Do different prompting methods yield a common task representation in language models?

Guy Davidson[1]
Todd M. Gureckis[2]
Brenden M. Lake[3]
Adina Williams[1]

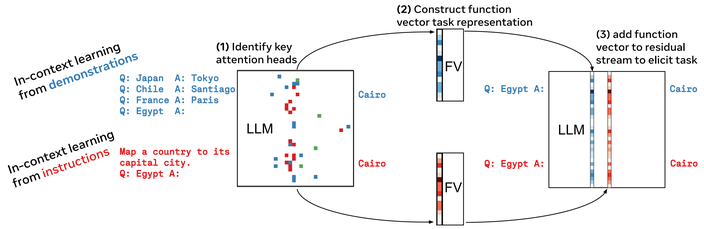
[1] Fundamental AI Research, Meta Superintelligence Labs
[2]Department of Psychology, New York University
[3]Department of Computer Science and Department of Psychology, Princeton University

Take a photo to learn more:

Motivation: Do identical tasks elicited in different ways result in similar representations of the task?

Case study: extracting function vectors from demonstrations and instructions.

Demonstrations: following the methodology of Todd et al. (2024)
Instructions: our contribution (which extends to any prompting approach)



Summary

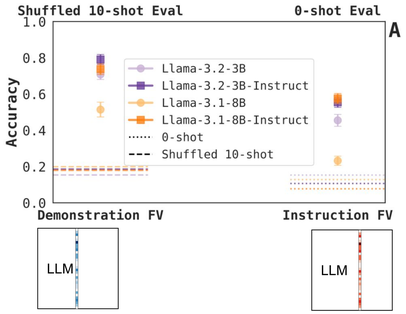
	Claim	Section
Question	Do different ways of presenting the same task elicit a common task representation?	
Method	Extend Function Vectors (FVs) from in-context demonstrations to instructions (and other) task presentations, and analyze their properties.	
Finding 1	Function vectors extracted from instructions facilitate zero-shot task accuracy.	3.1
Finding 2	Demonstration and instruction FVs are beneficial together.	3.2
Finding 3	Instruction and demonstration FVs mostly engage different attention heads.	3.3
Finding 4	Instruction-identified attention heads are more useful for building demonstration FVs than vice versa	3.4
Finding 5	Instruction FVs from post-trained models can steer base models and arise from supervised fine-tuning and preference optimization.	3.5

Function vector extraction technical details

Step	Demonstrations (Todd et al., 2024)	Instructions or other forms (our work)
From a task dataset $\mathcal{D}_t = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_t}, y_{N_t})\}$ **construct prompts** p_i^t	For a query example (x_{iq}, y_{iq}) $p_i^t = [(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \cdots, (x_{iK}, y_{iK}), x_{iq}]$ We use K = 10 (10-shot ICL)	Using task specifications Q_t , sample $q_m^t \in Q_t$ $p_i^t = [q_m^t, x_{iq}]$. We construct Q_t by sampling task instructions from an LLM from task examples.
Filter prompts to construct a set where the model succeeds $P_t = \{p_1^t, p_2^t, \cdots, p_N^t\}$	Randomly samples demonstrations where the model pry_{iq}icts $p_i^t$from	Select the best few task specifications on the training/demonstration set, and sample prompts with those where the model succeeds.
Over the set of successes P_t identified above, record the **mean output of each attention** head as $\bar{a}_{ij}^t = \frac{1}{\lvert P_t \rvert} \sum_{p_i^t \in P_t} a_{ij}(p_i^t)$		
Construct a set of **uninformative baseline in-context prompts** $\tilde{p}_i^t$ for the previously selected P_t	Shuffle the example-label assignments: $\tilde{p}_i^t = [(x_{i1}, \tilde{y}_{i1}), (x_{i2}, \tilde{y}_{i2}), \cdots, (x_{iK}, \tilde{y}_{iK}), x_{iq}]$ …	**(1)** Equiprobable token sequences under the LM **(2)** Real texts matched on length and probability **(3)** Specifications for other tasks (also matched)
Compute the **causal indirect effect** (CIE) of each head in the model f by intervening $a_{ij} := \bar{a}_{ij}^t$ with the mean head outputs: $CIE(a_{ij} \mid \tilde{p}_i^t) = f(\tilde{p}_i^t \mid a_{ij} := \bar{a}_{ij}^t)[y_{iq}] - f(\tilde{p}_i^t)[y_{iq}]$		
Construct a function vector v_t from a small set of attention heads $\mathcal{A}^D$ with the highest CIEs as $v_t = \sum_{a_{ij} \in \mathcal{A}^D} \bar{a}_{ij}^t$		

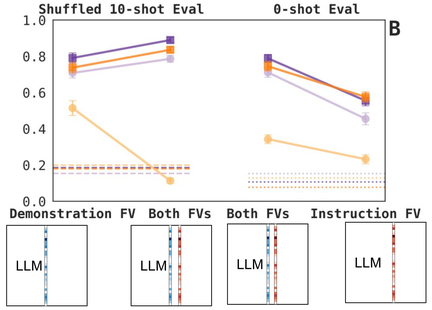
In all shared steps, we follow an identical procedure to Todd et al. (2024)

1 Instruction function vectors facilitate 0-shot accuracy



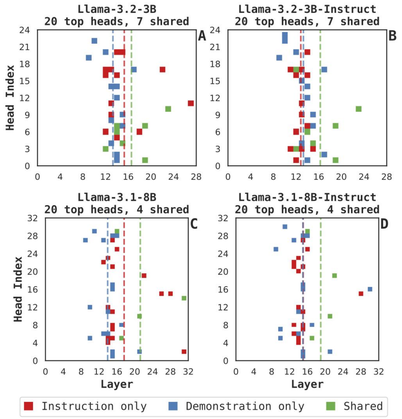
We intervene after the $\lvert L/3 \rvert$ layer (following Todd et al.), highly suboptimal for 3.1-8B. Similar results if we intervene more optimally.

2 Instruction & demonstration FVs are beneficial together



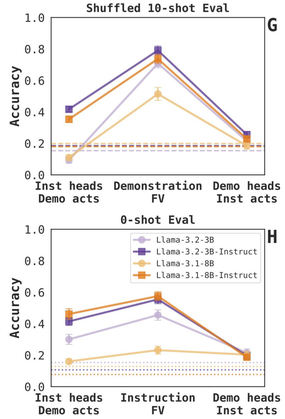
Is this just amplifying the intervention? We test adding the same FV twice and find it explains some, but not all of the benefits.

3 Inst. and demo. FVs mostly engage different attn. heads



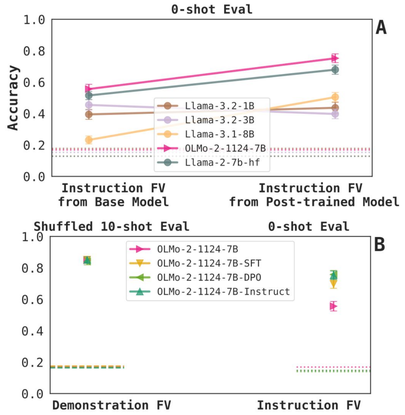
Across all models, 4-8 / 2- of the top attention heads are shared, the rest are distinct. Depths more similar in post-trained models.

4 Instruction heads better for demo FVs than vice versa



Construct ‘incongruent’ FVs: localize heads with one prompting approach, use mean activations from the other.

5 Instruction function vectors facilitate 0-shot accuracy



Transfer successful in the non-distilled models. Post-training stages show gains after SFT and DPO, with minimal RL effect.