THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Haolin Yang
Department of Statistics
The University of Chicago

"A Unified Mechanistic Framework for In-Context Learning: Hidden State Geometry, Task Recognition, and Task Learning"

May 6, 2026, at 9:00 AM
Jones 111, 5747 S. Ellis Avenue

Abstract

In-context learning (ICL) enables large language models to infer tasks from demonstrations at inference time, without gradient-based parameter updates. Although this capability has been studied extensively, existing mechanistic accounts often focus on only one level of explanation. Component-level studies identify special attention heads, such as induction heads and previous-token heads, as crucial to ICL; representation-level studies emphasize task vectors and hidden-state steering; and functional studies decompose ICL into task recognition and task learning based on perturbing input texts. Yet these perspectives are rarely integrated into a single framework that explains how model components shape hidden-state geometry across layers to produce ICL behavior.
This thesis develops such a unified account. The first part analyzes ICL through the geometry of query hidden states, showing that performance in classification settings is governed by two key properties: separability of hidden states across labels and alignment of these hidden states with label unembedding directions. Empirically, ICL exhibits a two-stage mechanism: early layers primarily increase separability, while middle-to-late layers refine alignment. This analysis further shows that previous-token heads are mainly responsible for inducing separability, whereas induction heads and task-vector-like updates primarily improve alignment.

The second part refines this mechanistic picture through the functional decomposition of ICL into task recognition and task learning. We introduce Task Subspace Logit Attribution (TSLA) to identify attention heads specialized for recognizing the task label space and for learning the mapping from inputs to labels. We show that task-recognition heads align hidden states toward the task subspace, while task-learning heads rotate hidden states within that subspace toward the

---

correct label. This framework reconciles prior observations on induction heads, task vectors, and steering-based interventions, and reveals how component-level operations realize the functional structure of ICL.

Taken together, the two parts of this thesis argue that ICL should be understood as a geometric and mechanistic process in which attention heads progressively reshape hidden states to first organize task-relevant structure and then orient that structure toward correct prediction. This perspective provides a coherent bridge between hidden-state geometry, attention-head circuits, task vectors, and the functional decomposition of ICL.
