StructLens: A Structural Lens for Language Models via Maximum Spanning Trees

Haruki Sakajo, Frederikus Hudi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST)

How are token representations organized?

- Logit Lens: examines token-level prediction
- Sparse Autoencoders: decompose representations into sparse features
- Structural Probe: recovers linguistic structures through linear transformations

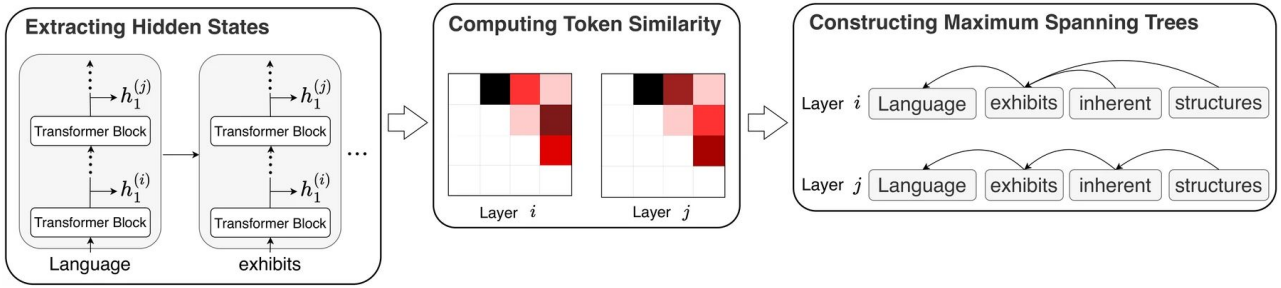
How are token representations organized?

- Logit Lens: examines **token-level** prediction
- Sparse Autoencoders: decompose representations into sparse features through **dictionary learning**
- Structural Probe: recovers linguistic structures through **linear transformations**

How do LLMs **naturally** organize tokens?

An unsupervised, training-free method is needed!

Method: StructLens



- Extract hidden states
- Compute similarity
- Build Maximum Spanning Trees

$$\boldsymbol{A}_{i,j} = \begin{cases} \frac{1}{1+\|\boldsymbol{h}_i - \boldsymbol{h}_j\|} & \text{if } i < j \\ 0 & \text{otherwise} \end{cases}$$

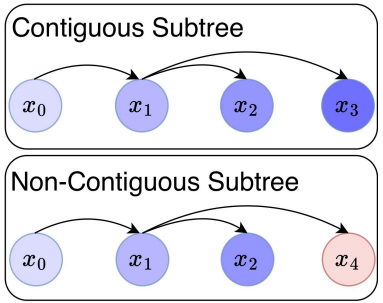
Characteristics of the Trees

Investigating subtrees (fragments)

Are connected fragments also contiguous token spans?

- Subtree Ratio
  - Fraction of k-node connected subtrees whose token indices form an interval
- Token Ratio
  - Fraction of tokens covered by at least one contiguous k-node subtree

We use k=3, 4, 5.

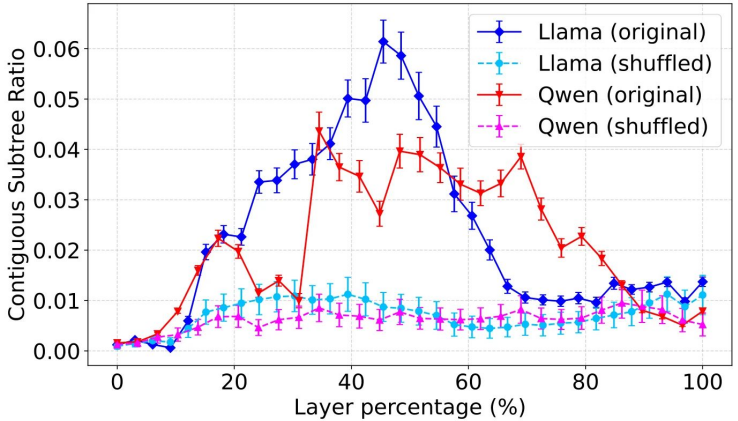


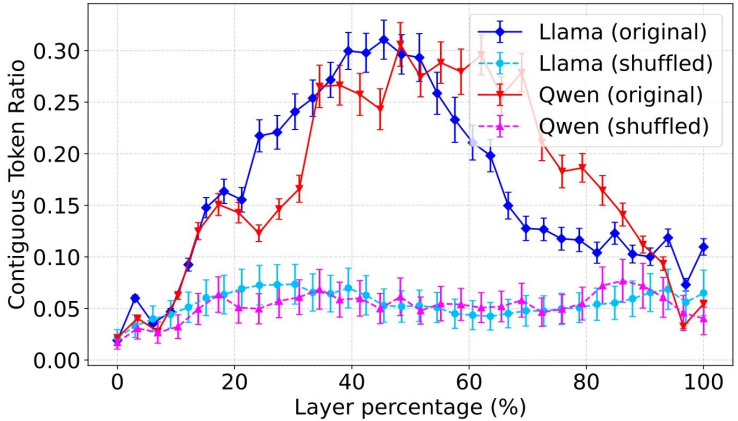
Experimental Settings

- Model
  - Llama 3.1 8B, Qwen 2.5 7B
  - OLMo 2 7B and intermediate checkpoints
    - Stage1-{1k, 10k, 101k, 500k, 750k, 928k}, Stage2-1k, Final
- Dataset
  - MMLU (QA task, 5-shot), Mutlinews (Summarization task)

Middle layers form local-span structures

Across Llama 3.1 8B and Qwen 2.5 7B, contiguous subtrees/tokens rise from lower to middle layers, then decline in upper layers.

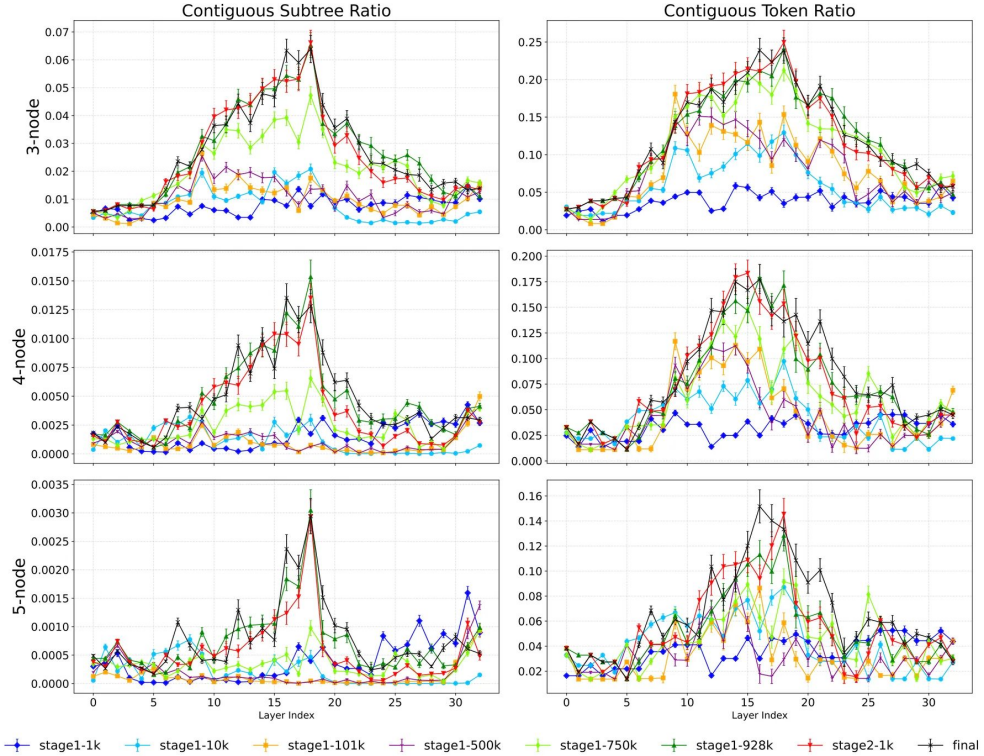




MMLU, k=4

Middle-layer structure emerges during training

- Smaller subtrees become detectable earlier
- Larger contiguous subtrees emerge later
- The emergence happens suddenly (at step 750k)



Layer pruning using structural signals

Measuring layer redundancy using structural similarity instead of cosine similarity

- We measure structural similarity using 3 methods
  - Details omitted
- We remove approximately 10% of layers

Structural signals can help layer pruning but not universally

Structural similarities are helpful for Llama, but not for Qwen

Model	Metric	MMLU Acc.
Llama 3.1 8B	Dense (Upper bound)	68.4
	Cosine Sim. (Baseline)	66.8
	CKA (Baseline)	66.9
	Structural Sim.	66.9 - **67.8**
Qwen 2.5 7B	Dense (Upper bound)	74.3
	Cosine Sim. (Baseline)	56.1
	CKA (Baseline)	**70.0**
	Structural Sim.	53.4 - 59.7

Key Takeaways

- We introduce StructLens, a framework to analyze token relationships in representation space
- StructLens reveals middle layer local-span structures
- StructLens also finds the evolution of token organization during pre-training
- Structural similarity is potentially useful for layer pruning