Layer by Layer: Uncovering Where Multi-Task Learning Happens in Instruction-Tuned Large Language Models

Zheng Zhao[1] Yftah Ziser[2] Shay B. Cohen[1]

[1] University of Edinburgh [2]Nvidia Research

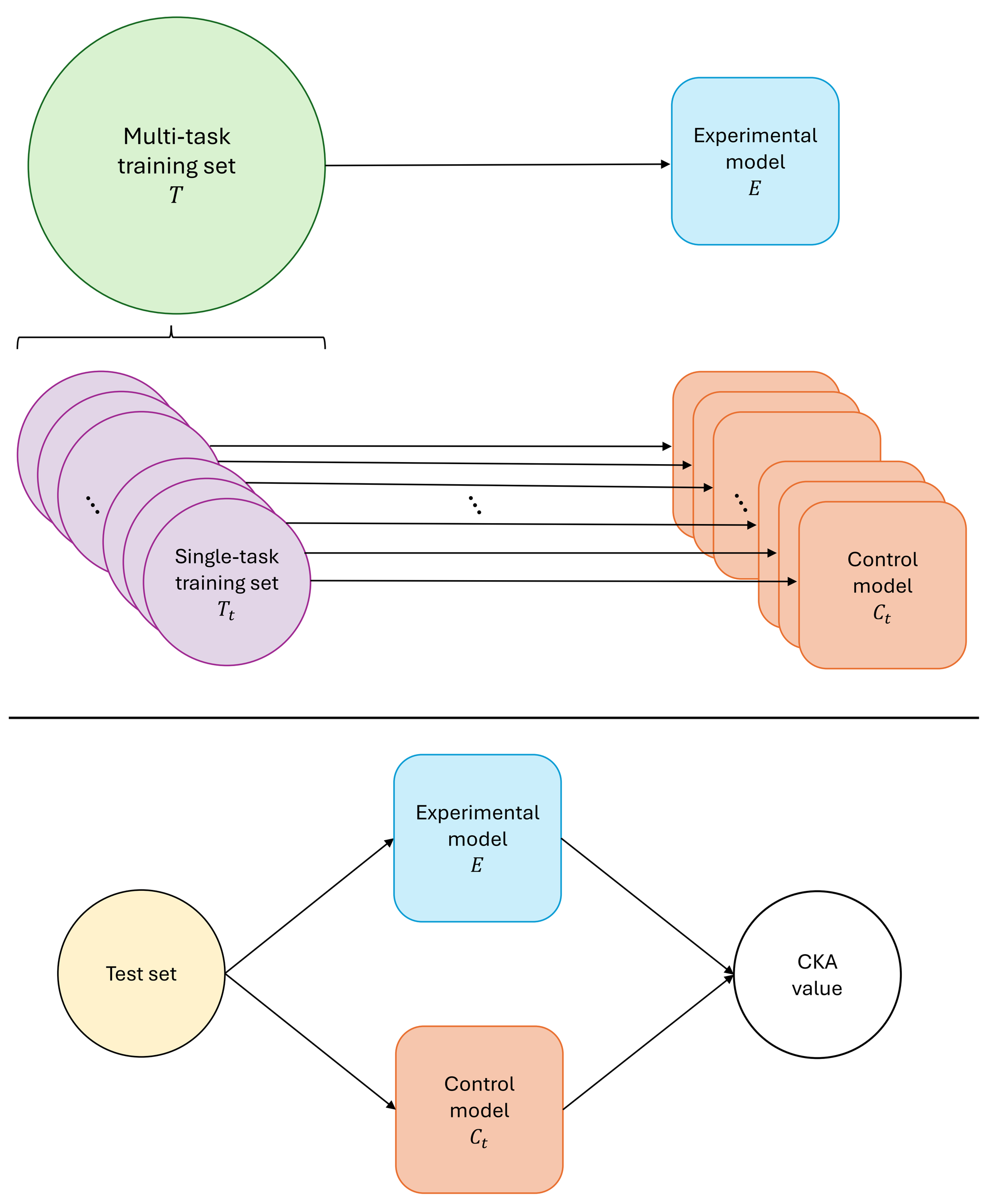
Overview

- Our study investigates *where and how multi-task learning occurs* in instruction-tuned large language models (LLMs), focusing on **task-specific information encoding** within different layers of the model.
- We explore the impact of **instruction tuning** on the representations learned by LLMs across over 60 NLP tasks, contrasting these with **pre-trained** and **task-specific fine-tuned** models.
- Using matrix analysis tools such as *Model-Oriented Sub-population and Spectral Analysis (MOSSA)* and *centered kernel alignment (CKA)*, we assess how instruction tuning modifies task representation in different layers.
- Our findings reveal three key functional groups in model layers: **shared layers** (for general representations), **transition layers** (for task-specific information), and **refinement layers** (for final task optimization).

Methodology

We use the MOSSA framework [1] as an alternative to probing methods. MOSSA compares latent representations, bypassing the challenges of directly comparing task-specific metrics in probing. Here is how MOSSA works:

- Two kinds of models: a multitask model $\mathbf{E}$, and specialized models $\mathbf{C}_t$ for $t \in [T]$;
- Two sets of representations (per task t): $\mathbf{Y}_t, \mathbf{Z}_t$ from examples fed to $\mathbf{E}$ and $\mathbf{C}_t$;
- Apply CKA between these two representations, and the CKA scores are used to quantify the task-specific information encoded in the $\mathbf{E}$ model:

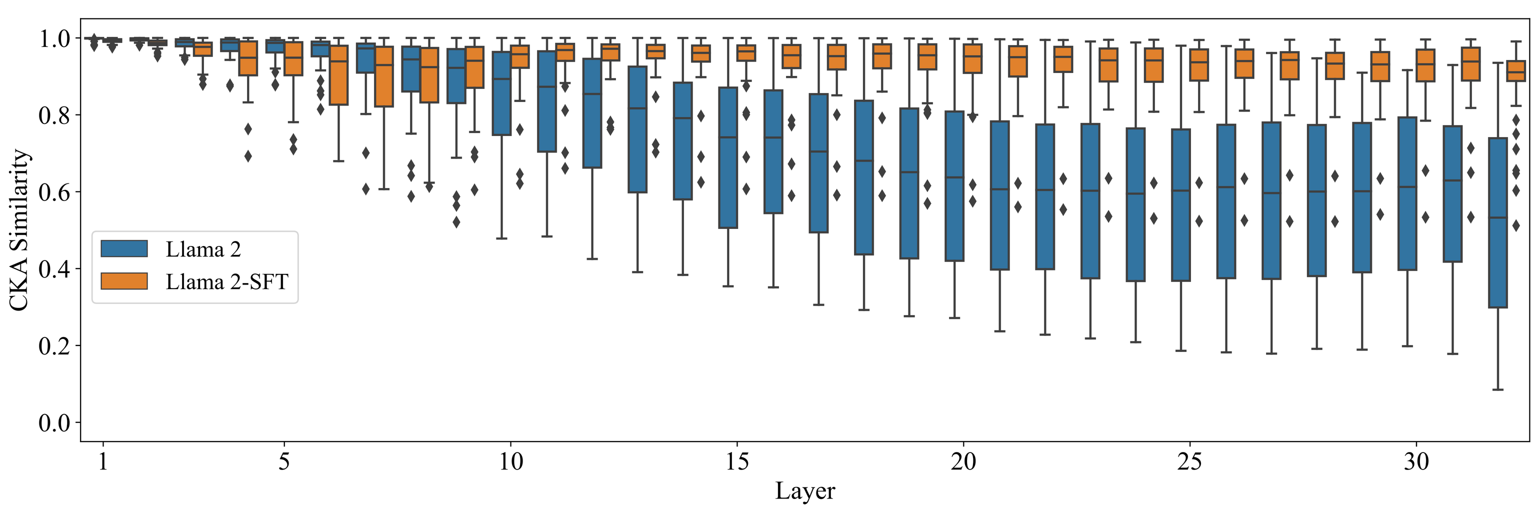


Experimental Setup

- **DATA:** FLAN 2021 for instruction tuning, 60+ NLP tasks over 12 task clusters.
- **MODEL:** Llama 2-7B for all models. All models are trained using LoRA. We refer the multi-task model $\mathbf{E}$ as Llama 2-SFT (instruction-tuned). In some experiments, $\mathbf{E}$ can be the pre-trained Llama 2 model.

Impact of Instruction Tuning

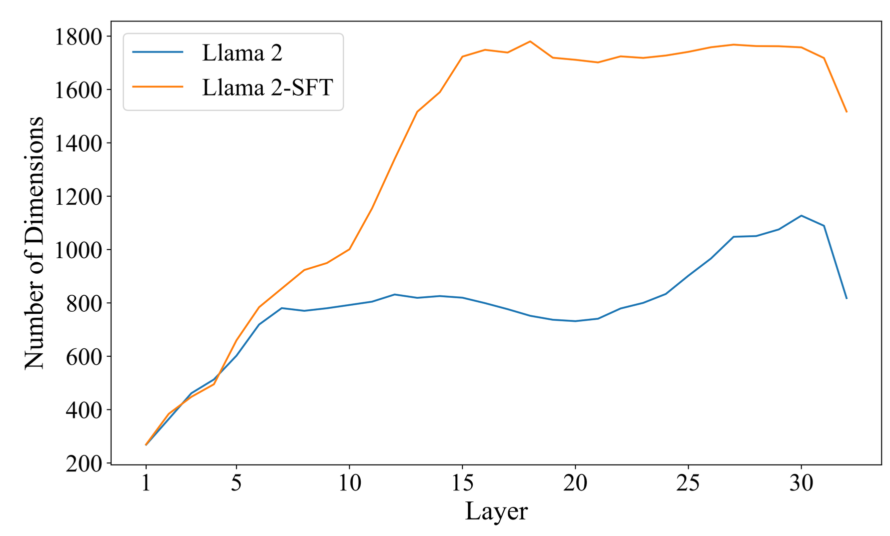
We provide the distribution of CKA similarities across all tasks and layers for the Llama 2 and Llama 2-SFT model.



- Early layers (1-9): Lower CKA scores in Llama 2-SFT vs Llama 2;
- Middle layers (10-15): Llama 2-SFT shows high similarity to control models;
- Final layers (16-32): Similar pattern continues with reduced intensity.

Do We Really Need All Dimensions?

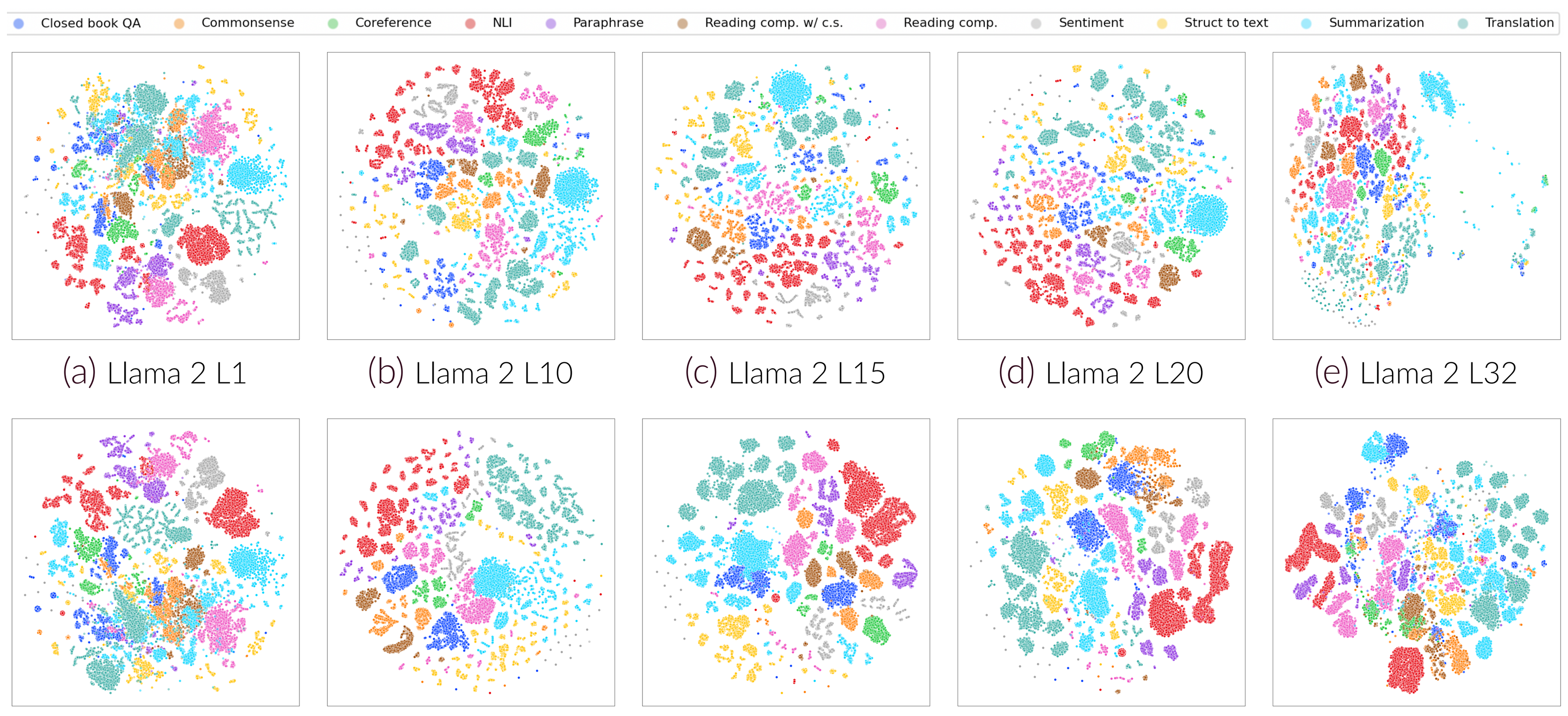
We perform SVD over the representation matrices from different models and calculate number of dimensions needed to explain 99% variance.



- Early layers (1-9): Both models require similar dimensions;
- Middle layers (10-15): Llama 2-SFT needs more dimensions for task-specific features.

Task Cluster Representation Across Layers

We use t-SNE to visualize task clusters across layers for Llama 2 and Llama 2-SFT.



(a) Llama 2 L1 (b) Llama 2 L10 (c) Llama 2 L15 (d) Llama 2 L20 (e) Llama 2 L32

(f) Llama 2-SFT L1 (g) Llama 2-SFT L10 (h) Llama 2-SFT L15 (i) Llama 2-SFT L20 (j) Llama 2-SFT L32

- Early layer (1): Similar clustering in both models;
- Middle layers (10 and 15): Llama 2-SFT shows more distinct task clusters;
- Higher layers (20 and 32): Task clustering intensifies for Llama 2-SFT.

Task Specific Information via Readability

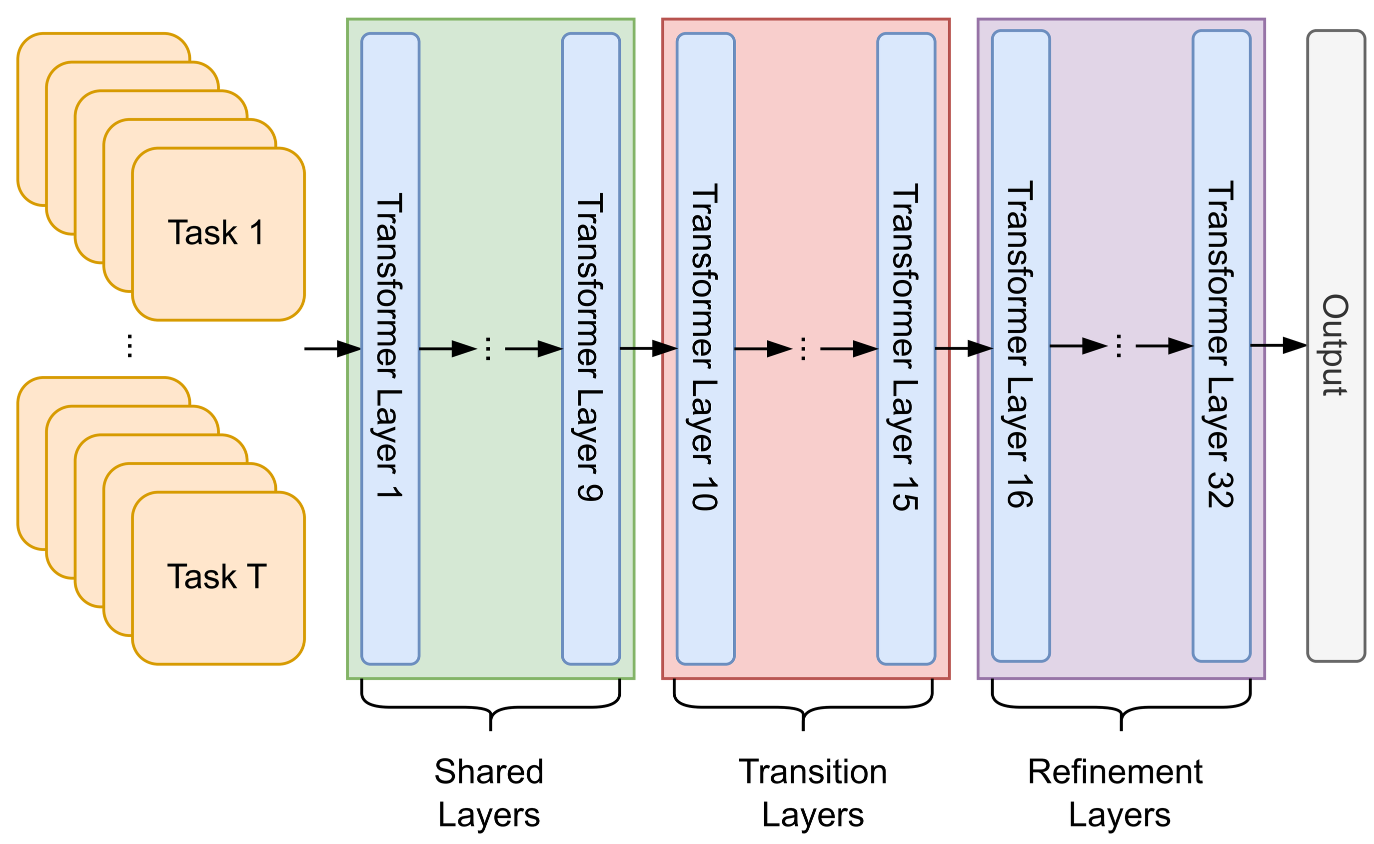
CKA values correlate with reading difficulty (Flesch-Kincaid grade level) in Llama 2-SFT, rising from layer 10, peaking at 15, then saturating.

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Llama 2	-0.03	0.01	0.10	-0.04	-0.09	0.02	-0.06	-0.03	-0.09	-0.11	-0.14	-0.18	0.06	0.11	0.16	0.15
Llama 2-SFT	0.17	0.26	0.25	-0.03	0.02	0.07	-0.01	-0.04	0.04	0.11	0.15	0.23	0.33	0.28	0.38	0.39

Layer	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Llama 2	0.13	0.11	0.11	0.09	0.07	0.06	0.06	0.05	0.05	0.06	0.07	0.07	0.06	0.06	0.02	-0.04
Llama 2-SFT	0.43	0.42	0.42	0.43	0.41	0.40	0.40	0.39	0.40	0.40	0.39	0.40	0.39	0.37	0.35	0.26

Key Takeaway

We discovered three functional groups in instruction-tuned models (Llama 2-SFT):



- **Shared layers** (1-9): Form general representations across all tasks;
- **Transition layers** (10-15): Transform representations into task-specific information;
- **Refinement layers** (16-32): Further refine task-specific representations.

References

[1] Zheng Zhao, Yftah Ziser, and Shay Cohen. Understanding domain learning in language models through subpopulation analysis. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–209, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.