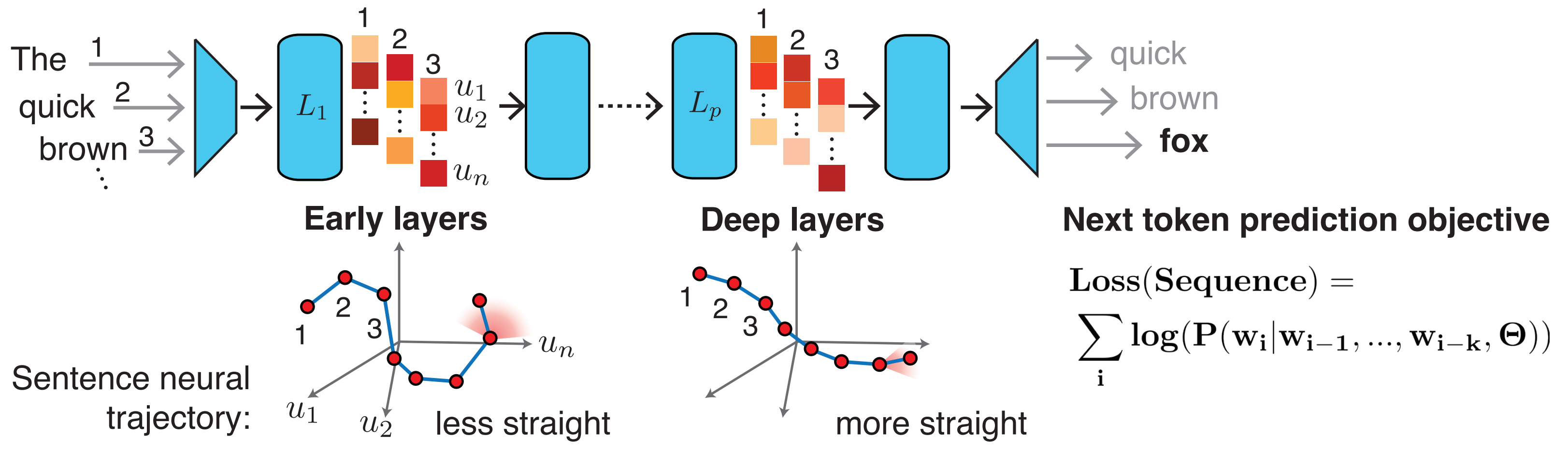
Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language

Eghbal A. Hosseini, Evelina Fedorenko
Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, MIT, Cambridge, MA
@eghbal_hosseini
ehoseini@mit.edu

Introduction

- Predicting upcoming events is critical for our ability to effectively interact with our environment and conspecifics.
- Transformer models in NLP, trained on next token prediction, develop versatile representations for various tasks
- However, we still lack an understanding of how a predictive objective shapes such representations.
- Drawing on vision neuroscience research (Henaff, 2019 [1]), we evaluate if **autoregressive transformer** [2] models' representations become more straight across network layers.



$$\text{Loss(Sequence)} = \sum_i \log(P(w_i|w_{i-1}, ..., w_{i-k}, \Theta))$$

Methods

Sentence corpus: comprised 8,408 Universal Dependencies sentences [2], 6-19 tokens long, spanning diverse topics and including the top 100K English words.

Sentence Curvature: involved translating words into transition vectors (1), calculating angles between them (2), and averaging angles for sentence curvature (3).

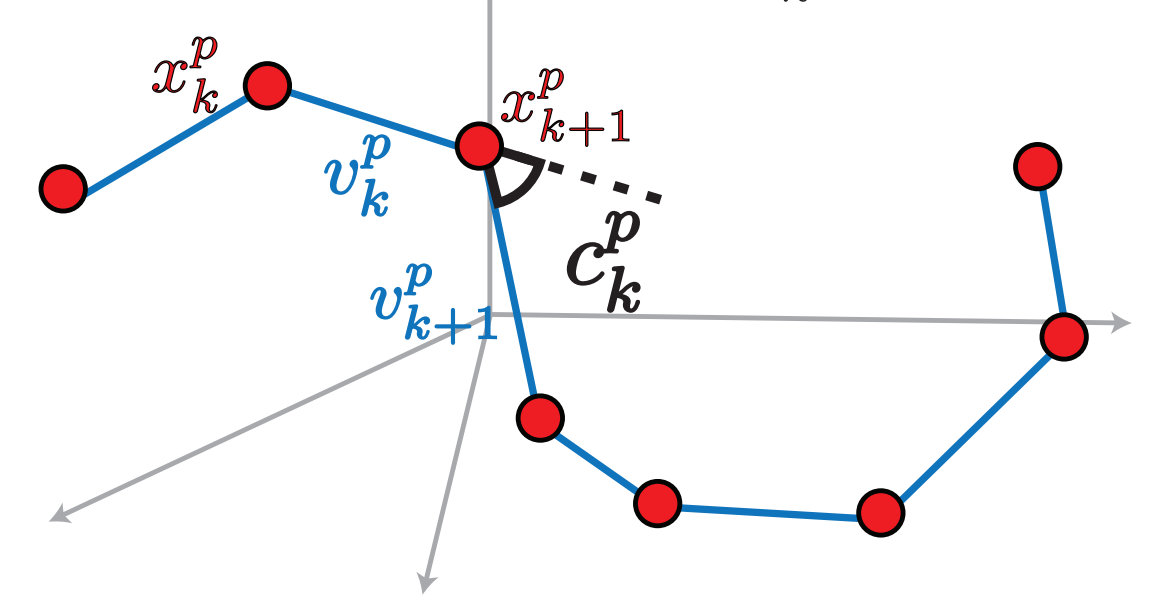
Layer Curvature: Calculated curvature change from the first layer (4), averaging these changes across sentences (5)

Transition vector

(1) $v_k^p = x_{k+1}^p - x_k^p$

Transition curvature

(2) $c_k^p = arccos\left(\frac{v_{k+1}^p . v_k^p}{||v_{k+1}^p||\ ||v_k^p||}\right)$



Sentence curvature

(3) $C_{s_n}^p = \frac{1}{k}\sum_{i=1}^{k} c_i^p$

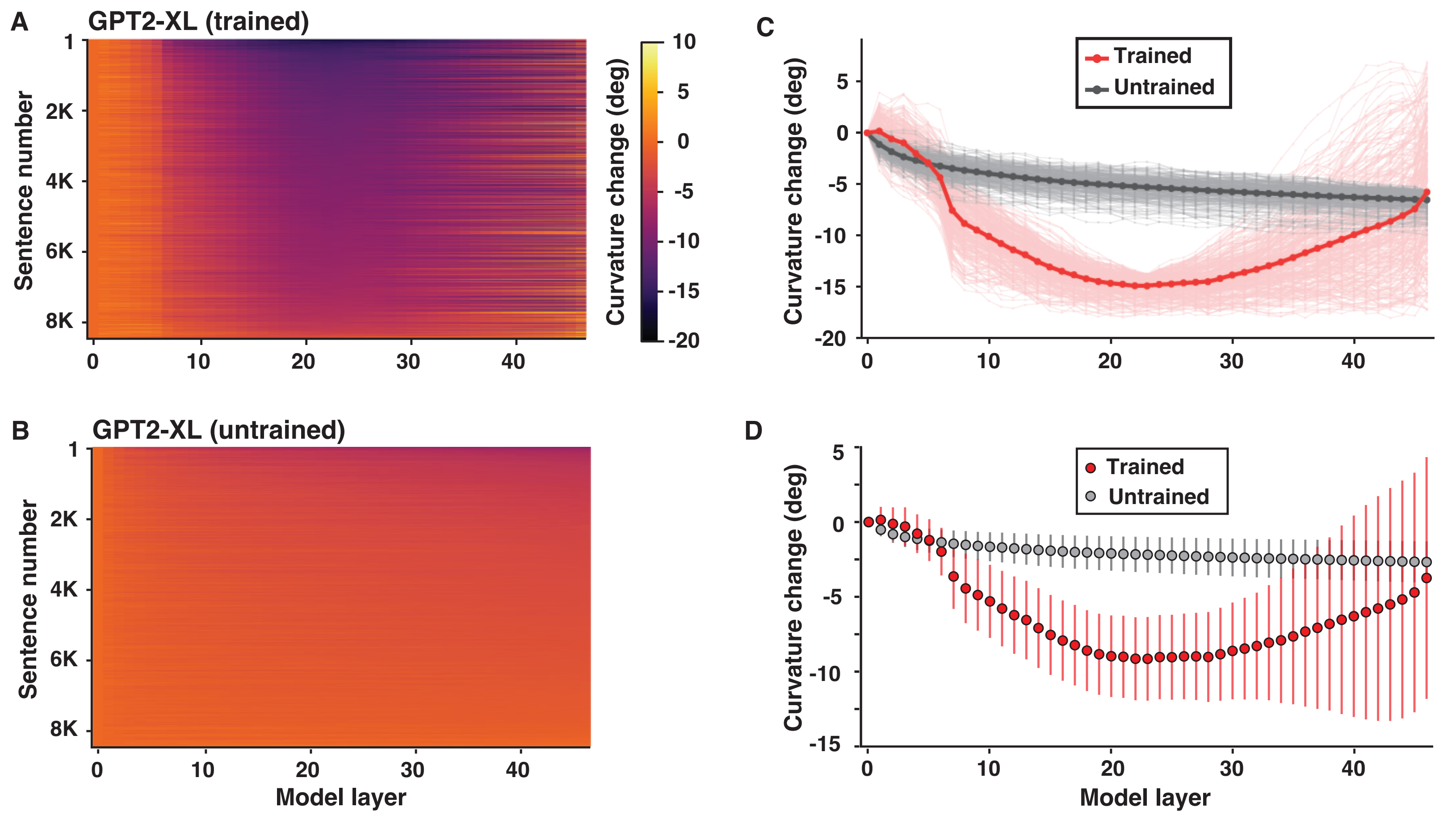
Sentence curvature change

(4) $\Delta C_{s_n}^P = C_{s_n}^P - C_{s_n}^1$

Layer curvature change

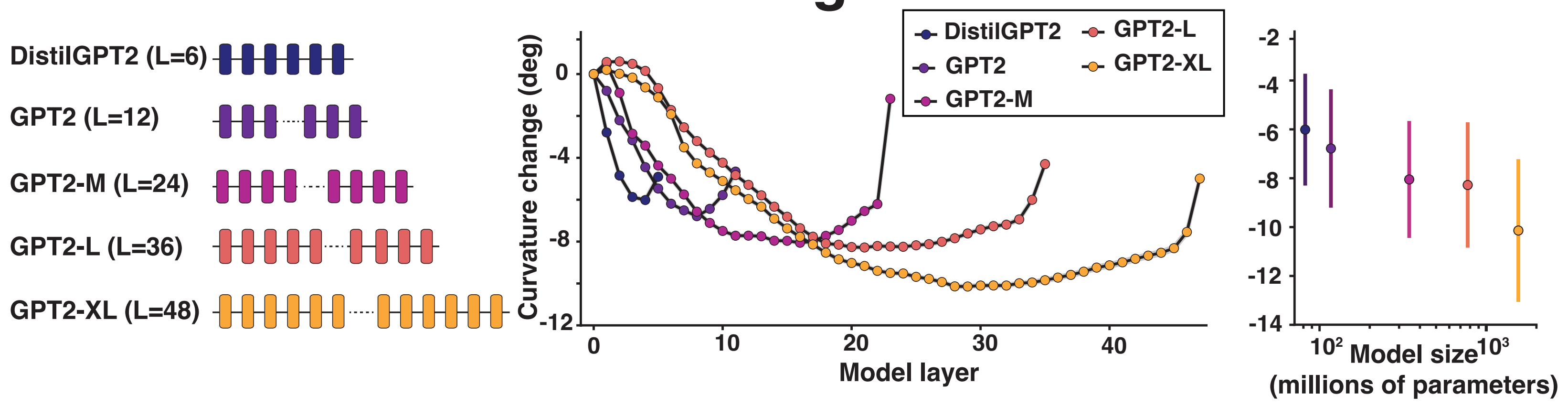
(5) $\Delta C^P = \frac{1}{N}\sum_{j=1}^{N} \Delta C_{s_n}^P$

Curvature decreases across the model layers

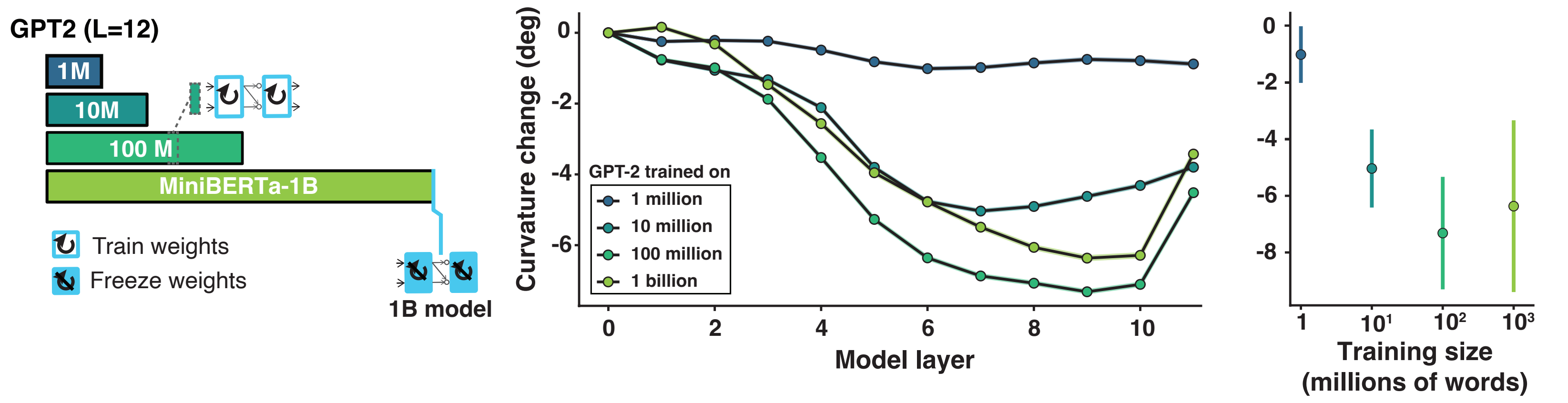


- **(A)** Trained model shows consistent curvature reduction from the first layer. **(B)** Minimal changes in untrained model. **(C)** Trained models display gradual curvature decrease layer-wise. **(D)** Average drop of about 8 degrees from the first layer

Curvature decreases to a greater extent in larger models and with more training

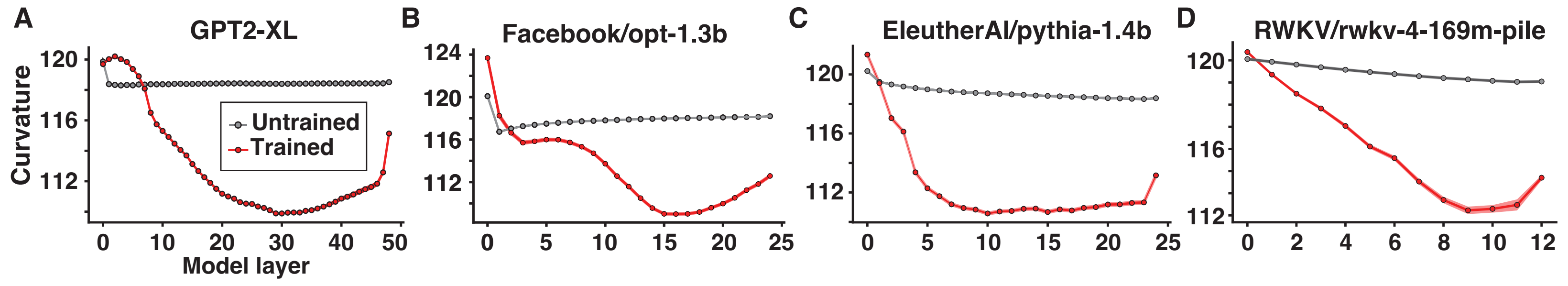


- All models consistently reduce curvature from early to deeper layers, with larger models (lower perplexity) showing greater reduction.

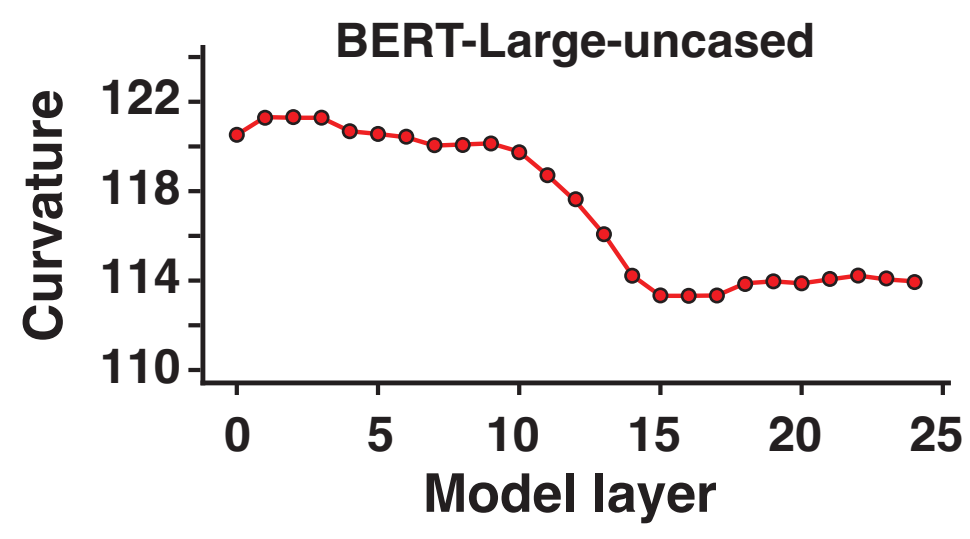


- Excluding the smallest dataset model (1M tokens), models trained on larger datasets increasingly reduce curvature

Straightening generalizes across autoregressive models

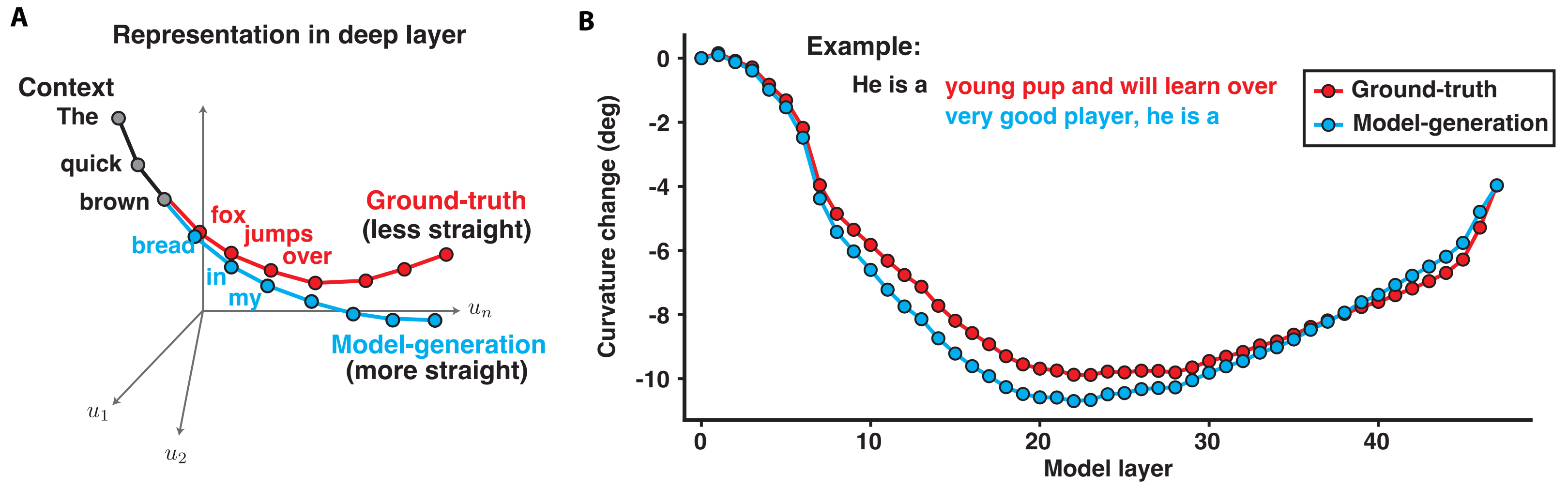


- Similar to the GPT2-XL model **(A)**, models employing different training **(B)**, input positional encoding **(C)**, and RNN-based attention **(D)** significantly reduces layer curvature, only after training.



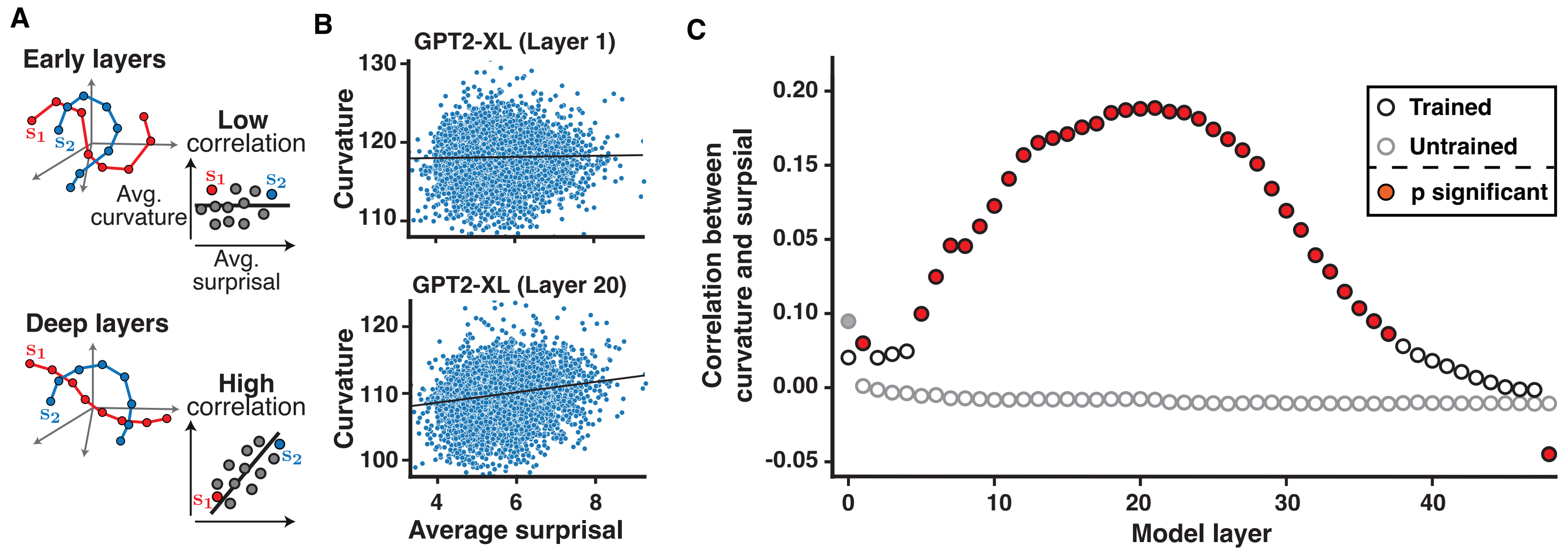
- Ulike autoregressive models trained on next-word prediction, bidirectional models, which trained on masked language modeling, do not exhibit the same straightening pattern across their layers.

Models favor straight trajectories during language generation



- **(A)**The hypothesis anticipates straighter model-generated sentence trajectories. Ground-truth sentences (blue) from UD corpus contrast with model-generated ones (red), derived from a three-token prompt and extended using greedy search. **(B)** Model-generated sentences show greater curvature reduction, aligning with the hypothesis.

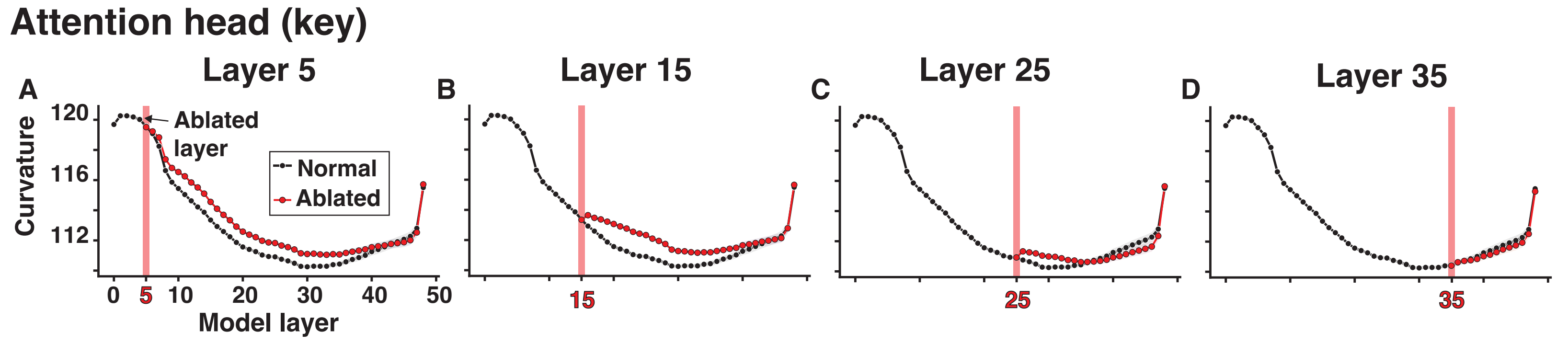
Model curvature is correlated with sentence surprisal



- **Surprisal:** Assessed word unexpectedness to link curvature with human behavior, averaging 3-word surprisal values per sentence[4,5].
- **(A)** Hypothesizes stronger correlation between curvature and surprisal in network's middle than early layers. **(B)** Middle layers show notably higher correlation than early layers, with **(C)** trained models exhibiting increasing correlation from early to middle layers, unlike untrained models.

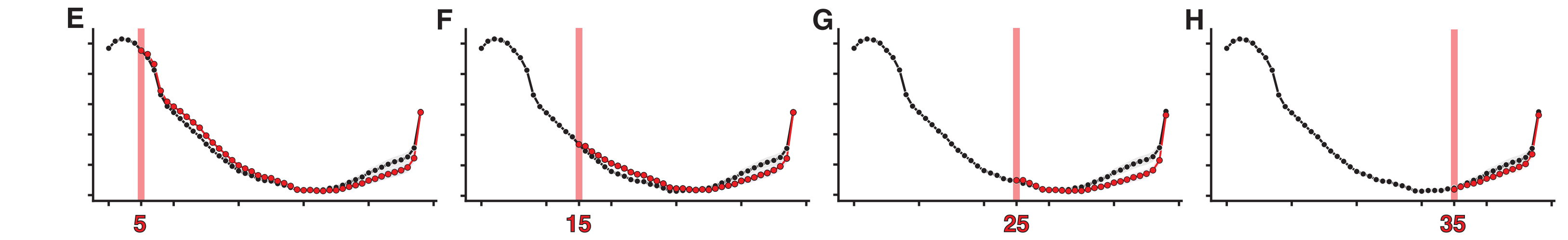
Model ablation reveals the role of self-attention in straightening

Attention head (key)



- The attention head (key) weight was replaced with an identity transform in just one layer, between layer 5 **(A)** to and 35 **(D)**. Ablating early layers resulted in a deficit in straightening.

Feed-Forward Network



- Both weights in the feed-forward modules were replaced such that the resulting operation is an identity transform. Unlike attention-head ablation, this change does not affect the models' ability to straighten.

Conclusions

- We propose neural trajectory straightening as a fundamental hypothesis for the mechanism underlying transformers models' ability to predict the next word
- Our findings fit within the efficient coding framework, suggesting that temporal prediction emerges as an efficient solution to the predictive objective in transformer networks.
- Exploring the universality of this mechanism in other predictive systems, both biological and artificial, is an area for future research.

References

[1] Hénaff, O. J., Goris, R. L. T., and Simoncelli, E. P. (2019). Perceptual straightening of natural videos. Nat. Neurosci.

[2] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

[3] de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. Comput. Linguist. Assoc. Comput. Linguist.

[4] Levy, R. (2008). Expectation-based syntactic comprehension. Cognition

[5] Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. PNAS

Link to paper