Google DeepMind

An evolutionary perspective on modes of learning in transformers

Alexander Y. Ku, Thomas L. Griffiths, Stephanie C.Y. Chan
Google DeepMind, Princeton University

Introduction

Transformers improve inference through two complementary strategies: in-weight learning (IWL) and in-context learning (ICL).

IWL involves the permanent refinement of model parameters, while ICL involves the ephemeral modulation of inferences by leveraging contextual information maintained in the model's activations.

Evolutionary biology suggests that environmental predictability across timescales determines the preferred adaptive strategy.

Genetic evolution adapts to stable environmental features over generations, whereas environmental volatility favors plasticity within a lifetime when reliable cues are present.

We hypothesize that analogous factors govern the competition between ICL and IWL in Transformers.

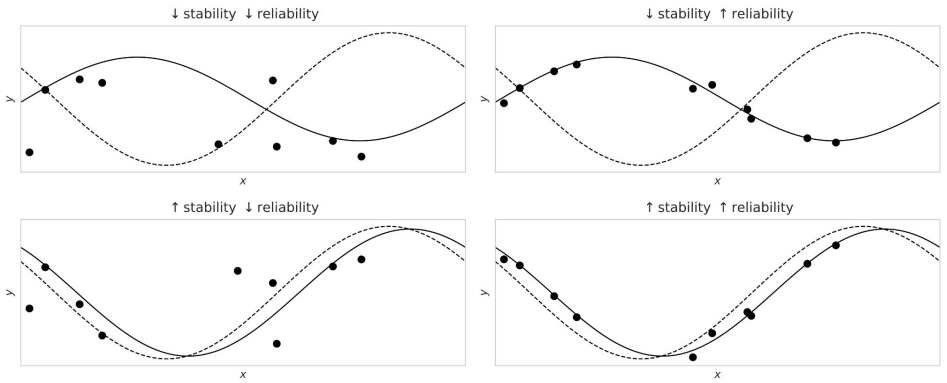
To test this, we operationalize environmental stability and cue reliability in controlled task settings, specifically sinusoid regression and Omniglot classification.

Experimental setup

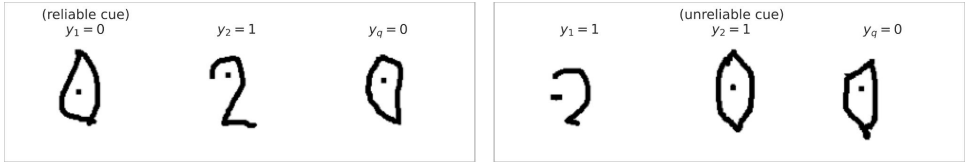
We operationalize two key dimensions of environmental predictability: cue reliability and environmental stability.

Cue reliability is the sufficiency of information within a single prompt to specify the target task, and environmental stability is the invariance of the target task across different prompts during training.

Sinusoid



Omniglot



Evaluation protocol

We quantify the model's preference for ICL versus IWL by computing prediction errors relative to conflicting targets.

We generate evaluation prompts where the underlying target task explicitly conflicts with the current training environment to calculate the errors EICL and EIWL.

The ICL preference score is defined as:

$$S_{\text{ICL}} = \frac{E_{\text{IWL}}}{E_{\text{ICL}} + E_{\text{IWL}}}$$

An SICL=1 indicates pure reliance on context, while SICL=0 indicates pure reliance on weights.

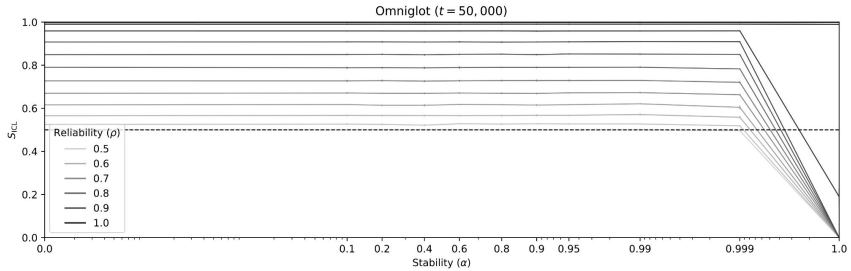
Determinants of asymptotic strategy

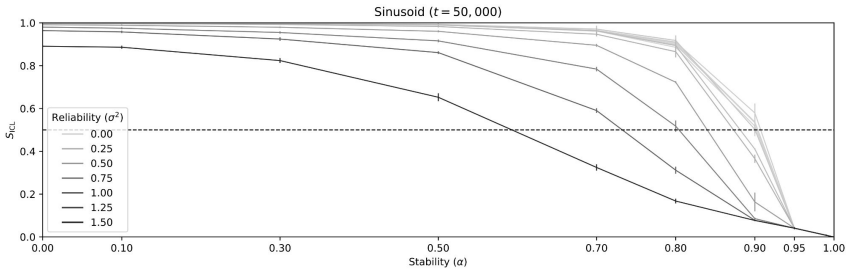
Stable environments favor IWL.

As environmental stability increases, models exhibit a precipitous decline in ICL preference, shifting almost entirely to encoding tasks into model parameters.

Volatile environments with reliable cues favor ICL.

When environmental stability is low, the model relies heavily on ICL, provided that accurate cues are available to guide the adaptive response.

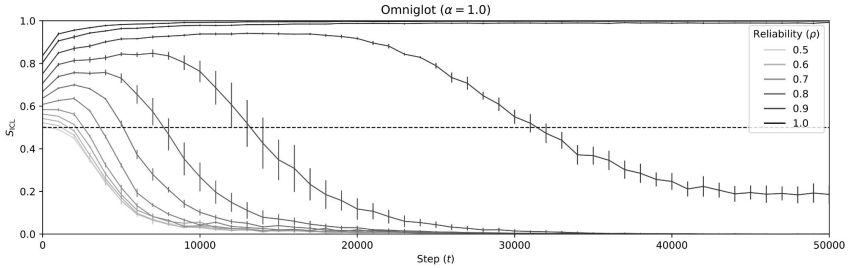




Task-dependent transience

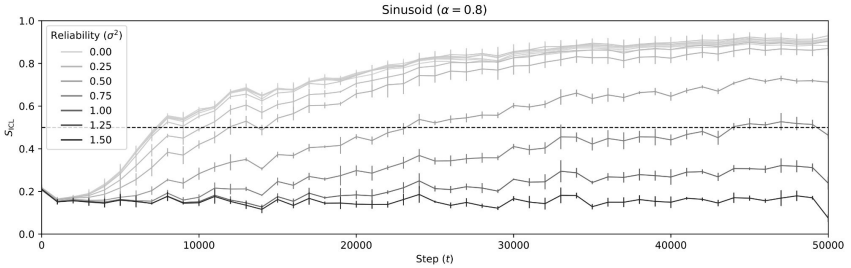
Omniglot classification exhibits ICL transience under high environmental stability.

The model exhibits a strong preference for ICL early in training, but gradually shifts towards an IWL solution as training progresses, mirroring genetic assimilation.



Sinusoid regression exhibits IWL transience, or delayed ICL.

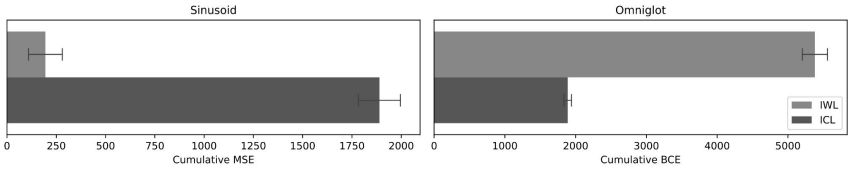
The model initially relies on IWL to approximate the function via weights, and a strong preference for context emerges only later in training.



Cost-based strategy arbitration

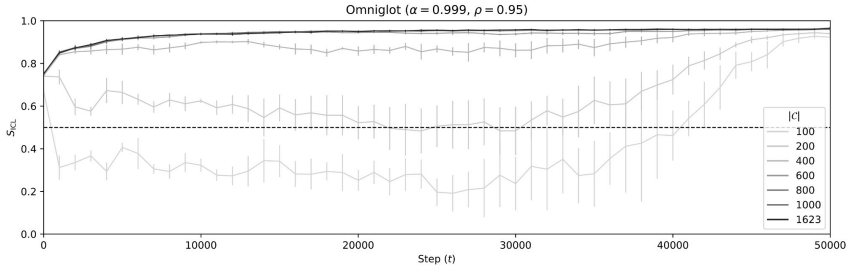
Strategy preference relies on both asymptotic optimality and acquisition cost.

Transformers initially gravitate toward the computationally cheaper strategy, even if it is long-term suboptimal. Prequential codelength reveals opposite cost profiles for our tasks.



Making IWL easier to acquire reverses transience in Omniglot.

Drastically reducing vocabulary size lowers IWL's learning cost below ICL's. Consequently, the model transiently adopts IWL first.



Discussion

An ecological perspective on model behavior paves the way for novel training methodologies.

By understanding how environmental predictability drives strategy selection, we can intentionally design training curricula that cultivate systems capable of flexibly navigating diverse and dynamic real-world environments.

Future research should explore if the early emergence of ICL serves as a scaffold for accelerating the acquisition of IWL solutions.

This would parallel the Baldwin Effect in evolutionary biology, where plasticity smooths the fitness landscape to guide subsequent genetic assimilation.

Thank you.

Google DeepMind