A Model of Errors in Transformers

Suvrat Raju and Praneeth Netrapalli

International Conference on Machine Learning

LLM Evaluation Example

Prompt

What is 9627541239 + 2928938366?

LLM Answer

12556479065

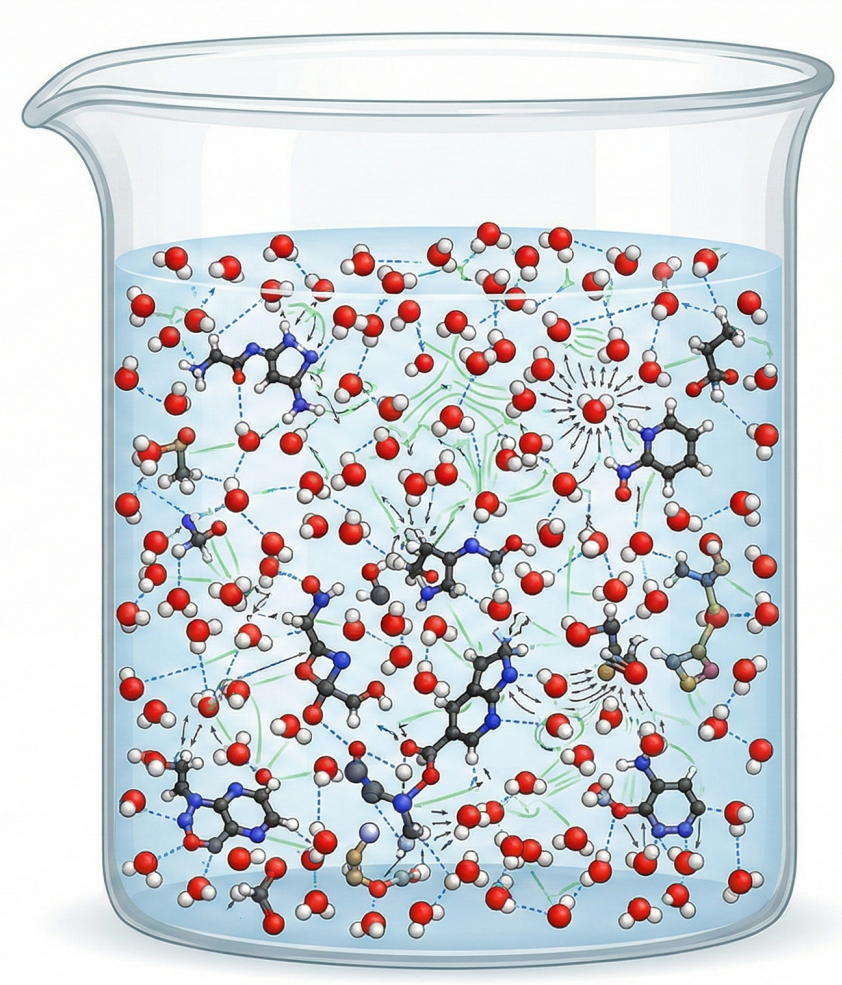
Correct Answer

12556479605

Class of tasks involve low entropy, deterministic sequences.

Models can use tools to solve these tasks. This study is theoretically motivated.

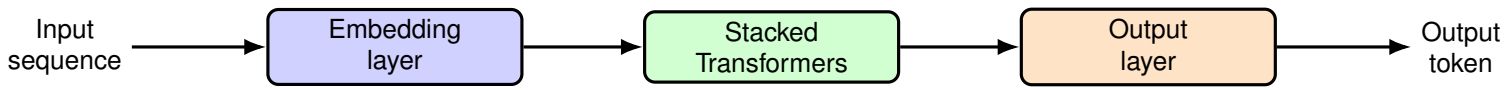
Effective dynamics



Aim: Abstract away from the raw functioning of the LLM to obtain effective dynamics.

Idealized and Effective model

Idealized transformer can be designed for tasks like addition.



We model the LLM using an “effective model”: same architecture as idealized model but erroneous parameters.

$$\mathcal{M}_{\text{eff}} = \mathcal{O} \circ \mathcal{L}_{\text{eff}} \circ \mathcal{E},$$

Accuracy Formula

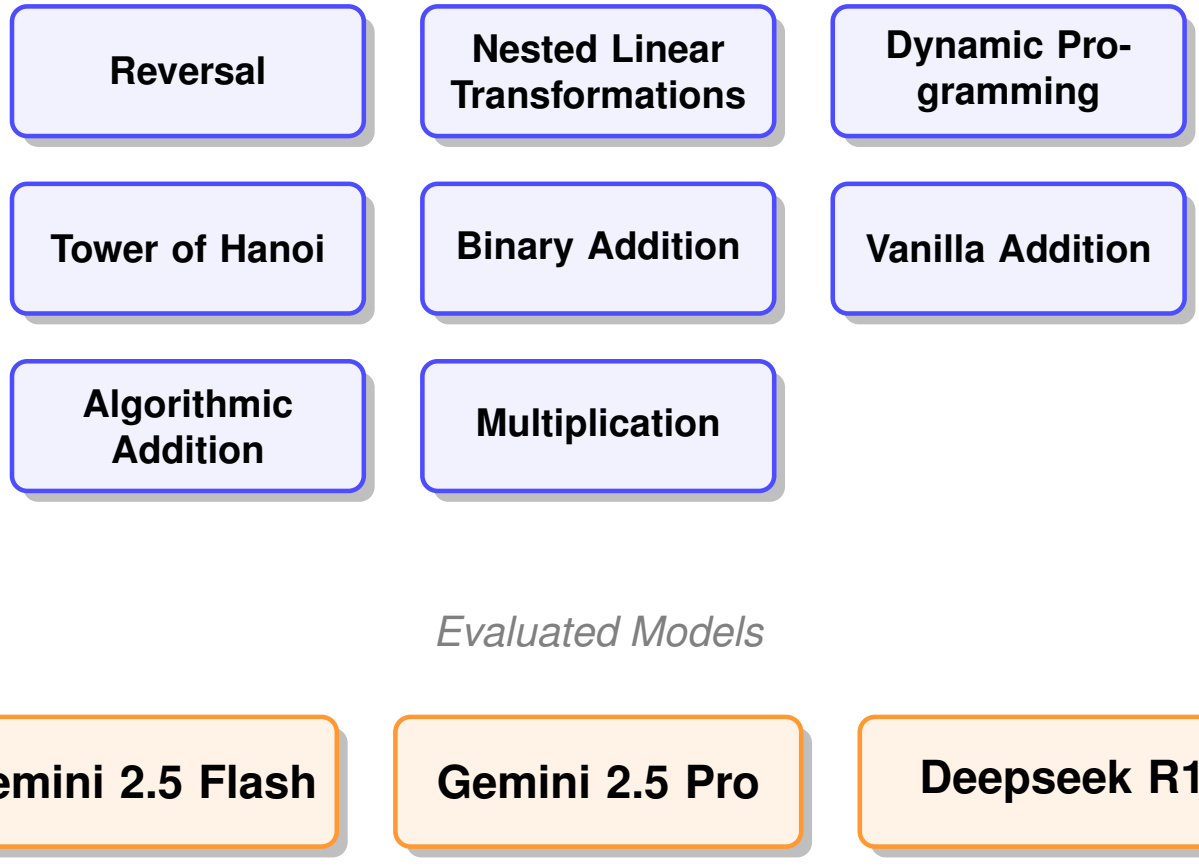
This theoretical analysis leads to a prediction for the model's accuracy on tasks of complexity c.

$$a = \frac{\gamma(\frac{q}{2}, \frac{q}{2rc^2})}{\Gamma(\frac{q}{2})}$$

$$\gamma(a, b) = \int_0^b t^{a-1} e^{-t} dt.$$

q, r have simple interpretation; depend on prompt, model.

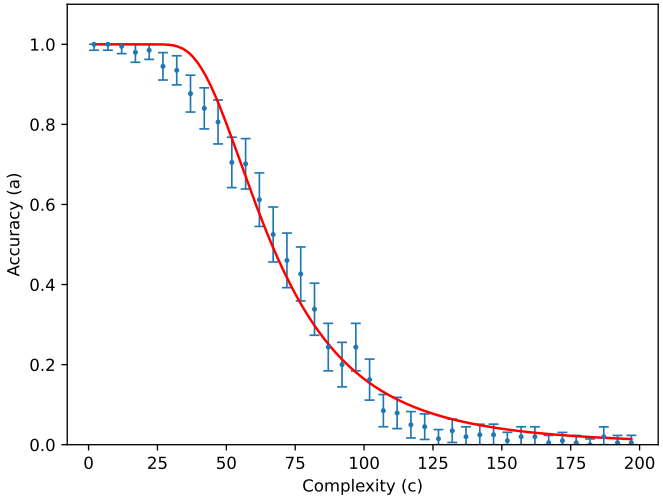
Model Evaluation on Tasks



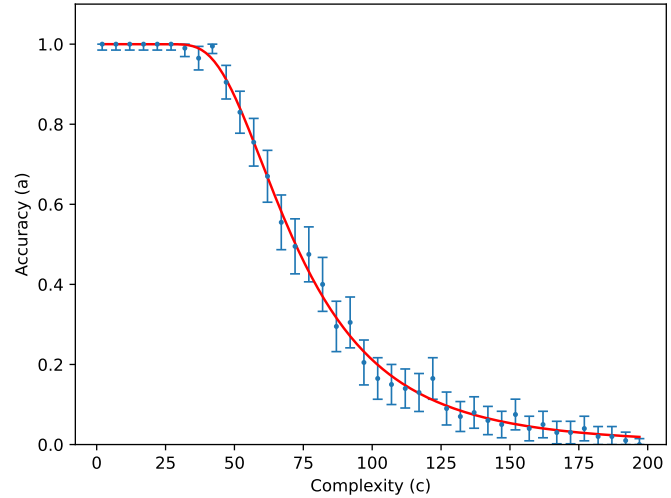
- 0.2 million randomized prompts.

List Reversal

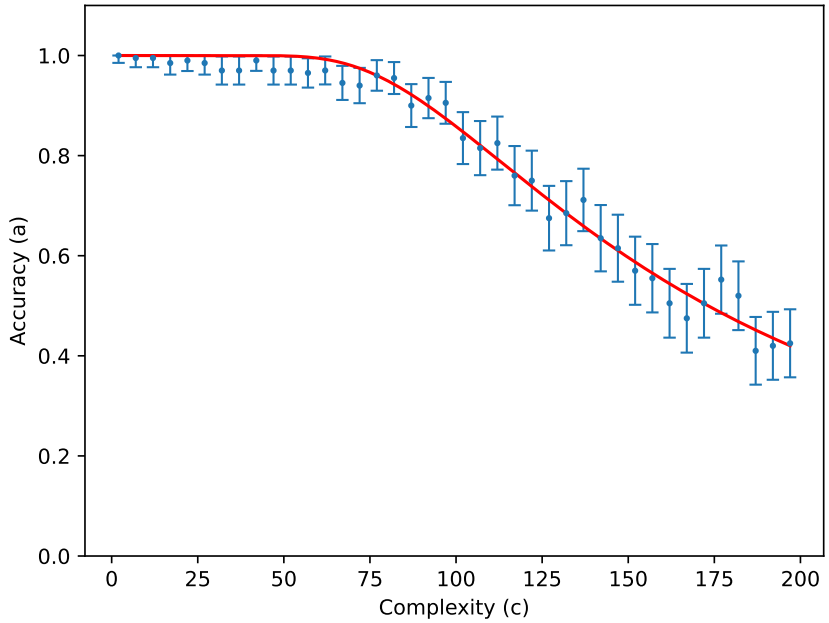
Flash



Deepseek

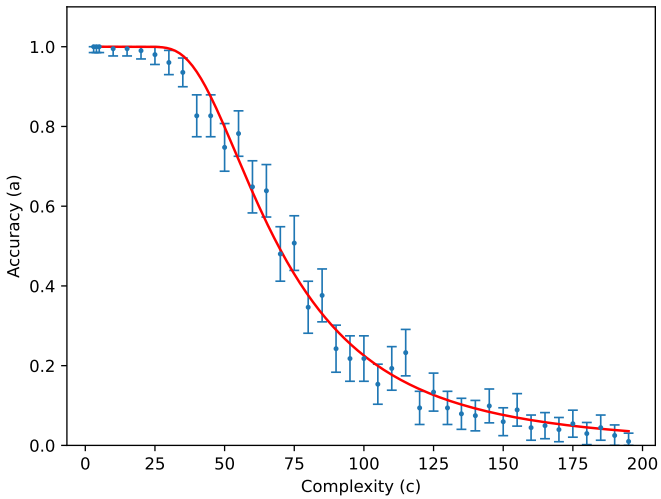


Pro

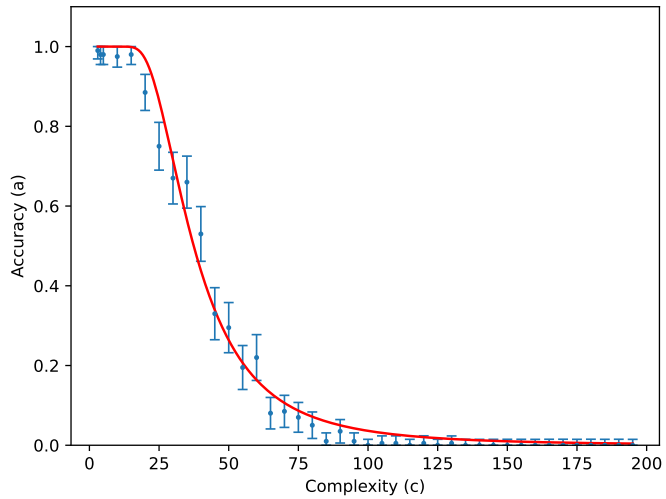


Tower of Hanoi

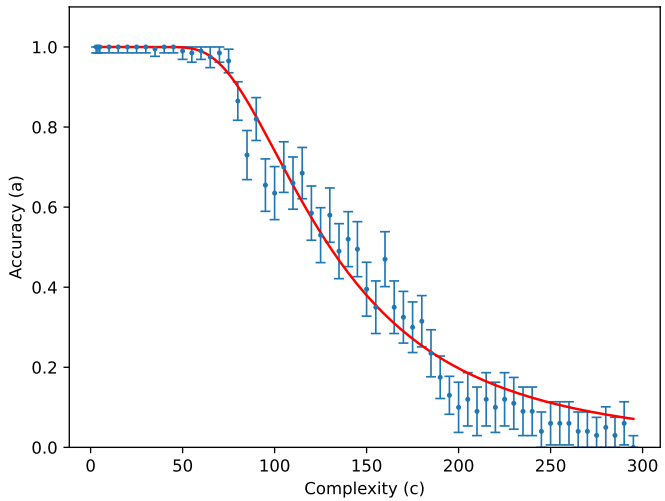
Flash



Deepseek



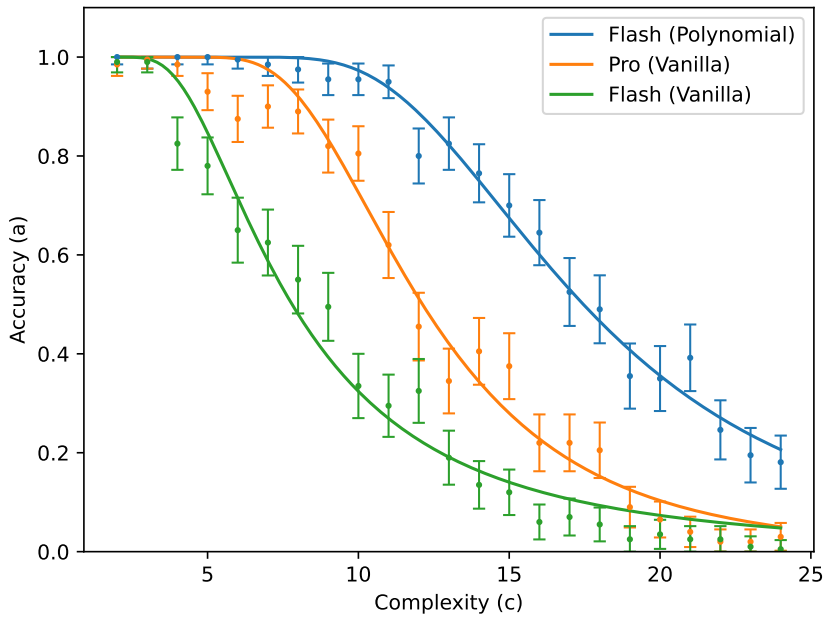
Pro



First *c* moves of Tower-of-Hanoi.

Improving model performance

Our insights lead to improved prompts.



More ambitiously, improve our understanding and also improve model architecture.