香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

TMLR Young Scientist SEMINAR

2026 SERIES

Trustworthy Machine Learning and Reasoning Group

Mr. Daniel Wurgaft

Ph.D. Student,
Stanford University

Date: 28 Jul 2026 (Tue)

Time: 11:00 – 13:00 (HKT)

Meeting: https://hkbu.zoom.us/j/6603117755

Rational Analysis of In-context Learning

ABSTRACT

This talk presents a normative view of in-context learning (ICL) from the perspective of *rational analysis*, where a learner's behavior is explained as a Bayes-optimal adaptation to data given computational constraints. In the first part, we study Transformers trained on mixtures of ICL tasks, where previous work has identified a broad set of strategies that describe model behavior in different experimental conditions. In aiming to explain these findings, we develop a hierarchical Bayesian model that casts pretraining as a process of updating posterior probability over different strategies, and inference-time behavior as a posterior-weighted mixture over strategies' predictions. Our model almost perfectly predicts Transformer next-token predictions throughout training—without assuming access to its weights—as well as captures a variety of previously-studied ICL phenomena. In the second part, we turn to LLMs and extend our perspective to prompt-based and activation-based control, modeling both in-context learning and activation steering as interventions on beliefs over latent concepts—prompts accumulate evidence, while steering shifts priors. This yields a closed-form Bayesian model that predicts LLM many-shot ICL behavior across both intervention types, accounting for sharp behavioral phase transitions. Taken together, these results highlight the value of taking a top-down Bayesian perspective to understanding neural networks, showing that such a perspective can both yield highly predictive accounts, as well as offer potential explanations for *why* models behave the way they do.

BIOGRAPHY

Daniel Wurgaft is a PhD candidate at Stanford University, advised by Noah Goodman, and a research fellow at Goodfire AI. He studies in-context learning, representational geometry, and reasoning in both language models and humans, drawing on normative frameworks to explain why these systems behave and represent information as they do.

ENQUIRY

Email: bhanml@comp.hkbu.edu.hk