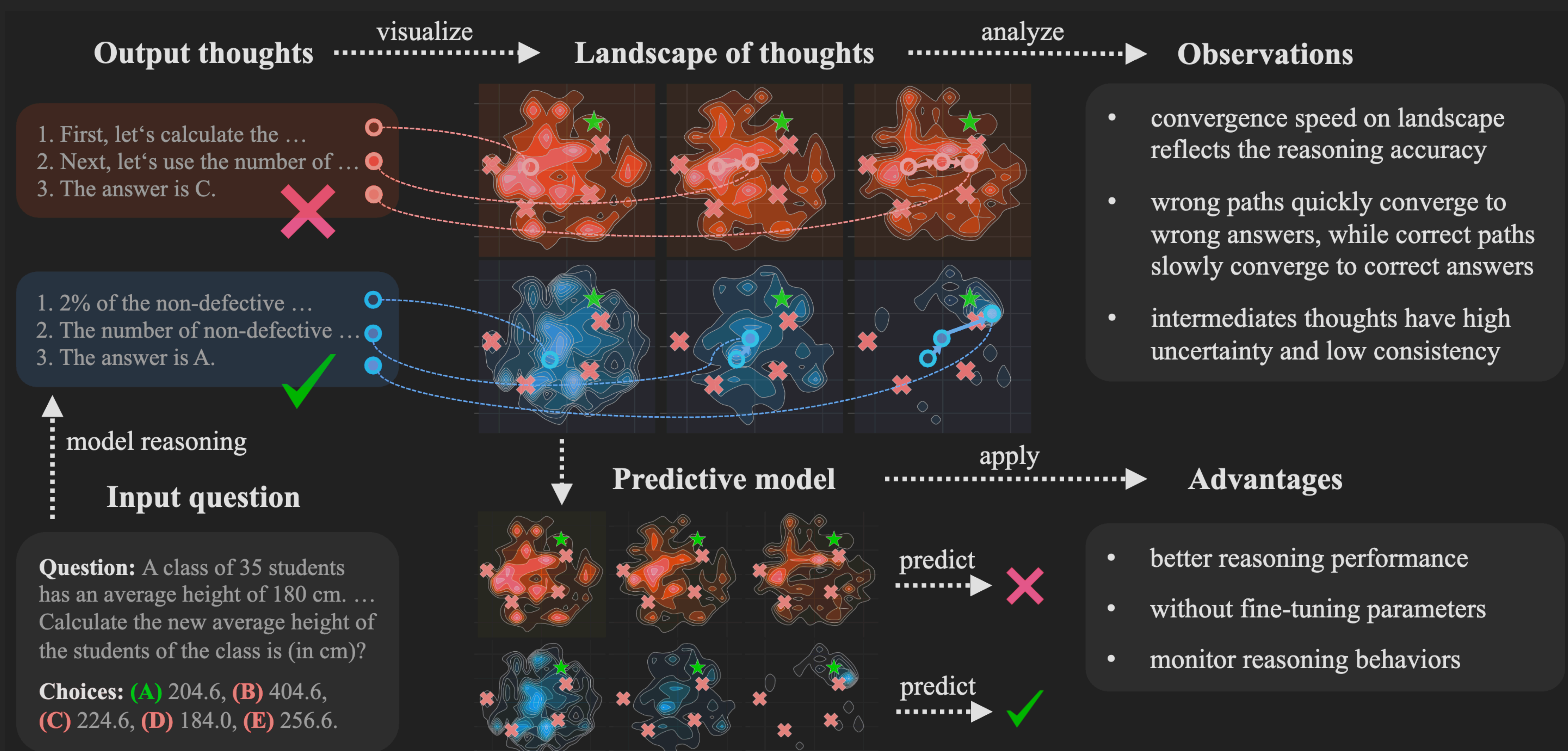
香港浸會大學 HONG KONG BAPTIST UNIVERSITY · Stanford University · Mila · Université de Montréal · HEC MONTRÉAL · intel

Landscape of Thoughts: Visualizing the Reasoning Process of Large Language Models

Zhanke Zhou*, Zhaocheng Zhu*, Xuan Li*, Mikhail Galkin, Xiao Feng, Sanmi Koyejo, Jian Tang, Bo Han

Motivation: the reasoning behavior of LLMs remains poorly understood

- **Reading texts (reasoning outputs)** is tedious and time-consuming ❌
- **Analysis with visualization plots** is more easy and intuitive ✅

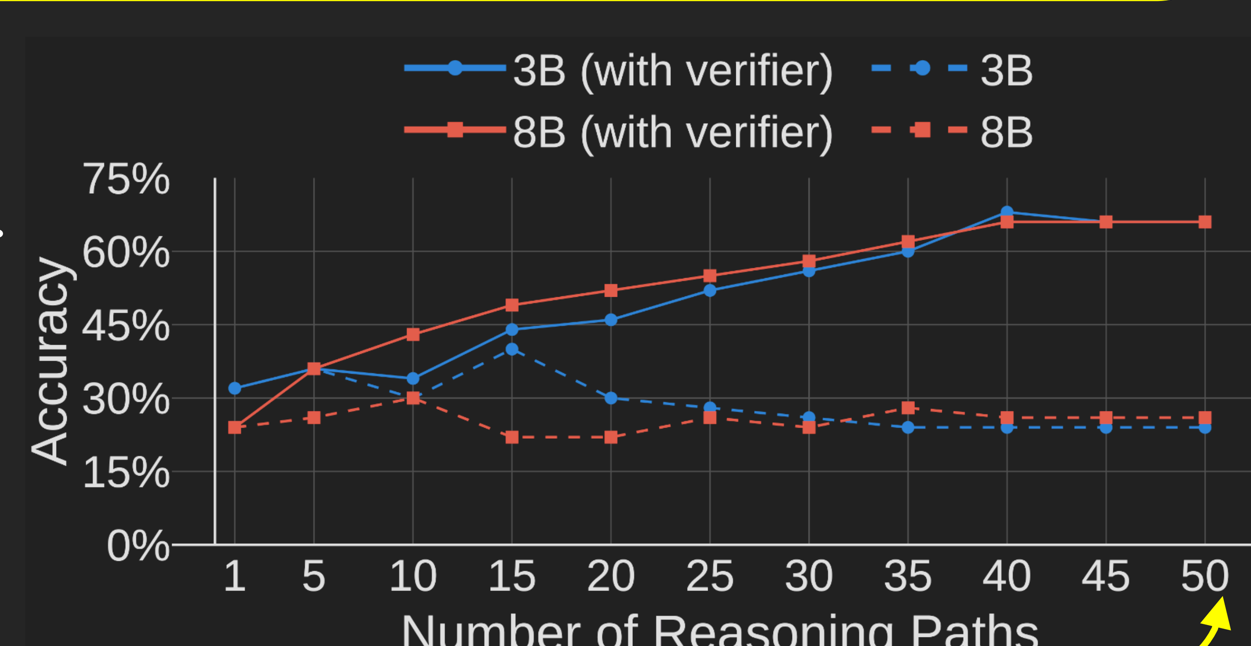


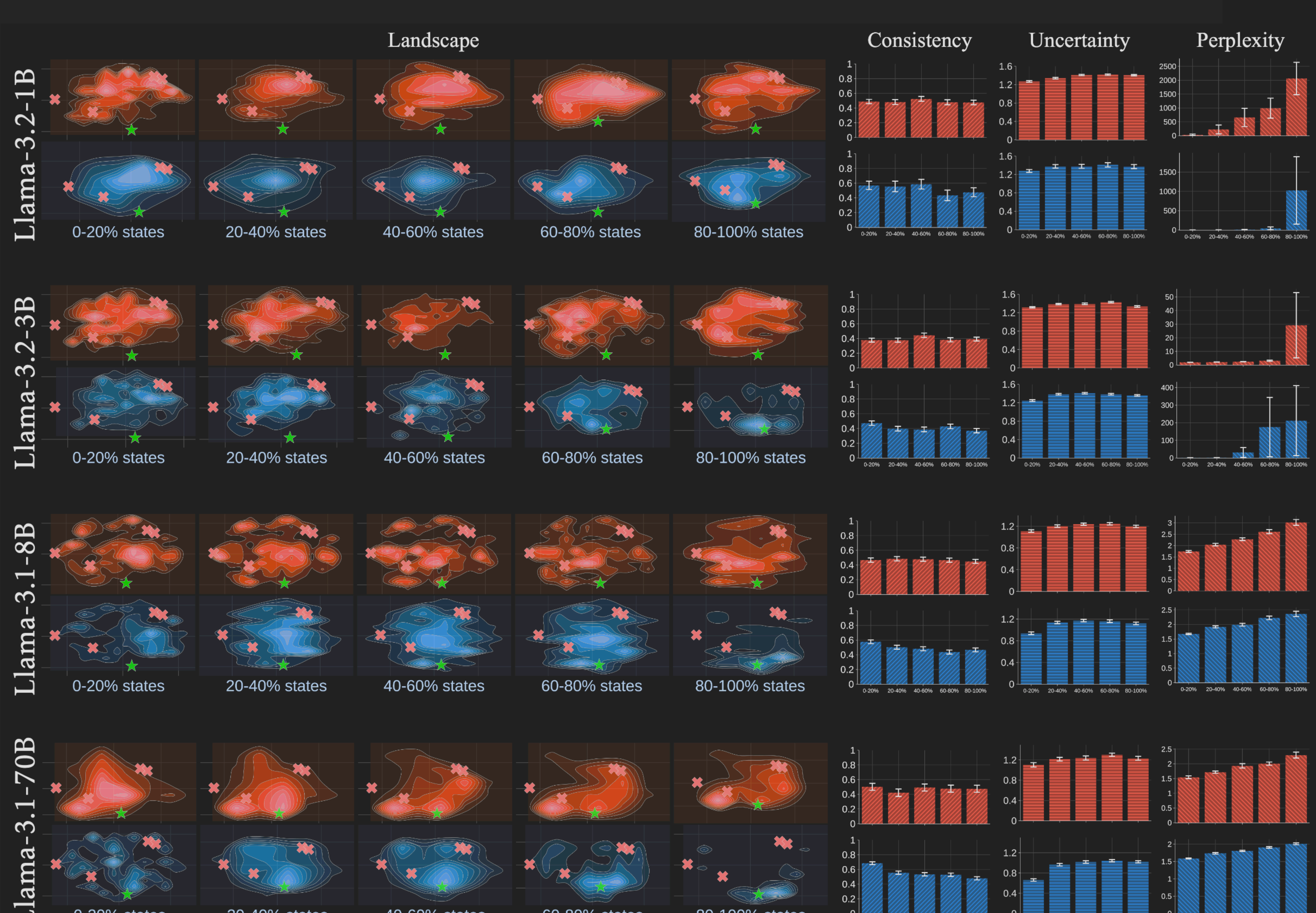
Key: Project each state from texts to numerical feature s_i (distances to the k choices of this question)

$s_i = [d(s_i, c_1), d(s_i, c_2), \ldots, d(s_i, c_k)]^\top$, where $d(s_i, c_j) = p_{\text{LLM}}(c_j|s_i)^{-\frac{1}{|c_j|}}$ (the perplexity of decoding choice c_j given state s_i) then, we obtain the feature matrix including all states and choices, and project it to 2-dimensional space via t-SNE

Observations from the landscape

1. Faster convergence to the correct answers is tied to higher reasoning accuracy.
2. Wrong paths quickly converge to wrong answers, while correct paths slowly converge to correct answers.
3. The landscape converges faster as the model size increases. Larger models have higher consistency, lower uncertainty, and lower perplexity.





Application of the landscape

Build up a lightweight verifier with the feature matrix of landscape

Consistency: whether the model knows the answer in the middle

$$\text{Consistency}(s_i) = \mathbb{I}(\arg\min s_i = \arg\min s_n)$$

Uncertainty: how confident the model is about its prediction (the information entropy)

$$\text{Uncertainty}(s_i) = -\sum_{d \in s_i} d \log d$$

Perplexity: how confident the model is about its generated thoughts

$$\text{Perplexity}(t_i) = p_{\text{LLM}}(t_i|s_{i-1})^{-\frac{1}{|t_i|}}$$

paper · code · slides