Dissertation Abstract・学位論文の内容の要旨

In-context Learning (ICL) has recently emerged as a powerful paradigm enabling Language Models (LMs) to perform few-shot learning without parameter updates. However, despite the success in application, the underlying mechanism of ICL, i.e., why and how LMs perform ICL operation, remains poorly understood. Therefore, this dissertation, which consists of 4 major works, aims to present a systematic investigation into the inner dynamics of LMs under the ICL scenario, and also utilize the gained insights to develop practical applications to improve ICL performance.

(Work 1) Motivated by the gap between the current interpretation of ICL and the observed ICL inference phenomena, first, we establish a comprehensive inference circuit based on the previous works of induction circuits, revealing how the induction circuits work in ICL on real data samples and real LMs. Through careful measurements and causal analysis, we identify three essential operations: (1) Input Text Encode: LMs encode every input text (in the demonstrations and queries) into a linear representation in the hidden states with sufficient information to solve ICL tasks. (2) Semantics Merge: LMs merge the encoded representations of demonstrations with their corresponding label tokens to produce joint representations of labels and demonstrations. (3) Feature Retrieval and Copy: LMs search the joint representations of demonstrations similar to the query representation on a task subspace, and copy the searched representations into the query. Then, language model heads capture these copied label representations to a certain extent and decode them into predicted labels. This decomposition unifies previously fragmented observations of ICL and is empirically validated via ablation, which shows substantial degradation when any step is disabled. Additionally, we confirm and list some bypass mechanisms that solve ICL tasks in parallel with the proposed circuit.

(Work 2) However, operation 3 above requires the ground-truth label in the demonstrations as the source for copying; otherwise, according to our proposed theoretical framework, LMs should fail to make correct predictions in the unseen-label scenario, where the ground-truth labels are absent from the demonstrations. Surprisingly, we find that LMs can still achieve significantly higher accuracy even under this unseen-label scenario, which suggests the existence of another significant mechanism in the aforementioned bypasses. Therefore, we hypothesize another mechanism, called Task-oriented Information Removal, which runs parallel with induction circuits. Specifically, we demonstrate that in the zero-shot scenario, LMs encode queries into non-selective representations in hidden states containing information for all possible tasks, leading to arbitrary outputs without focusing on the intended task, resulting in low accuracy. Meanwhile, we find that selectively removing specific information from hidden states by a low-rank filter effectively steers LMs toward the intended task. Building on these findings, by measuring the hidden states on carefully designed metrics, we observe that few-shot ICL effectively simulates such task-oriented information removal processes, selectively removing the redundant information from entangled non-selective representations, and improving the output based on the demonstrations, which constitutes a key mechanism underlying ICL. Moreover, we identify essential attention heads inducing the removal operation, termed Denoising Heads, which enables the ablation experiments blocking the information removal operation from the inference, where the ICL accuracy significantly degrades, especially when the correct label is absent from the few-shot demonstrations, confirming both the critical role of the information removal mechanism and denoising heads. Notice that such an information removal mechanism does not require the ground-truth labels in the demonstrations, thus explaining the ICL performance in the unseen-label scenario.

(Work 3) Then, we leverage the mechanistic insights gained from the above studies to develop practical applications to enhance ICL performance. Based on the conclusion of “LMs encode every input text into a linear representation in the hidden states with sufficient information to solve ICL tasks”, we develop a new output calibration method called Hidden Calibration, which renounces token probabilities and uses the nearest centroid classifier on the LM's last hidden states. In detail, we assign the label of the nearest centroid previously estimated

from a calibration set to the test sample as the predicted label. Our experiments on 6 models and 10 classification datasets indicate that Hidden Calibration consistently outperforms current token-based baselines by about 20%~50%, achieving a strong state-of-the-art in ICL. Our further analysis demonstrates that Hidden Calibration finds better classification criteria with less inter-class overlap, and LMs provide linearly separable intra-class clusters with the help of demonstrations, which supports Hidden Calibration and gives new insights into the principle of ICL.

(Work 4) Moreover, based on the conclusion of "induction head copy label information to the output", we develop an ICL-oriented fine-tuning method called Attention Behavior Fine-Tuning, building training objectives on the attention scores instead of the final outputs, to force the attention scores of induction heads to focus on the correct label tokens presented in the context and mitigate attention scores from the wrong label tokens. Our experiments on 9 modern LMs and 8 datasets empirically find that ABFT outperforms in performance, robustness, unbiasedness, and efficiency, with only around 0.01% data cost compared to the previous methods. Moreover, our subsequent analysis finds that the end-to-end training objective contains the ABFT objective, suggesting the implicit bias of ICL-style data to the emergence of induction heads. Also, ABFT demonstrates the possibility of controlling specific module sequences within LMs to improve their behavior, opening up the future application of mechanistic interpretability.

In summary, this dissertation improves the mechanistic interpretation of ICL by fitting induction circuits onto the previous observations, then proposing a new mechanism to supply the unseen label scenario. Moreover, based on such a mechanistic understanding, we propose some applicable methods to improve the performance of ICL, and also open up a new research direction, Mechanistic Controllability by the proposed prototypes.

Keywords: Mechanistic Interpretability, In-context Learning, Language Models, Transformer, Low-resource Model Controllability, Representation Learning