trends for LM performance and brain alignment

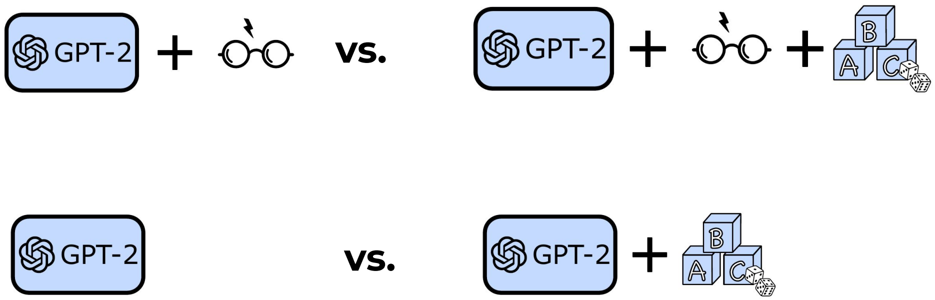


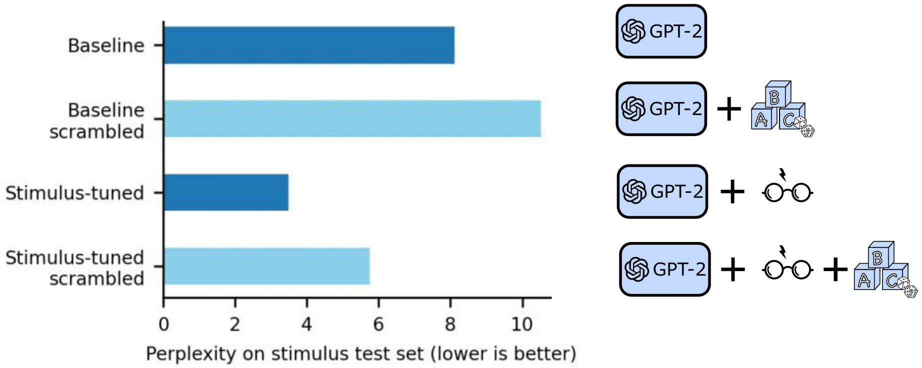





# Contrast to control for word-level semantics **and** next-word prediction performance

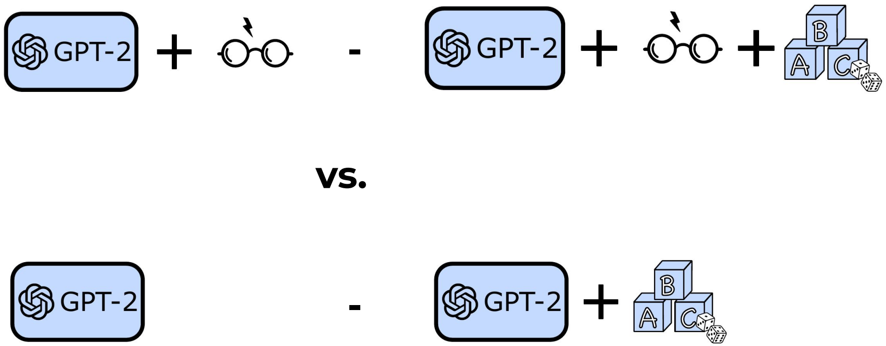


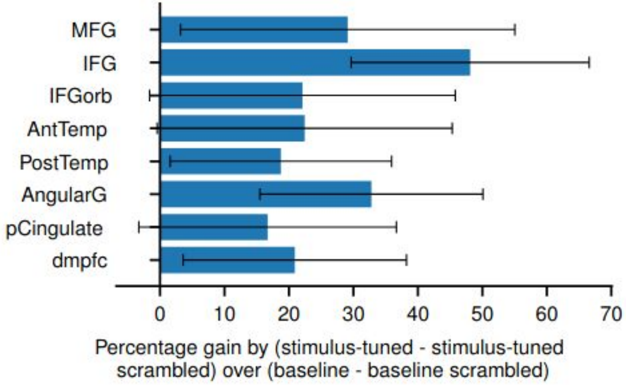

Controls for word-level semantics,
But **not** next-word prediction

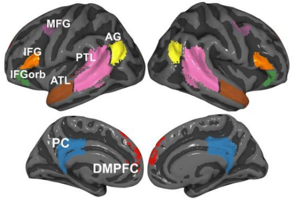



**Key idea:** make use of constant difference in next-word prediction

# Alignment with Angular Gyrus and IFG due to more than word-level semantics and next-word prediction



# Today: evidence from 3 perturbation case studies

1. Alignment due to more than next-word prediction & word-level semantics
   [Merlin & Toneva, 2022 arXiv https://arxiv.org/abs/2212.00596]
2. **Joint processing of linguistic properties**
   [Oota, Gupta, and Toneva 2022 arXiv https://arxiv.org/abs/2212.08094]
3. Training to summarize narratives improves brain alignment [Aw & Toneva, ICLR 2023 https://arxiv.org/abs/2212.10898]

# Case study 2

- Best brain alignment observed with middle layers of LMs [Jain and Huth 2018, Toneva and Wehbe 2019, Caucheteux and King 2020]
- Thought to be because of high-level information equally-distant from word-level input
- But BERTology tells us that middle layers best for syntactic processing [Jawahar et al. 2019, Rogers et al. 2020]

Subba Reddy Oota

What linguistic properties underlie brain alignment, across all layers but also specifically in middle layers?

Joint processing of linguistic properties in brains and language models, Reddy, Gupta, and Toneva 2022 arXiv https://arxiv.org/abs/2212.08094

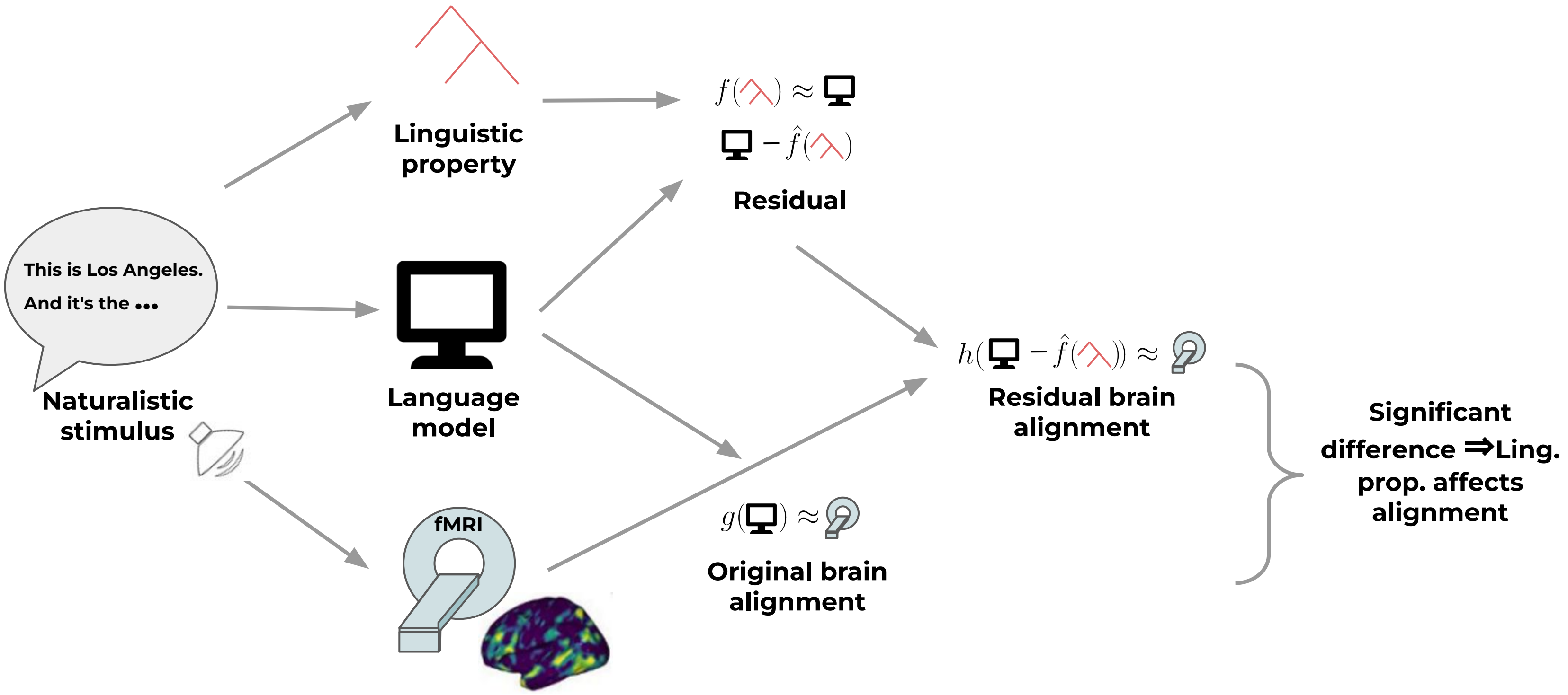
# Investigating effect of surface, syntactic, and semantic linguistic properties

| Layer | SentLen (Surface) | WC (Surface) | TreeDepth (Syntactic) | TopConst (Syntactic) | BShift (Syntactic) | Tense (Semantic) | SubjNum (Semantic) | ObjNum (Semantic) | SOMO (Semantic) | CoordInv (Semantic) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 93.9 (2.0) | 24.9 (24.8) | 35.9 (6.1) | 63.6 (9.0) | 50.3 (0.3) | 82.2 (18.4) | 77.6 (10.2) | 76.7 (26.3) | 49.9 (-0.1) | 53.9 (3.9) |
| 2 | 95.9 (3.4) | 65.0 (64.8) | 40.6 (11.3) | 71.3 (16.1) | 55.8 (5.8) | 85.9 (23.5) | 82.5 (15.3) | 80.6 (17.1) | 53.8 (4.4) | 58.5 (8.5) |
| 3 | **96.2 (3.9)** | 66.5 (66.0) | 39.7 (10.4) | 71.5 (18.5) | 64.9 (14.9) | 86.6 (23.8) | 82.0 (14.6) | 80.3 (16.6) | 55.8 (5.9) | 59.3 (9.3) |
| 4 | 94.2 (2.3) | **69.8 (69.6)** | 39.4 (10.8) | 71.3 (18.3) | 74.4 (24.5) | 87.6 (25.2) | 81.9 (15.0) | 81.4 (19.1) | 59.0 (8.5) | 58.1 (8.1) |
| 5 | 92.0 (0.5) | 69.2 (69.0) | 40.6 (11.8) | 81.3 (30.8) | 81.4 (31.4) | 89.5 (26.7) | 85.8 (19.4) | 81.2 (18.6) | 60.2 (10.3) | 64.1 (14.1) |
| 6 | 88.4 (-3.0) | 63.5 (63.4) | **41.3 (13.0)** | 83.3 (36.6) | 82.9 (32.9) | 89.8 (27.6) | **88.1 (21.9)** | 82.0 (20.1) | 60.7 (10.2) | 71.1 (21.2) |
| 7 | 83.7 (-7.7) | 56.9 (56.7) | 40.1 (12.0) | **84.1 (39.5)** | 83.0 (32.9) | 89.9 (27.5) | 87.4 (22.2) | **82.2 (21.1)** | 61.6 (11.7) | 74.8 (24.9) |
| 8 | 82.9 (-8.1) | 51.1 (51.0) | 39.2 (10.3) | 84.0 (39.5) | 83.9 (33.9) | 89.9 (27.6) | 87.5 (22.2) | 81.2 (19.7) | 62.1 (12.2) | 76.4 (26.4) |
| 9 | 80.1 (-11.1) | 47.9 (47.8) | 38.5 (10.8) | 83.1 (39.8) | **87.0 (37.1)** | **90.0 (28.0)** | 87.6 (22.9) | 81.8 (20.5) | 63.4 (13.4) | **78.7 (28.9)** |
| 10 | 77.0 (-14.0) | 43.4 (43.2) | 38.1 (9.9) | 81.7 (39.8) | 86.7 (36.7) | 89.7 (27.6) | 87.1 (22.6) | 80.5 (19.9) | 63.3 (12.7) | 78.4 (28.1) |
| 11 | 73.9 (-17.0) | 42.8 (42.7) | 36.3 (7.9) | 80.3 (39.1) | 86.8 (36.8) | 89.9 (27.8) | 85.7 (21.9) | 78.9 (18.6) | 64.4 (14.5) | 77.6 (27.9) |
| 12 | 69.5 (-21.4) | 49.1 (49.0) | 34.7 (6.9) | 76.5 (37.2) | 86.4 (36.4) | 89.5 (27.7) | 84.0 (20.2) | 78.7 (18.4) | **65.2 (15.3)** | 74.9 (25.4) |

Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

Jawahar et al. 2019

# Perturbation approach to evaluate effect of linguistic property on brain alignment



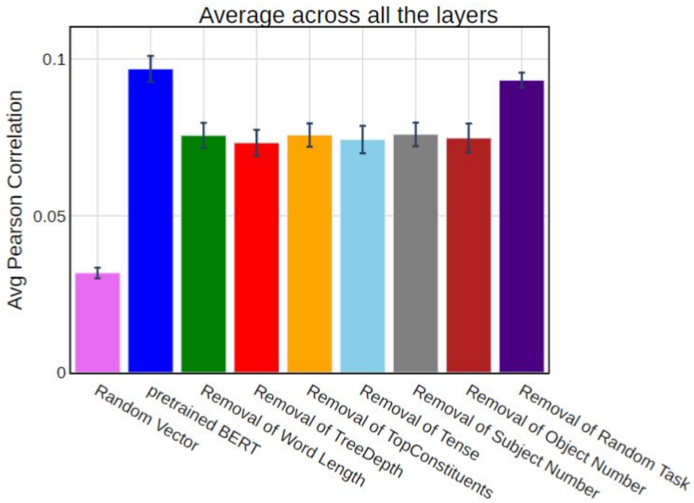
# Datasets & Model

- Brain: fMRI recordings from Narratives [Nastase et al. 2021]
    - listening to a story
    - n=18
- Annotate linguistic properties using Stanford core-NLP stanza library [Manning et al. 2014]
- BERT base (12 layers, pretrained)

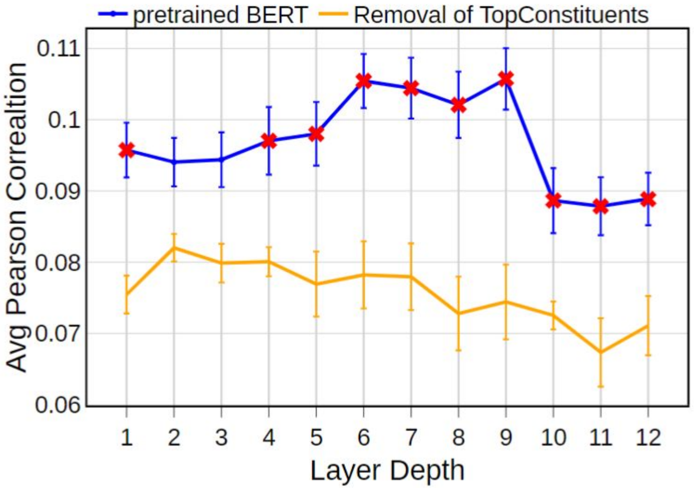
# (Linear) contribution of linguistic properties is successfully removed from BERT

| Layers | Word Length 5-classes (Surface) | | TreeDepth 8-classes (Syntactic) | | TopConst 20-classes (Syntactic) | | Tense 2-classes (Semantic) | | SubjNum 2-classes (Semantic) | | ObjNum 2-classes (Semantic) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | before | after | before | after | before | after | before | after | before | after | before | after |
| 1 | 32.14 | 03.40 | 32.41 | 13.20 | 42.13 | 20.71 | 70.53 | 53.51 | 86.16 | 41.80 | 88.39 | 53.94 |
| 2 | 30.80 | 14.48 | 32.73 | 15.56 | 52.05 | 30.35 | 68.30 | 56.50 | 88.39 | 54.37 | 85.50 | 58.17 |
| 3 | 31.69 | 17.14 | 32.19 | 15.51 | 54.41 | 31.69 | 70.08 | 56.28 | 87.94 | 52.51 | 84.82 | 61.07 |
| 4 | 38.83 | 14.72 | 30.05 | 07.73 | 57.01 | 22.34 | 69.64 | 58.07 | 89.41 | 48.50 | 88.16 | 50.71 |
| 5 | **39.73** | 10.82 | 32.73 | 12.15 | 69.55 | 20.55 | **74.10** | 60.07 | 90.62 | 52.07 | 89.28 | 50.16 |
| 6 | 39.19 | 16.54 | **35.94** | 18.12 | 69.94 | 23.58 | 71.43 | 53.56 | 90.17 | 48.33 | **90.17** | 55.16 |
| 7 | 38.39 | 11.94 | 34.01 | 17.34 | **80.04** | 27.12 | 72.32 | 59.30 | 89.28 | 36.91 | 88.39 | 60.71 |
| 8 | 37.05 | 03.52 | 31.55 | 09.16 | 79.13 | 26.03 | 73.21 | 58.28 | 89.73 | 45.50 | 87.05 | 57.71 |
| 9 | 33.92 | 01.70 | 31.55 | 07.27 | 72.62 | 26.11 | 71.42 | 56.26 | **91.51** | 54.31 | 88.83 | 56.96 |
| 10 | 32.58 | 09.48 | 31.55 | 12.67 | 70.41 | 29.04 | 73.21 | 60.85 | 91.07 | 53.05 | 88.39 | 55.83 |
| 11 | 36.16 | 12.04 | 32.62 | 08.03 | 67.12 | 28.07 | 71.87 | 56.50 | 88.93 | 56.26 | 86.60 | 54.73 |
| 12 | 33.03 | 10.01 | 29.41 | 14.13 | 60.05 | 24.62 | 73.21 | 58.96 | 87.94 | 53.50 | 84.82 | 53.82 |

# Removal of linguistic properties significantly decreases alignment




Largest effect in mid layers



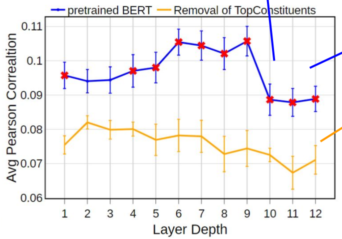

Red dots: significant difference

# Top constituents and word length contribute the most to the alignment trend across layers

Correlations across layers

| Tasks | Decoding Task with pretrained BERT vs Brain Recordings Residuals | Brain Recordings (BERT vs Residuals) |
|---|---|---|
| Word Length | 0.571 | 0.182 |
| TreeDepth | 0.299 | 0.491 |
| TopConstituents | 0.884 | 0.407 |
| Tense | 0.377 | 0.606 |
| Subject Number | 0.681 | 0.589 |
| Object Number | 0.449 | 0.359 |

| TopConst 20-classes (Syntactic) | |
|---|---|
| before | after |
| 42.13 | 20.71 |
| 52.05 | 30.35 |
| 54.41 | 31.69 |
| 57.01 | 22.34 |
| 69.55 | 20.55 |
| 69.94 | 23.58 |
| **80.04** | 27.12 |
| 79.13 | 26.03 |
| 72.62 | 26.11 |
| 70.41 | 29.04 |
| 67.12 | 28.07 |
| 60.05 | 24.62 |




- If col 2 is high -> ling. prop. less important for trend
- If col 1 is high & col 2 is low -> ling. prop. important for trend

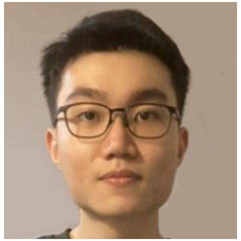
# Today: evidence from 3 perturbation case studies

1. Alignment due to more than next-word prediction & word-level semantics
   [Merlin & Toneva, 2022 arXiv https://arxiv.org/abs/2212.00596]
2. Joint processing of linguistic properties
   [Oota, Gupta, and Toneva 2022 arXiv https://arxiv.org/abs/2212.08094]
3. **Training to summarize narratives improves brain alignment**
   [Aw & Toneva, ICLR 2023 https://arxiv.org/abs/2212.10898]

# Case study 3

- To achieve deeper understanding of language, recent works train language models to summarize narrative datasets [Kryscinski et al. 2021, Sang et al. 2022]
- Are these models truly learning deeper understanding of language?

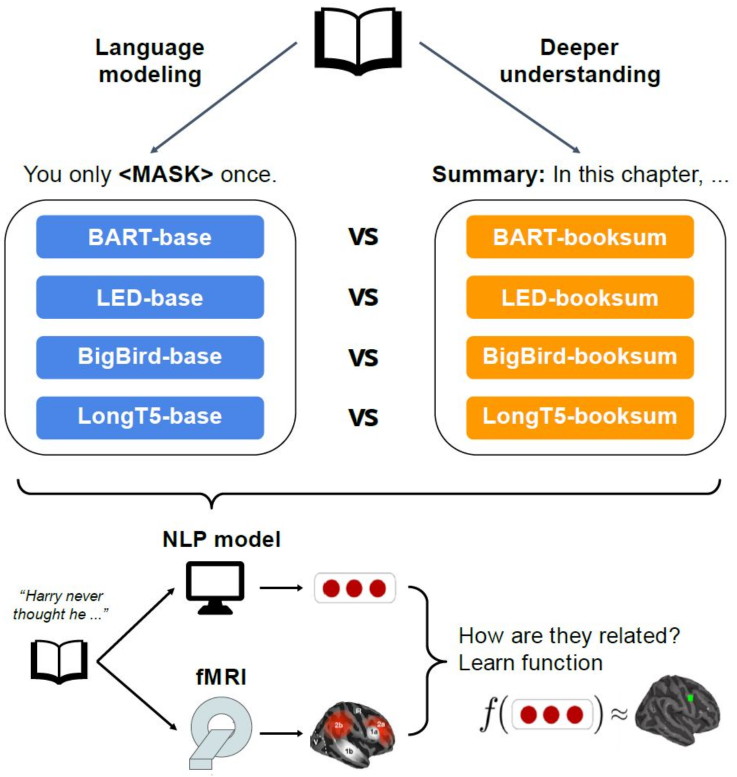
Investigate with one system that truly understands language: the human brain



Khai Loong Aw

Training language models for deeper understanding improves brain alignment, Aw and Toneva ICLR 2023 [https://arxiv.org/abs/2212.10898](https://arxiv.org/abs/2212.10898)

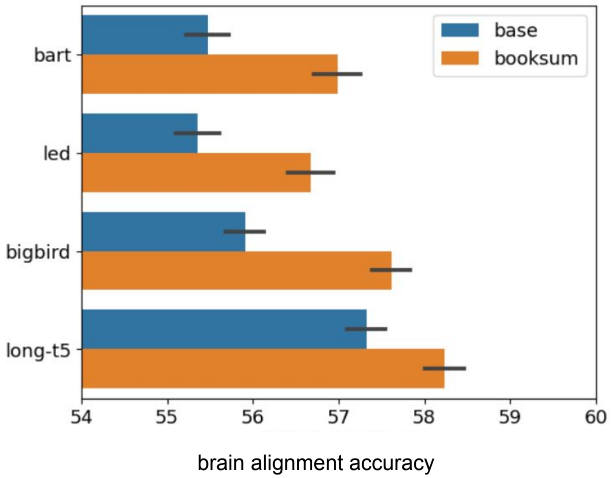
# Perturbation: training to summarize narratives

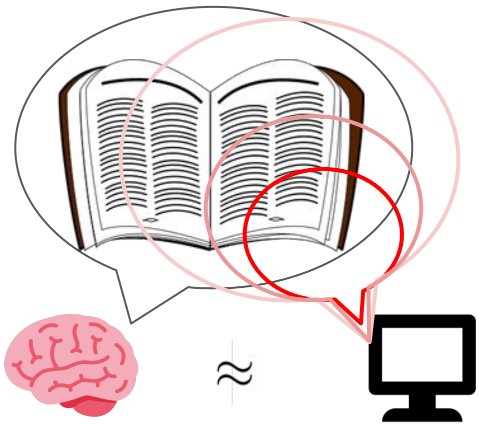



Base: pretrained models

Booksum: fine-tuned on BookSum dataset (summarization of narrative chapters)

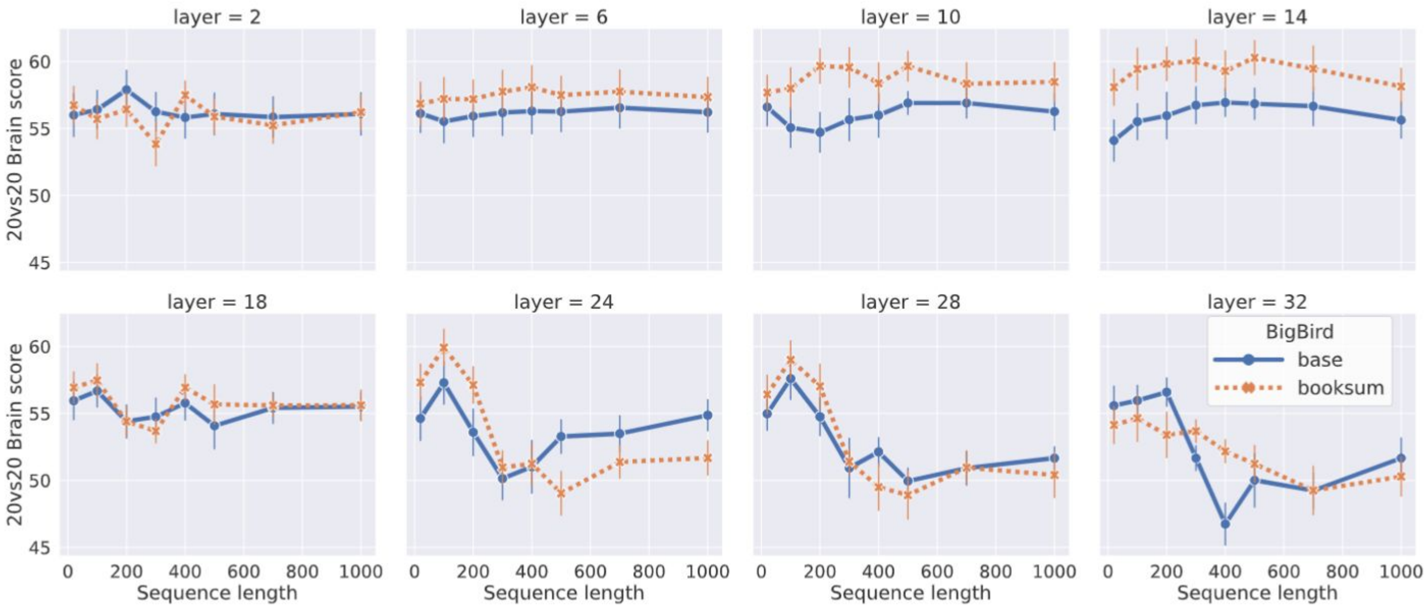
# Models trained to summarize narratives align better with brain recordings



Vary context provided to NLP model and observe how alignment with fMRI recordings changes

# 10-fold increase in context length that results in the peak of brain alignment



# Why do models that learn to summarize narratives align better with brain recordings?

- Not because of next-word prediction
  - BookSum << pretrained at LM
- Not (entirely) because of greater similarly of text domain to brain dataset
  - BookSum >= LM-BookSum at brain alignment
- Partially because of summarization
  - CNNSum >= pretrained at brain alignment, but BookSum >> CNNSum
- More brain-aligned representation of important discourse elements
  - BookSum >> pretrained for sentences with Characters, more than other sentences

# Today: evidence from 3 perturbation case studies

1. Alignment due to more than next-word prediction & word-level semantics
   [Merlin & Toneva, 2022 arXiv https://arxiv.org/abs/2212.00596]
2. Joint processing of linguistic properties
   [Oota, Gupta, and Toneva 2022 arXiv https://arxiv.org/abs/2212.08094]
3. Training to summarize narratives improves brain alignment [Aw & Toneva, ICLR 2023 https://arxiv.org/abs/2212.10898]

# Supplementary Slides

# Linguistic Property Similarity