Internal Planning in Language Models: Characterizing Horizon and Branch Awareness

Muhammed Ustaomeroglu*, Baris Askin*, Gauri Joshi, Carlee Joe-Wong, Guannan Qu

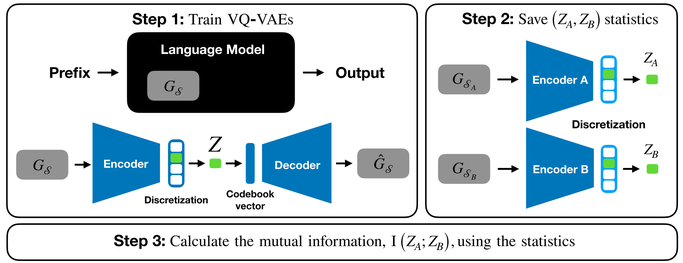
Carnegie Mellon University

Motivation

- Even simple next-token predictors show signs of **planning ahead**
- Focus on the model's internal **planning horizon** and **branch awareness**
- **Hidden states** encode rich planning-related information

Method

- Mutual Information as a signal of planning
- Discretize hidden states and measure mutual information between them
- Interpretation by comparing mutual information patterns



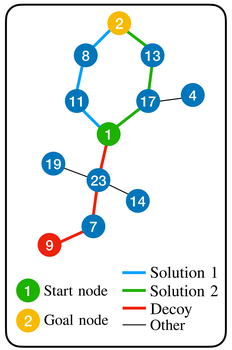
LM Tasks

CFG: controlled synthetic dataset for learning low-level textual/syntactic structure, similar to local relations in natural text.

Path-finding: graph-based dataset for multi-step relational reasoning, where the model must infer a valid path from a start node to a goal node.

A sample prompt: "1923,132,118,417,123,97,28,171,132,14 23 , 23 7 , 1 23 :"
Correct responses: "1 11 8 2" or "1 17 13 2"

NLP (OpenWebText): real-world natural text for next-token prediction.



Do LMs internally plan their outputs, even without special prompting or fine-tuning?

How far ahead do they plan, and how much branching do they consider?

MI-based pipeline to analyze:

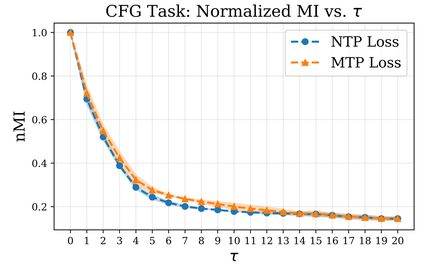
- planning horizon
- awareness of alternative answers

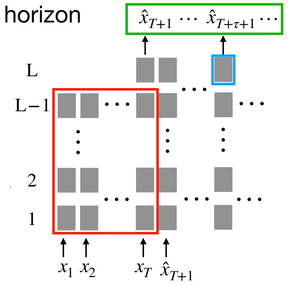
Takeaways

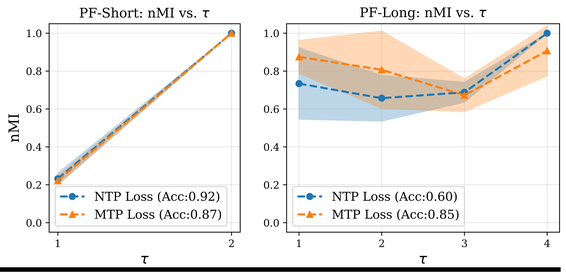
- LMs exhibit internal planning involving both future horizon and alternative correct answers
- Planning depth varies by task
- Hidden states are aware of alternative answers
- Generations depend primarily on later layers and recent context

How much do LMs plan about future tokens before deciding on the immediate next token?

- Mutual information (MI) between pre-output computations and future token decisions
- Sharply decreasing MI indicates a short planning horizon
- Slowly decreasing or stable MI indicates a longer planning horizon







Do models consider alternative correct answers?

- Mutual information (MI) between pre-output computations and possible correct and incorrect answers
- Stronger relation to alternative correct answers than to incorrect answers indicates branch awareness

Model	PF-Short		PF-Long	
	$\frac{\mathcal{I}(Z_H;Z_{\text{alt}})}{\mathcal{I}(Z_H;Z_{\text{decoy}})}$	Accuracy	$\frac{\mathcal{I}(Z_H;Z_{\text{alt}})}{\mathcal{I}(Z_H;Z_{\text{decoy}})}$	Accuracy
$\mathcal{M}_{\text{NTP}}$	7.60 ± 0.78	0.92 ± 0.03	1.45 ± 0.01	0.60 ± 0.01
$\mathcal{M}_{\text{MTP}}$	6.29 ± 0.17	0.88 ± 0.05	1.82 ± 0.27	0.85 ± 0.02

To what extent do LMs rely on earlier computations when generating new tokens?