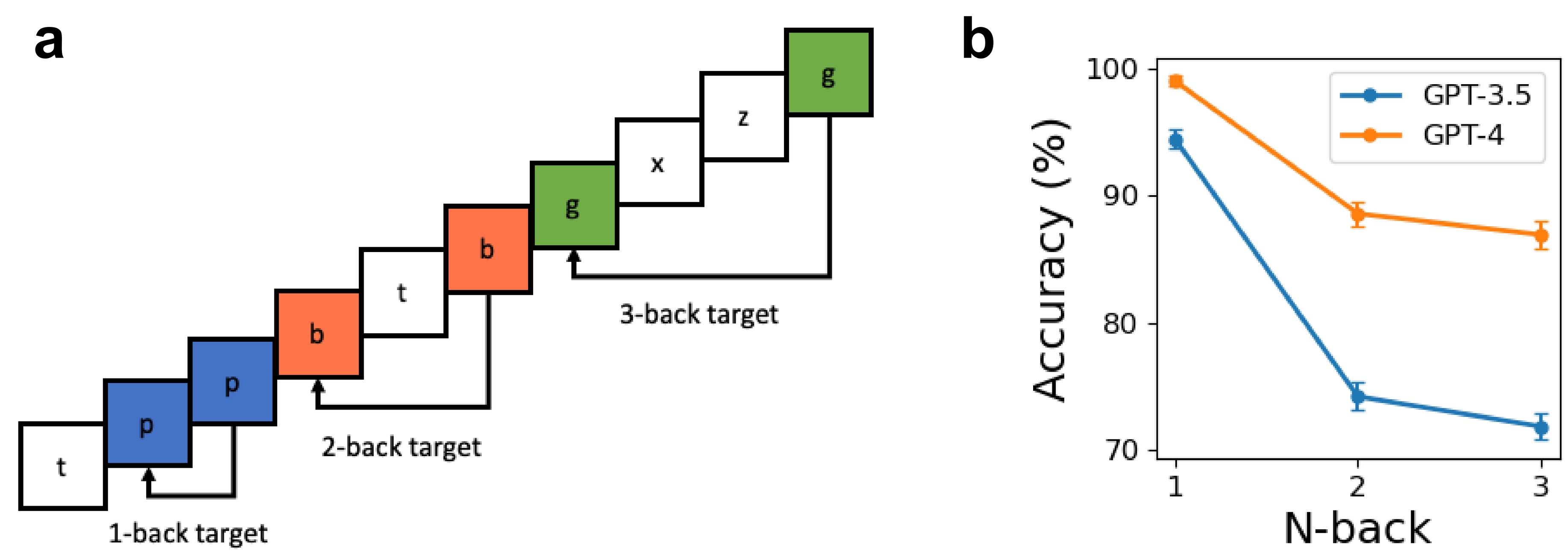
Self-Attention Limits Working Memory Capacity of Transformer-Based Models

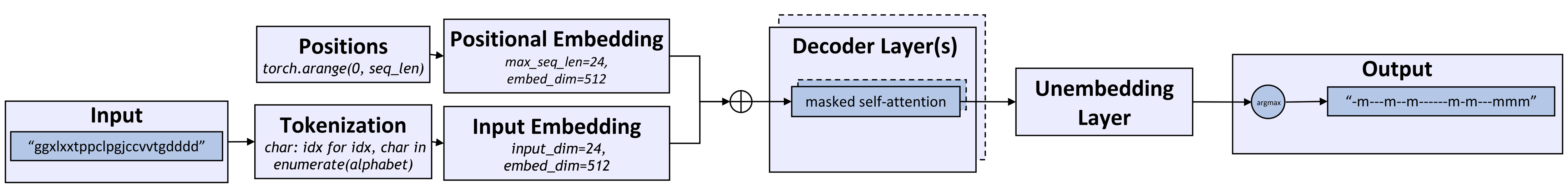
Dongyu Gong, Yale University (dongyu.gong@yale.edu); Hantao Zhang, Yale University

Background

Transformer-based large language models (LLMs) has striking limits in their working memory capacity, as measured by N-back tasks in cognitive science [1]. However, there is still a lack of mechanistic interpretability as to why this phenomenon would arise.

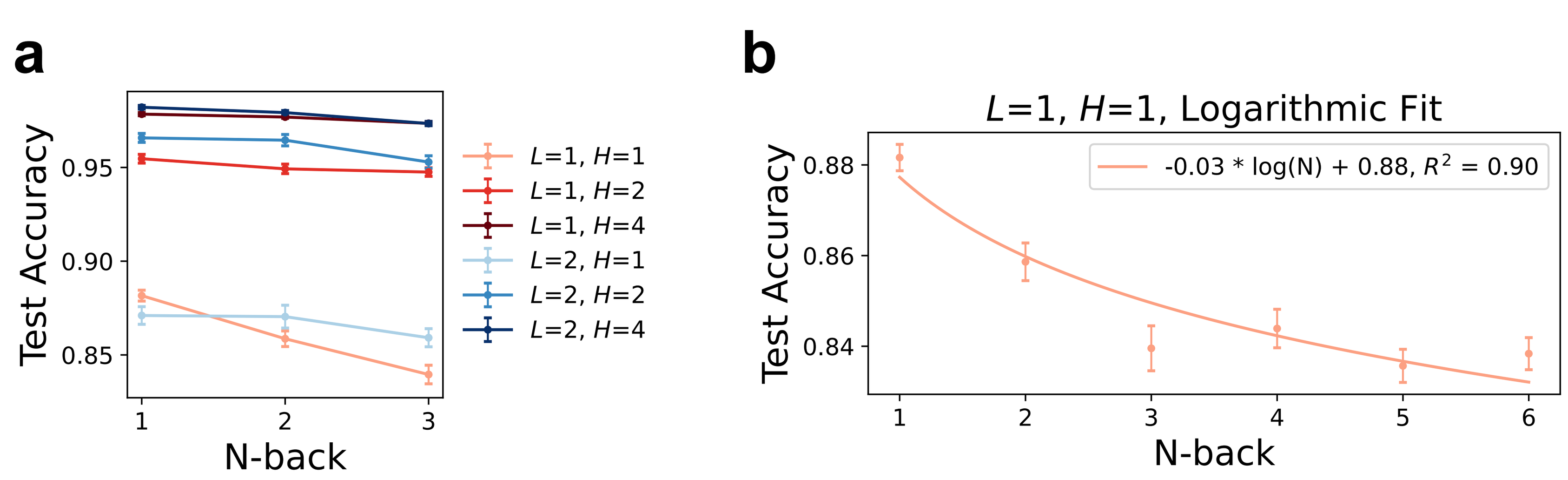


Methods

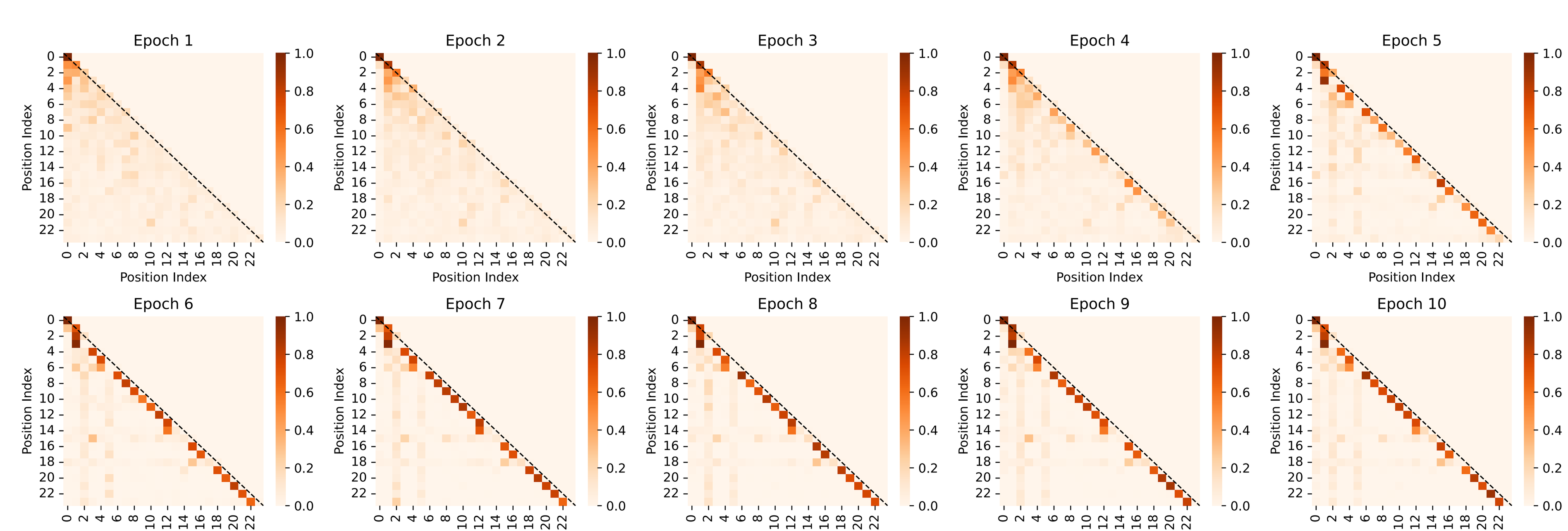


Inspired by the **executive attention theory** in cognitive science, we hypothesize that the self-attention mechanism within Transformer-based models might be responsible for their working memory capacity limits. To test this hypothesis, we train vanilla decoder-only transformers to perform N-back tasks. We mainly focus our analysis on a causal Transformer containing one decoder layer with only one attention, although we also test a few architectural variants in the number of decoder layers (L) and number of attention heads per layer (H) for comparisons.

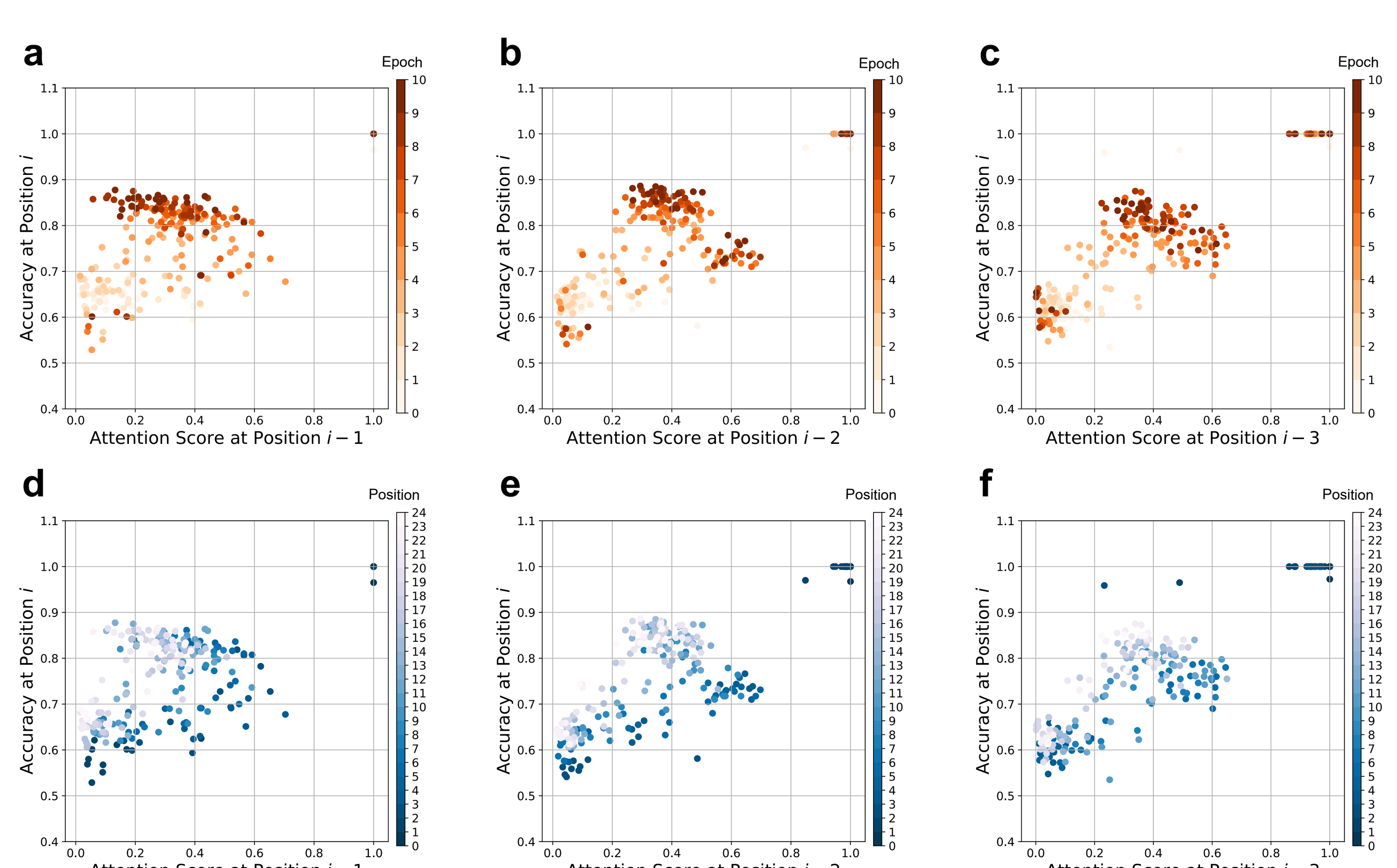
Results



1. Model accuracy decreases as N increases. We find a significant decline in model performance as N increases for the 1-layer 1-head model. To further confirm this pattern, we extend the task to N = 6 and find a significant logarithmic decline in the test accuracy as N increases.



2. Attention scores during training reflect the trajectory of learning. Starting with almost uniformly distributed attention scores in each row, attention scores gradually aggregate to a line corresponding to the N-back positions.



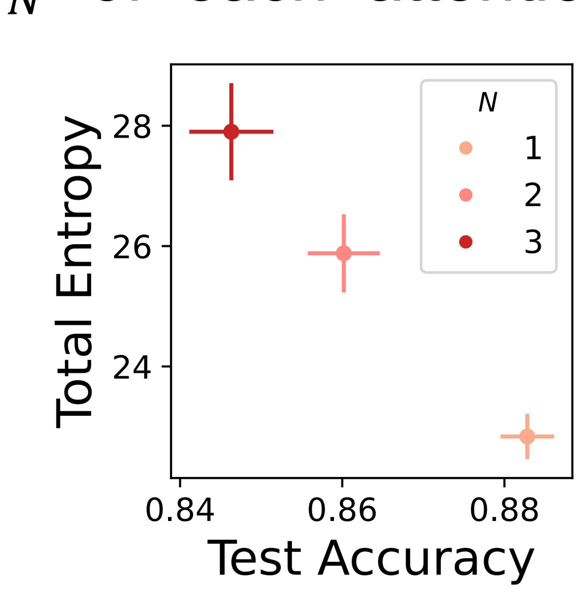
3. Attention score at position *i* − N increases with test accuracy at position *i*. Over training epochs, the attention score at position i − N increases along with the accuracy at position i (panel **a**-**c**). When using the same data but assigning colors to the dots according to which position each dot belongs to (panel **d**-**f**), there is a clear pattern that attention scores get dispersed at later locations.

4. Total entropy of attention scores increases as N increases. We define the total entropy H_N of each attention score matrix $A \in \mathbb{R}^{24\times 24}$ as

$$H_N(A) = -\sum_{i=1}^{24}\sum_{j=1}^{i} A_{i,j}\log(A_{i,j})$$

where

$$A_{i,j} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)_{i,j}$$



We find that H_N increases as N increases, leading to the decrease in test accuracy.

Discussion

Our findings suggest a shared role of attention in the working memory capacity of humans and LLMs. The mechanistic interpretability of working memory capacity limits in Transformer-based models could inform future efforts to design more powerful model architectures with enhanced cognitive capabilities [2].

References

[1] Dongyu Gong, Xingchen Wan, and Dingmin Wang. Working memory capacity of ChatGPT: An empirical study. *AAAI 2024*.

[2] Graeme S Halford, Nelson Cowan, and Glenda Andrews. Separating Cognitive Capacity from Knowledge: A New Hypothesis. *Trends in Cognitive Sciences 2007*.