How Post-Training Reshapes LLMs: A Mechanistic View on Knowledge, Truthfulness, Refusal, and Confidence

Hongzhe Du[1,*,†], **Weikai Li**[1,*], **Min Cai**[2], **Karim Saraipour**[1], **Zimin Zhang**[3], **Himabindu Lakkaraju**[4], **Yizhou Sun**[1], **Shichang Zhang**[4]
[1]University of California, Los Angeles [2]University of Alberta
[3]University of Illinois at Urbana-Champaign [4]Harvard University

Abstract

Post-training is essential for the success of large language models (LLMs), transforming pre-trained base models into more useful and aligned post-trained models. While plenty of works have studied post-training algorithms and evaluated post-training models by their outputs, it remains understudied how post-training reshapes LLMs internally. In this paper, we compare base and post-trained LLMs mechanistically from four perspectives to better understand post-training effects. Our findings across model families and datasets reveal that: (1) Post-training does not change the factual knowledge storage locations, and it adapts knowledge representations from the base model while developing new knowledge representations; (2) Both truthfulness and refusal can be represented by vectors in the hidden representation space. The truthfulness direction is highly similar between the base and post-trained model, and it is effectively transferable for interventions; (3) The refusal direction is different between the base and post-trained models, and it shows limited forward transferability; (4) Differences in confidence between the base and post-trained models cannot be attributed to entropy neurons. Our study provides insights into the fundamental mechanisms preserved and altered during post-training, facilitates downstream tasks like model steering, and could potentially benefit future research in interpretability and LLM post-training. Our code is publicly available at HZD01/post-training-mechanistic-analysis.

[*] Equal contribution [†] Correspondence: hongzhedu@cs.ucla.edu