

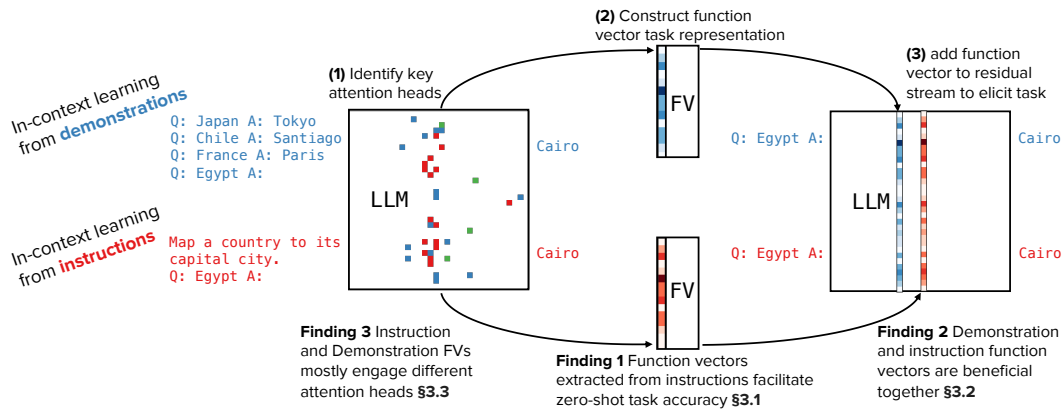
# Do different prompting methods yield a common task representation in language models?

Guy Davidson<sup>1</sup>, Todd M. Gureckis<sup>2</sup>, Brenden M. Lake<sup>3</sup>, Adina Williams<sup>1</sup>

<sup>1</sup>FAIR at Meta, <sup>2</sup>New York University, <sup>3</sup>Princeton University  
guy.davidson@nyu.edu  
todd.gureckis@nyu.edu, brenden@princeton.edu  
adinawilliams@meta.com

## Abstract

Demonstrations and instructions are two primary approaches for prompting language models to perform in-context learning (ICL) tasks. Do identical tasks elicited in different ways result in similar representations of the task? An improved understanding of task representation mechanisms would offer interpretability insights and may aid in steering models. We study this through *function vectors* (FVs), recently proposed as a mechanism to extract few-shot ICL task representations. We generalize FVs to alternative task presentations, focusing on short textual instruction prompts, and successfully extract instruction function vectors that promote zero-shot task accuracy. We find evidence that demonstration- and instruction-based function vectors leverage different model components, and offer several controls to dissociate their contributions to task performance. Our results suggest that different task prompting forms do not induce a common task representation through FVs but elicit different, partly overlapping mechanisms. Our findings offer principled support to the practice of combining instructions and task demonstrations, imply challenges in universally monitoring task inference across presentation forms, and encourage further examinations of LLM task inference mechanisms.



**Figure 1: Language model task representations depend on the form of task presentation.** We compare in-context learning task representations formed from **demonstrations** with those formed from **instructions** using function vectors (FVs). The process of extracting FVs is shown in Steps (1)-(3). We highlight several findings: We successfully extract FVs from instructions (§3.1); Instruction FVs offer complementary benefits when applied with demonstration FVs (§3.2); Different prompting methods yield distinct task representations (§3.3; highlighted squares on the left LLM are Llama-3.1-8B-Instruct attention heads: those identified by **demonstrations** only, those identified by **instructions** only, and **shared** ones; columns are head indices, rows are head indices, see Figure 2D).

# 1 Introduction

If you prompt a large language model (LLM) with in-context examples “Q: Japan A: Tokyo Q: Chile A: Santiago Q: France A: Paris Q: Egypt A:” or with instructions “Map a country to its capital city: Q: Egypt A:”, you expect to get the same answer (Figure 1). These two prompts share only the final query in common, but imply the same underlying task for the network to perform. This paper explores how these two prompting methods result in similar or different representations of the specified task in the language model. We consider these representations *task representations*, as they reflect information in the network that induces the correct mapping for a given task, rather than the answer itself.

We focus on two prevalent approaches to specifying tasks to LLMs: demonstrations and instructions. The ability of LLMs to perform in-context learning (ICL) from demonstrations has been of considerable interest since GPT-3 (Brown et al., 2020; Lampinen et al., 2024). With appropriate fine-tuning, language models can also follow textual instructions, facilitating a far broader range of use cases (Chung et al., 2022). We follow a considerable literature studying the mechanisms that govern ICL (e.g., Olsson et al., 2022; Chen et al., 2024; Akyürek et al., 2024) and those that promote instruction-following (Stolfo et al., 2024; Wu et al., 2024). Prior work studied either ability by itself; here, we examine to what extent the two share representations and mechanisms.

We approach this question by leveraging a new interpretability method known as *function vectors* (FVs, Todd et al., 2024). FVs are *causal* patterns identified from the intermediate layers of a language model that mediate its ability to perform a task. Todd et al. demonstrate that FVs, as a single additive intervention to a model’s latent activity, successfully cause the execution of a task in a different (or empty) context. We extend Todd et al.’s FV extraction method from using specifically in-context demonstrations to any form of task presentation, including instruction prompts. We then compare the representations elicited by demonstrations and textual instructions, examining their effectiveness at inducing task-following behavior, latent activity similarity, and elicited internal mechanisms.

Table 1 summarizes our key findings. Our extension of the function vector identification procedure successfully extracts instruction function vectors, and these promote zero-shot task accuracy (§3.1). We validate that demonstration and instruction FVs contain complementary information by intervening with both simultaneously, and find it conveys task performance benefits beyond using either FV alone (§3.2). Next, we examine the attention heads identified in demonstration and instruction FVs, and find that most are identified by one type of FV, with only a few shared by both (§3.3). We evaluate the functional implication of the different attention heads and find an asymmetry between the relevance of demonstration-identified attention heads to instructions and instruction-identified heads to demonstrations (§3.4). Finally, noting that instruction FVs function better in post-trained models, we find that we can steer base models with post-trained model instruction FVs, and we identify the relevant post-training stages for instruction FV extraction (§3.5).

Overall, our results suggest that **different task presentations do not induce a common task representation through function vectors, but activate partly overlapping mechanisms and induce jointly beneficial representations**. Our findings provide new insights into how LLMs represent tasks and offer an explanation for why combining instructions and in-context examples often improves model performance.

Table 1: Guiding research question and summary of findings

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

## 2 Methods

We describe the function vector extraction procedure outlined by Todd et al. (2024) and our extension from in-context demonstrations to arbitrary task prompts (see Todd et al., 2024, for additional details). The procedure identifies a small set of causally relevant heads in a model  $f$  for performing a given task  $t$ , and uses them to compute the function vector  $v_t \in \mathbb{R}^{d_{\text{model}}}$ , where  $d_{\text{model}}$  is the model’s latent dimension. Each task  $t \in \mathcal{T}$  (the set of tasks considered) consists of a supervised dataset  $\mathcal{D}_t = \{(x_1, y_1), (x_2, y_2), \dots, (x_{N_t}, y_{N_t})\}$ , where  $x_i$  and  $y_i$  are (tokenized) string inputs and outputs respectively. We construct  $K$ -shot in-context demonstration prompts  $p_i^t$  for a query example  $(x_{iq}, y_{iq})$  as  $p_i^t = [(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \dots, (x_{iK}, y_{iK}), x_{iq}]$ . To identify causally relevant attention heads, consider a set of prompts  $P_t$  for a given task  $t$  on which a model  $f$  succeeds in predicting  $y_{iq}$  from  $p_i^t$ ,  $P_t = \{p_1^t, p_2^t, \dots, p_N^t\}$ . Denote the output of attention head  $a_{lj}$  (head  $j$  in layer  $l$ ) in processing the final token of prompt  $p_i^t$  as  $a_{lj}(p_i^t) \in \mathbb{R}^{d_{\text{model}}}$  (projected to the model’s latent dimension). We compute the mean task-conditioned activation of each attention head,  $\bar{a}_{lj}^t = \frac{1}{|P_t|} \sum_{p_i^t \in P_t} a_{lj}(p_i^t)$ . Next, we construct uninformative baseline in-context demonstration prompts  $\tilde{p}_i^t$  by shuffling the labels  $\tilde{y}_{ik}$  assigned to each  $\tilde{x}_{ik}$ :  $\tilde{p}_i^t = [(\tilde{x}_{i1}, \tilde{y}_{i1}), (\tilde{x}_{i2}, \tilde{y}_{i2}), \dots, (\tilde{x}_{iK}, \tilde{y}_{iK}), x_{iq}]$ . We seek to score each head by its causal indirect effect toward predicting the correct  $y_{iq}$  from the shuffled prompt  $\tilde{p}_i^t$ ; that is, attention heads that promote task accuracy in the context of the shuffled prompt. To compute this, denote  $a_{lj} := \bar{a}_{lj}^t$  the intervention of setting the output of attention head  $a_{lj}$  to its task-conditioned activity  $\bar{a}_{lj}^t$ . Then compute the difference in probabilities assigned to the first token of the correct  $y_{iq}$  between when we intervene on the model  $f$  and when we do not:  $CIE(a_{lj} \mid \tilde{p}_i^t) = f(\tilde{p}_i^t \mid a_{lj} := \bar{a}_{lj}^t)[y_{iq}] - f(\tilde{p}_i^t)[y_{iq}]$ . Todd et al. (2024) find that a small set of heads consistently achieve high causal scores across their task set. Using  $\mathcal{A}^D$  to denote this set of top heads, FVs are computed for each task using the heads’ task-conditioned means:  $v_t = \sum_{a_{lj} \in \mathcal{A}^D} \bar{a}_{lj}^t$ .

### 2.1 Generalizing function vectors beyond in-context demonstrations

One of our contributions is to generalize the function vector method described above from demonstrations to alternative task specifications. We denote by  $Q_t$  a set of task specifications (of any form) for the task  $t$  (e.g., the instructions in Figure 1). For a query example  $(x_{iq}, y_{iq})$  (Egypt  $\Rightarrow$  Cairo in Figure 1), we sample  $q_m^t \in Q_t$  and construct prompts  $p_i^t$  as  $p_i^t = [q_m^t, x_{iq}]$ . Next, we address the challenge of creating uninformative baselines  $\tilde{q}_m^t$  for the task specifications  $q_m^t$ . We consider three approaches to generating these baselines, and see further details and examples in Appendix B:

- **Equiprobable token sequences.** *Intuition:* sample token sequences that are similarly likely under the model but are unrelated to the task. Each  $q_m^t$  is encoded as a token sequence  $q_m^t = [w_1, w_2, \dots, w_L]$  whose probability under the model  $f$  is  $P(q_m^t) = \prod_{l \leq L} P(w_l \mid w_{<l}) = f(w_{<l})[w_l]$ . Starting from the BOS token, we sample  $\tilde{w}_l$  to approximately match the conditional probabilities  $f(\tilde{w}_{<l})[\tilde{w}_l] \approx f(w_{<l})[w_l]$ , and construct  $\tilde{q}_m^t = [\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_L]$ .
- **Real texts.** *Intuition:* sample texts from a natural corpus that convey no task information but otherwise match the task specifications. We score token sequences from a chosen corpus under the model and sample  $\tilde{q}_m^t$  with approximately the same length and probability as  $q_m^t$ .
- **Other task specifications.** *Intuition:* sample task specifications for other tasks  $t' \neq t$  that likely convey no information for task  $t$ . We score task specifications  $q_{m'}^{t'}$  generated for other tasks  $t' \in \mathcal{T}$ , and again sample ones with approximately the same length and probability as  $q_m^t$ .

Following Todd et al.’s (2024) selection of prompts in which the model successfully performs the demonstrated task, we focus on a best-performing small set of alternative task specifications we denote as  $Q_t^* = \{q_1^{t*}, \dots, q_J^{t*}\} \in Q_t$ , with the intuition that these facilitate forming the most salient representation of the task  $t$ . We construct prompts  $p_i^t = [q_m^t, x_{ik}]$ , keep only ones in which the model successfully predicts  $y_{ik}$ , use these to select  $Q_t^*$ , and narrow down our set of prompts to a final  $P_t^*$ , where  $p_i^{t*} = [q_m^{t*}, x_{ik}]$ . From there we follow the original function vector procedure: we compute task-conditioned activations (over  $P_t^*$ ), generate uninformative baselines  $\tilde{p}_i^{t*} = [\tilde{q}_m^{t*}, x_{ik}]$ , compute the indirect effects for each attention head, and use these to select the top heads (in practice, we average over these different baselines in computing head causal effects; see below).

**Constructing function vectors from in-context instructions** Although the method is more general, here we use zero-shot textual instructions as our  $Q_t$ . We generate a candidate set of textual

instructions for each task by querying Llama-3.1-405B (Llama Team, 2024) using a  $K$ -shot ICL prompt for the task, and ask it to generate 10 instruction prompts for the task it infers from the provided demonstrations (see Appendix A for the templates). We repeat this procedure 20 times for each task  $t$  and deduplicate exact repetitions to arrive at a task instruction set  $Q_t$ . We repeat this procedure twice, once encouraging the model to generate short instructions and once with no such encouragement. We provide sample generated instructions of both lengths in Appendix A.1 and corresponding uninformative baselines (using all three approaches) in Appendix B.2. For our ‘real texts’ uninformative baseline, we sample texts from WikiText-103-v1 (Merity et al., 2016).

## 2.2 Experimental conditions

**Hyperparameters.** We match the settings used by Todd et al. (2024): we use 100 examples to compute mean task-conditioned activations and 25 for the indirect effects. We do so over the  $J = 5$  top textual instructions for each task, splitting the examples evenly between the top instructions.

**Models.** We focus on the base and instruction-tuned versions of the 3B Llama-3.2 and 8B Llama-3.1 models, with the full list in Table 4. We also report some results with the weaker 1B Llama-3.2 and 7B Llama-2 models (Touvron et al., 2023), the latter of which matches Todd et al.’s (2024). Finally, to examine the roles of post-training stages, we evaluate four OLMo-2 models (OLMo et al., 2024).

**Tasks.** We consider the same set of tasks and datasets used by Todd et al. (2024). We omit a few classification datasets where successfully predicting the next token requires an understanding of the format that is facilitated by demonstrations but not necessarily by minimal textual instructions (we retain a total of 50 datasets; see Appendix D for the full list). We follow Todd et al. (2024) in computing the top sets of heads only over datasets where a model surpasses chance performance. As our approach requires 20 successful prompts for each of the five best instructions to compute the mean task-conditioned activations, we also omit tasks where a model fails to pass this number of prompts (which, in most cases, means it is also below chance accuracy and was already omitted).

**Textual instructions and uninformative baselines.** We evaluate each model on each task using one random seed six separate times, for the short ( $\leq 16$  tokens) vs. longer (unbounded) instructions crossed with all three uninformative baseline approaches (§2.1). We observe minimal deviation in top heads across these conditions (see Appendix H.3), so we average the causal effects across all six to compute the top heads. We then report final evaluation accuracies averaged over the results with both sets of mean activations, those generated with short instructions and those from longer ones.

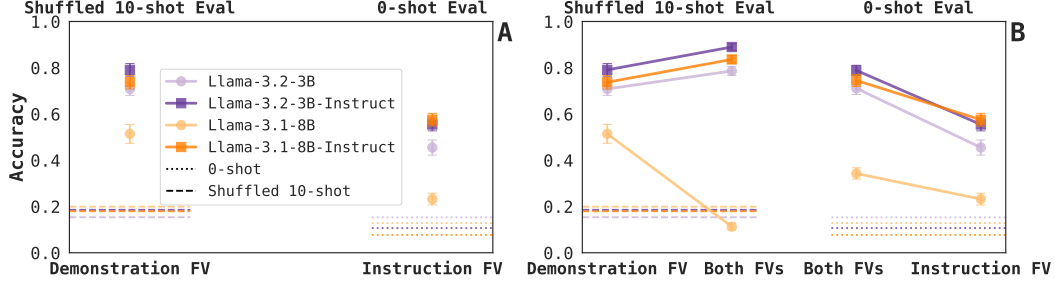
**Results using in-context demonstrations.** To enable comparison, we also replicate the original, demonstration-based function vector evaluation with all models we consider, using the same random seeds (and therefore, same train-evaluation splits) as used in the textual instruction setting.

**Evaluation and Comparison Logic.** We structure our evaluations to facilitate comparison between FVs derived from demonstrations and instructions. We use the  $|\mathcal{A}| = 20$  top heads in both settings, consistent with Todd et al. (2024) for the 7B models. We follow Todd et al. (2024) in evaluating function vectors as an additive intervention to the residual stream after the  $\lfloor L/3 \rfloor$  layer (layer 9 for the Llama-3.2-3B models and layer 11 for the Llama-3.1-8B and OLMo-2-1123-7B ones). We report two evaluation settings: 10-shot with shuffled labels ( $\tilde{p}_i^t$  in §2.1) and 0-shot (with no instructions). We focus our assessment of each FV in the setting matching its extraction: shuffled 10-shot for demonstration FVs and 0-shot for instruction FVs. In addition to the FV evaluations, we report a baseline of the evaluation setting without adding the function vector. We report accuracies using informative task presentations for each evaluation in Appendix F: instructed 0-shot (averaging over the top  $J = 5$  instructions for each model and task), and 10-shot (without label shuffling).

## 3 Results

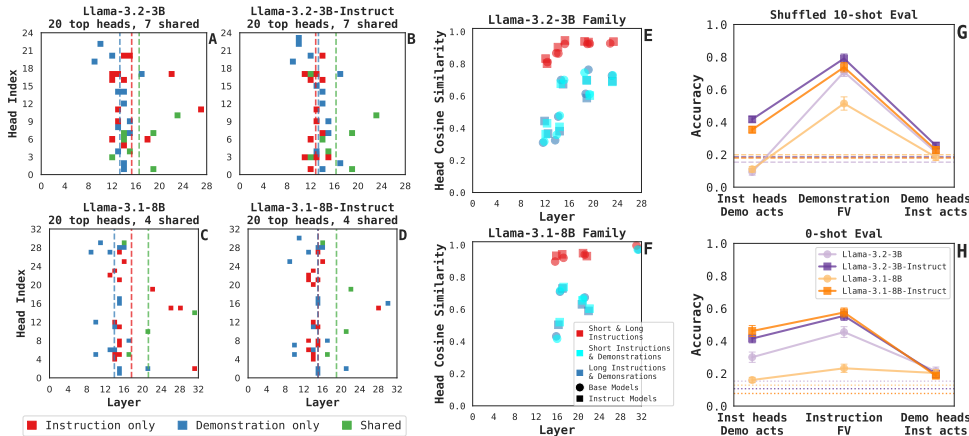
### 3.1 Function vectors extracted from instructions facilitate zero-shot task accuracy

Figure 2A summarizes our evaluation of demonstration function vectors (left) and instruction-based FVs (right). We find that **our adaptation of the FV extraction procedure succeeds and increases zero-shot task accuracy** from below 20% to above 50% in the best models. A notable exception is the Llama-3.1-8B model, for which the  $\lfloor L/3 \rfloor$  intervention depth appears suboptimal (see results from other intervention layers in Figure 11). On the other hand, the Llama-3.2 models may benefit



**Figure 2: Instruction-based function vectors are effective in zero-shot evaluation; both function vectors are beneficial together.** (A) We evaluate each FV in the setting matching its extraction: shuffled 10-shot for demonstrations and 0-shot for instructions. Both FV types are effective at their respective evaluations, though demonstration ones fare better. This procedure is effective for both post-trained models and for the distilled Llama-3.2-3B base model; less so for the Llama-3.1-8B base model. (B) We examine the effect of jointly adding both FVs together. The joint intervention outperforms either one by itself (with the exception of the base Llama-3.1-8B model, for which the  $|L/3|$  intervention depth appears highly suboptimal; see Figure 11). Dotted lines represent baselines with no FV intervention and error bars reflect standard errors of the mean (SEMs). See Figure 11 for results when intervening at optimal depths, rather than a fixed  $|L/3|$  depth, and Figure 12 for results evaluating the function vectors in the setting opposite to their extraction.

from being distilled from larger, post-trained Llama-3.1 models. The top  $|A| = 20$  attention heads also represent a slightly larger fraction of attention heads in the smaller models (about 3% in the 3B models vs. 2% in the 8B ones). We also observe that, unsurprisingly, instruction FVs are substantially more effective in post-trained models. Finally, while instruction FVs are highly effective in the zero-shot setting, they fail to match the accuracy demonstration FVs attain in the shuffled 10-shot condition, and struggle when evaluated in the shuffled 10-shot condition themselves (see Figure 12). We replicate findings 1 and 2 (below) with other models: see Figures 13-15, in appendices G.3-G.4.



**Figure 3: (A-D) Demonstrations and instructions elicit mostly distinct attention heads.** We visualize the sets of top heads identified only with **demonstrations**, only with **instructions**, and **shared** in both cases. Only a few top heads are shared each in model (see Appendix H.1 for additional models). We observe that top instruction FV heads tend to be in earlier layers of the post-trained (-Instruct) models, compared to the base versions, and that shared heads tend to arise in later layers. (E-F) **Shared head activation similarity depends on layer, but not on post-training.** We plot the cosine similarity of activations in shared heads between three sets of task-conditioned activations: those elicited by demonstrations, by short instructions, and by longer instructions. We observe increasing similarity between demonstrations and instructions in later model layers, but show no effect of post-training on similarity. (G-H) **‘Heterogeneous’ function vectors show demonstration-instruction asymmetry.** We construct ‘heterogeneous’ FVs using attention heads localized with instructions and demonstration task activations (left data points) to the opposite combination, using demonstration-localized heads and instruction task activations (right). Compared to the ‘homogeneous’ function vectors (middle), we mostly observe a smaller accuracy drop when using instruction FV heads to read from demonstration mean activations than vice versa.

We attribute these differences to two factors. Demonstration FVs arise from the attention heads that most improve shuffled 10-shot accuracy (as those were the uninformative prompts  $\tilde{p}_i^t$ ; §2.1). In contrast, 0-shot evaluation differs from the instruction FV baseline prompts, as it contains no instructions (and hence, has fewer tokens than the uninformative instructions  $\tilde{q}_i^t$ ; §2.2). Further, ICL from demonstrations appears more primal to LLMs than from instruction-following. The former appears in base models, while the latter requires post-training; we support this claim with later results.

Constructing function vectors seems to benefit from model capabilities. Smaller Llama-3.2-1B models show a lower accuracy increase with instruction FVs; conversely, newer OLMo-2-1124-7B models show a greater increase (see Appendix G.4). This capacity is not unique to the latest models; results with Llama-2-7B models qualitatively match newer models. Finally, we evaluate the effect of choosing a fixed intervention layer in Figure 11. If we choose the optimal layer for each model and evaluation, the accuracy numbers rise as expected, but the overall patterns remain the same.

### 3.2 Demonstration and instruction function vectors are beneficial together

If demonstrations and instructions elicit different task representations in a language model, to what extent are their representations jointly beneficial? To study this, we intervene with both FVs, adding them to the residual stream after the same  $\lfloor L/3 \rfloor$  layer. Our results in Figure 2B demonstrate that **adding both function vectors appears consistently beneficial**, except for the base Llama-3.1-8B model (for which the choice of layer seems crucial). Surprisingly, adding both FVs after the same layer does not appear to induce interference, even though these FVs were extracted independently. In Figure 11B, we report results from the highest accuracy intervention layers in a sweep over layer pairs; adding both FVs to the same layer often proved optimal. Finally, to establish how much of this effect is simply due to amplifying the intervention, we report the results of adding the same FV twice in Figure 18. For some models and evaluations, adding the same FV twice performs similarly to adding both FVs. This suggests that some of the benefit is due to the magnitude of the intervention rather than the different information carried in both FVs; we leave studying the extent of these relative effects to future work. Having found behavioral evidence that these different function vectors convey different information, we next examine to what extent they share mechanisms within the model.

### 3.3 Instruction and demonstration FVs mostly engage different attention heads

In Figure 3A-D, we visualize the  $|\mathcal{A}| = 20$  top attention heads identified in each model by the FV extraction procedure. We observe that **demonstrations and instructions elicit mostly distinct sets of heads**. Of the 20 top heads identified, instruction and demonstration FVs share few of them – 7 in both Llama-3.2-3B models and only 4 in the Llama-3.1-8B ones. Post-training appears to move instruction FV heads closer to demonstration FV heads. The most notable change between base and post-trained model versions is in the average layer of the instruction-only heads — from a mean several layers deeper than demonstrations in base models, to almost identical depths in post-trained ones (other model classes are similar, see Appendix H). We observe higher causal scores in demonstration top heads. In §2, we define the causal score of each head as its contribution to correctly predicting the next token. Demonstration FV top heads receive substantially higher scores than instruction ones (Table 7). This suggests that task inference from demonstrations is more localized to a small set of heads than from instructions and that instruction-based task inference is more diffuse.

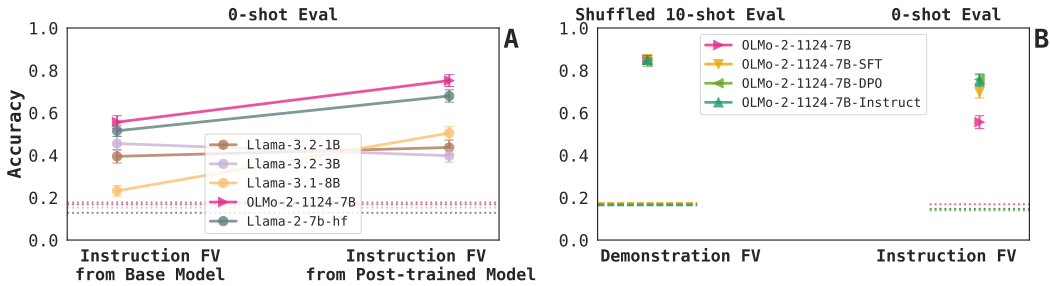
We also examine the mean task-conditioned activation patterns in the top heads shared between instructions and demonstrations. We compute the cosine similarities between the patterns elicited by demonstrations vs. by shorter/longer instructions (see Figure 3E-F). Expectedly, we find higher similarities between the two instruction-driven activation patterns than between the instruction and demonstration patterns. Additionally, later layers are generally more similar. Finally, post-training appears *not* to adapt demonstration FV heads. We observe no difference in demonstration-instruction similarities between base and post-trained models, and equal numbers of shared heads between instructions and demonstrations across model versions. Instead, post-training appears to induce a different, separate mechanism for instruction task inference, one we hope to explore in future work.

### 3.4 Instruction-identified attention heads are more useful for building demonstration FVs than vice versa

If demonstrations and instructions share a common task inference mechanism, we would not observe a significant difference if we constructed ‘incongruent’ function vectors using top heads identified by one with activations from the other. We borrow from neuroimaging and consider the FV extraction procedure through the lens of functional localizers, viewing head identification as localization and mean activation computation as recording (Saxe et al., 2006; Berman et al., 2010). We evaluate both types of ‘incongruent’ FVs, those constructed with demonstration top heads and those using instruction top heads, and summarize our results in Figure 3G-H. We observe that, as expected, **accuracy with these incongruent FVs falls below the regular FVs in all cases**. We note an asymmetry: **using instruction-localized top heads from with demonstration mean activations seems preferable to the opposite combination** (with the exception of the shuffled 10-shot evaluation on the base models). This effect is sharpened when we evaluate each FV in the layer resulting in the highest accuracy, rather than fixing to  $\lfloor L/3 \rfloor$  (Figure 28). To explain this, we report the causal indirect effects of the top heads from each task presentation in the opposite presentation format – that is, the scores of instruction-localized heads when using demonstrations, and vice versa (Table 8). Examining the scores, we observe that instruction-localized heads are more helpful in the demonstration setting than demonstration-localized heads are in the instruction setting. We take this as evidence that the mechanism for task inference from instruction leverages attention heads that play a minor role in demonstration task inference. The development of such a mechanism may also explain the depth alignment shifts we previously observed between base and post-trained models (subsection 3.3, Figure 3A-D). In contrast, the demonstration task inference mechanism uses attention heads that are less useful for instruction task inference. This supports our earlier claim about the primacy of demonstration ICL. We conclude with two control experiments to ensure that this observed asymmetry is meaningful (and not an artifact of selecting arbitrary sets of heads). We select sets of heads that are either unrelated to both FV types, or that have the *lowest* causal scores (Appendix J.1). In both cases, accuracy falls to task baseline or below, suggesting the observed effect is meaningful.

### 3.5 Instruction FVs from post-trained can steer base models, arise from SFT and DPO

Given the effectiveness of instruction FVs in post-trained models, we examine whether instruction FVs from post-trained models can steer base models. We repeat the instruction FV evaluations, intervening on each base model with the FV generated by its post-trained version (multiple post-trained variants exist for OLMo-2-1124-7B; we use the final model, OLMo-2-1124-7B-Instruct). We report our results in Figure 4B. In three of the base models, **we find substantial accuracy increases, nearing the zero-shot accuracy elicited by instruction FVs in the post-trained models** (consistent with Stolfo et al.’s (2024) cross-model steering results). This effect is even more striking when



**Figure 4: Instruction-based function vectors can transfer from post-trained to base models and are facilitated by SFT and DPO.** (A) We apply the instruction FVs extracted from post-trained models to their respective base models. We find weak impacts on the distilled Llama-3.2 models, and substantial benefits in the other ones, almost recovering the post-trained model FV evaluation accuracy. (B) We examine the OLMo-2-1124-7B family of models and observe meaningful increases in the efficacy of instruction-based function vectors in both the SFT and DPO stages over the base model; conversely, the final RL stage appears minimally impactful. In both panels, dotted lines represent baselines with no FV intervention, and error bars reflect standard errors of the mean (SEMs). See Figure 33 for results evaluating the function vectors in the setting opposite to their extraction, and Figure 34 for results when intervening at optimal depths, rather than a fixed  $\lfloor L/3 \rfloor$  depth.

evaluating the steering instruction FVs in the shuffled 10-shot setting (Figure 33A). We take this as further evidence that this procedure identifies the task representations elicited by instructions. We find minimal or mildly negative effects on the Llama-3.2 models.

We conclude by examining which post-training stages contribute to the ability to extract instruction FVs using the OLMo-2-1124-7B family of models (OLMo et al., 2024). The authors release four versions of the model: a base version, one following supervised fine-tuning (-SFT), another following preference fine-tuning (-DPO), and a final version (-Instruct). Figure 4B depicts results with these four models. All OLMo models are conducive to extracting function vectors from demonstrations, but the base model instruction FVs show lower accuracies. We observe two accuracy increases, once with the SFT model and another with the DPO model. The DPO model also offers an increase in the efficacy of instruction FVs in mismatched evaluations (see Figure 33B). Finally, we note that post-training appears to have no impact on demonstration-based ICL, as evidenced by the lack of variability in Figure 4B (left).

## 4 Related Work

**In-Context Learning.** The ability to perform tasks from demonstrations has been in the foreground of language model research following its identification by Brown et al. (2020). Substantial subsequent research has studied LLMs’ ability to perform in-context learning, from perspectives such as the data distribution (Chan et al., 2022; Chen et al., 2024), learning algorithms (Xie et al., 2021; Akyürek et al., 2022, 2024), mechanics and learning dynamics (Von Oswald et al., 2023; Zhang et al., 2023; Park et al., 2024; Han et al., 2024; Li et al., 2025), and latent representations (Pan et al., 2023; Todd et al., 2024; Hendel et al., 2023; Yin and Steinhardt, 2025). Other lines of work examined factors to ICL contributing (Min et al., 2022), such as sensitivity to in-context prompt formatting choices (Sclar et al., 2023; Su et al., 2025). Most prior work focuses on ICL from demonstrations; our work bridges between those and instructions as a gateway to exploring other forms of ICL (Lampinen et al., 2024).

**Instruction-following** refers to executing a task given a natural language, similar to how one person may instruct another. Although fine-tuning is necessary to elicit this capability, it has been shown to facilitate generalization and usefulness following the introduction of the Flan models (Chung et al., 2022). Instruction-following has since been the focus of dataset development (Mishra et al., 2022; Wang et al., 2022a,b; Zhou et al., 2023a; Taori et al., 2023), model evaluation (Zhou et al., 2023b; Liu et al., 2024; Lyu et al., 2024), and interpretability work (Stolfo et al., 2024; Wu et al., 2024). Our work explores whether instruction tuning leverages the mechanisms of demonstration-based ICL.

**Task Representations.** Computational neuroscience has long sought to understand how neural networks learn to represent abstract tasks (Cohen et al., 1990; Botvinick and Plaut, 2002; Yang et al., 2019; Flesch et al., 2021; Farrell et al., 2023; Hummos et al., 2024). In tandem, computer scientists devised methods to explicitly introduce these representations to models (Lampinen and McClelland, 2020; Ilharco et al., 2022; Shao et al., 2022). Our work follows a recent line of inquiry extracting task representations that arise in LLMs (Todd et al., 2024; Hendel et al., 2023; Saglam et al., 2025) or VLMs (Luo et al., 2024; Hojel et al., 2024; Huang et al., 2024) as they complete tasks. Most task representation extraction approaches fall into one of a few categories. Some treat the model’s latent (residual) representation of a delimiter token at some depth as the task representation (e.g., Hendel et al., 2023; Luo et al., 2024; Li et al., 2025). Another approach learns a per-layer task representation (Saglam et al., 2025). Finally, Todd et al. (2024) combine activity from multiple attention heads across different depths and intervene with the resultant function vector at a single depth. We focus on contrasting the representations that arise from different textual presentations of the same task.

## 5 Discussion

We empirically test for the presence of common, presentation-agnostic task representations in large language models. Our results suggest that different presentations of a task *do not* elicit a common task representation through function vectors. We extend the FV extraction procedure from in-context demonstrations to arbitrary task presentations and successfully construct instruction FVs that promote zero-shot accuracy. We find these convey different information than demonstration FVs, evidenced by the benefits of intervening with both forms of function vectors together over either by itself. We also find that mostly distinct (though partly overlapping) sets of attention heads causally mediate task



performance in these two settings. Our results offer support (in a limited, but controlled setting) for the widely used practice of prompting language models with both demonstrations and instructions. Our findings support two other takeaways. First, ICL from demonstrations may be more inherent to LLMs. We offer preliminary evidence that instruction task inference leverages attention heads that are peripherally useful for demonstration-based ICL. The contrary effect is weaker; demonstration FV attention heads are less helpful for instructions. Second, we show evidence that instruction FVs transfer from post-trained models to steer their base versions better than base model instruction FVs.

We found the effectiveness of instruction-derived FVs surprising. Given the diversity of plausible instructions for a given task, we would have considered it a priori likely that instructions induced relatively diffuse representations that would not support FV extraction from a few attention heads. Indeed, the representation of textual instructions appears more diffuse in language models, as evident from the distributions of attention head causal scores (Table 7); yet, our approach is successful.

The accuracy discrepancy when function vectors are evaluated in a setting incongruent with their extraction further supports the notion that task representations are dependent on the form of presentation (compare Figure 2 and Figure 12). This evaluation form-dependency portends a challenge in identifying, monitoring, or amplifying task representations in naturalistic settings, such as prompts combining multiple demonstration forms, obfuscated or adversarially-presented tasks, or other more benign presentations forms, such as role prompts or explanations (Lampinen et al., 2024).

In our results, we average over the three uninformative baselines we propose: sampling tokens directly from the model, natural texts, and instructions from other tasks. While the selected top heads did not greatly differ between the three, one is unlike the other two (Appendix H.3). Instructions from other tasks still convey the existence of a task, just a different one, and may help localize task-specific information. Conversely, tokens sampled from the model and natural texts do not convey ‘task-ness,’ hence helping to localize the existence of a task in general. The differences between attention heads identified with each baseline alone may elucidate the differing roles of instruction FV attention heads.

We identify which post-training stages contribute to the ability to extract instruction FVs. However, our results do not explain *what* takes place in post-training that facilitates inferring tasks from instructions. Wu et al. (2024) propose that the primary drivers could be changes in processing words that convey instructions—which suggests the hypothesis that the instruction FV attention heads should strongly respond to these tokens. In another recent work, Yin and Steinhardt (2025) compare induction heads (Elhage et al., 2021) and FV heads—do instruction FV heads share anything in common with induction heads as well? On that note, what can we learn from the heads that change in importance between base and post-trained models? Our results point a spotlight at a set of heads relevant to this capacity; future work may study their contributions and mechanisms in greater depth.

Our finding that we can steer base models with post-trained models instruction FV is consistent with recent findings by Stolfo et al. (2024) on activation steering in an instruction-following context. Our current results do not suffice to offer a potential savings in the effort required to post-train a model, as so far, we only demonstrate the ability to steer a base model with its own post-trained variation. Recent results by Lee et al. (2025) propose a method for transferring steering vectors between models — combined with our results, this suggests a potential to confer some of the benefits of post-training on a model that has not been post-trained, a connection we hope to explore in future work.

Finally, we highlight a few other proposals on how our work might inform prompt and intervention design. Our finding regarding the diffuseness of instruction FVs (subsection 3.3) implies that interventions focused on changing or improving instruction-following may benefit from intervening at multiple locations and layers; presently, most steering methods perform a single additive residual stream intervention, which may be more meaningfully suboptimal for instructions, than for example, for demonstrations. The observation that demonstrations and instructions elicit different FV attention heads (also in subsection 3.3) suggests a speculative potential pathway to improving prompt design. In contexts where both instructions and demonstrations are provided, it may be possible to monitor their separate sets of implicated attention heads, and try to classify from their activity whether or not the model successfully formed a task representation. This classifier could serve as a signal that helps guide toward improving the instructions or selecting different or additional demonstrations.

## 5.1 Limitations

**Task set.** We use a limited set of minimal, fairly simple tasks, following Todd et al.’s (2024) choices. We omit a few classification tasks where predicting the correct token requires understanding the expected answer format, information conveyed clearly by demonstrations, but not necessarily by instructions. This could be addressed by combining demonstrations and instructions, or by encouraging the instruction-generating model to include output formatting information. Furthermore, it would be of interest to examine how these findings generalize to longer, more complex, open-ended, and naturalistic tasks, such as the ones present in MMLU (Hendrycks et al., 2020) or BBH (Srivastava et al., 2022; Suzgun et al., 2022).

**Choice of task representation.** We study a particular proposal for task representations — function vectors as formulated by Todd et al. (2024). Other task representation extraction approaches exist in the literature, such as task vectors (Hendel et al., 2023) — we hope future work examines whether our findings translate to their formulation as well. More broadly, we cannot conclusively prove the negative, that common task representations do not exist; we can only demonstrate that our methods did not identify them. We also do not explore the working of attention heads identified in greater depth, such as identifying which circuits (Elhage et al., 2021) they may contribute to.

**Arbitrary intervention depth.** We follow Todd et al. (2024) in intervening at a fixed  $|L/3|$  depth. While we also present (in the appendices) results at empirically optimal intervention depths for each model, the choice retains a measure of arbitrariness. It is possible, for instance that different types of tasks are better served better by different intervention depths, and our analyses do not reflect that.

**Model sizes.** We focus our investigation on smaller (large) language models. While we offer fairly consistent evidence across the models we examine, we do not explore scaling with model size. We note that Todd et al. (2024) successfully extracted demonstration function vectors to Llama-2 models at the 7B, 13B, and 70B sizes, which offers promising evidence for scaling function vectors.

**Limited replications.** We evaluate many experimental conditions (Appendix E.1). For each model we study, we evaluate each of the approximately 50 tasks six times — shorter and longer instructions, each with each of the three baselines. We also ran the additional control conditions reported with all tasks for the Llama-3.2-3B and Llama-3.1-8B models. However, we only replicate each model on each dataset in each setting once—given the qualitatively consistent results observed by Todd et al. (2024), replicating over random seeds (used primarily in train-test splits and query example sampling) seemed untenable, if not wasteful. We use a single prompt template (Figure 9), relying on consistent previous results with different prompt templates (p. 6 and Appendix C in Todd et al., 2024).

## Acknowledgments

We thank members of the Computation and Cognition lab at NYU and the Human and Machine Intelligence lab at Princeton for valuable feedback at various stages of this work. We thank Eric Todd, Eshika Saxena, and Nicola Cancedda for feedback on this project and manuscript. Author GD was funded by the Meta AI Mentorship program during his work on this manuscript.

## References

- Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2022). What learning algorithm is in-context learning? investigations with linear models. *arXiv [cs.LG]*.
- Akyürek, E., Wang, B., Kim, Y., and Andreas, J. (2024). In-context language learning: Architectures and algorithms. *arXiv [cs.CL]*.
- Berman, M. G., Park, J., Gonzalez, R., Polk, T. A., Gehrke, A., Knaffla, S., and Jonides, J. (2010). Evaluating functional localizers: the case of the FFA. *Neuroimage*, 50(1):56–71.
- Botvinick, M. and Plaut, D. C. (2002). Representing task context: proposals based on a connectionist model of action. *Psychol. Res.*, 66(4):298–311.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C.,

- McCandlish, S., Radford, A., Sutskever, I., and Amodi, D. (2020). Language models are few-shot learners. *arXiv [cs.CL]*.
- Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A. K., Richemond, P. H., McClelland, J., and Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. *Neural Inf Process Syst*, abs/2205.05055:18878–18891.
- Chen, Y., Zhao, C., Yu, Z., McKeown, K., and He, H. (2024). Parallel structures in pre-training data yield in-context learning. *arXiv [cs.CL]*.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv [cs.LG]*.
- Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychol. Rev.*, 97(3):332–361.
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv [cs.CL]*.
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodi, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for transformer circuits. <https://transformer-circuits.pub/2021/framework/index.html>. Accessed: 2025-4-30.
- Farrell, M., Recanatesi, S., and Shea-Brown, E. (2023). From lazy to rich to exclusive task representations in neural networks and neural codes. *Curr. Opin. Neurobiol.*, 83(102780):102780.
- Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., and Summerfield, C. (2021). Rich and lazy learning of task representations in brains and neural networks. *bioRxiv*, page 2021.04.23.441128.
- Han, S., Song, J., Gore, J., and Agrawal, P. (2024). Emergence and effectiveness of task vectors in in-context learning: An encoder decoder perspective. *arXiv [cs.CL]*.
- Hendel, R., Geva, M., and Globerson, A. (2023). In-context learning creates task vectors. *arXiv [cs.CL]*.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv [cs.CY]*.
- Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. (2023). Linearity of relation decoding in transformer language models. *arXiv [cs.CL]*.
- Hojel, A., Bai, Y., Darrell, T., Globerson, A., and Bar, A. (2024). Finding visual task vectors. *arXiv [cs.CV]*.
- Huang, B., Mitra, C., Arbelle, A., Karlinsky, L., Darrell, T., and Herzig, R. (2024). Multimodal task vectors enable many-shot multimodal in-context learning. *arXiv [cs.CV]*.
- Hummos, A., del Río, F., Wang, B. M., Hurtado, J., Calderon, C. B., and Yang, G. R. (2024). Gradient-based inference of abstract task representations for generalization in neural networks. *arXiv [cs.LG]*.
- Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. (2022). Editing models with task arithmetic. *arXiv [cs.LG]*.
- Lampinen, A. K., Chan, S. C. Y., Singh, A. K., and Shanahan, M. (2024). The broader spectrum of in-context learning. *arXiv [cs.CL]*.
- Lampinen, A. K. and McClelland, J. L. (2020). Transforming task representations to perform novel tasks. *Proc. Natl. Acad. Sci. U. S. A.*, 117(52):32970–32981.
- Lee, A., Weber, M., Viégas, F., and Wattenberg, M. (2025). Shared global and local geometry of language model embeddings. *arXiv [cs.CL]*.
- Li, Y., Campbell, D., Chan, S. C. Y., and Lampinen, A. K. (2025). Just-in-time and distributed task representations in language models. *arXiv [cs.CL]*.
- Liu, Y., Shi, K., Fabbri, A. R., Zhao, Y., Wang, P., Wu, C.-S., Joty, S., and Cohan, A. (2024). ReIFE: Re-evaluating instruction-following evaluation. *arXiv [cs.CL]*.

- Llama Team, A. I. . M. (2024). The llama 3 herd of models. *arXiv [cs.AI]*.
- Luo, G., Darrell, T., and Bar, A. (2024). Vision-language models create cross-modal task representations. *arXiv [cs.CV]*.
- Lyu, X., Wang, Y., Hajishirzi, H., and Dasigi, P. (2024). HREF: Human response-guided evaluation of instruction following in language models. *arXiv [cs.CL]*.
- Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *arXiv [cs.CL]*.
- Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work?
- Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. (2022). Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.
- Nguyen, K. A., Schulte im Walde, S., and Vu, N. T. (2017). Distinguishing antonyms and synonyms in a pattern-based neural network. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.
- OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Guerin, M., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., and Hajishirzi, H. (2024). 2 OLMo 2 furious. *arXiv [cs.CL]*.
- Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2022). In-context learning and induction heads. *arXiv [cs.LG]*.
- Pan, J., Gao, T., Chen, H., and Chen, D. (2023). What in-context learning “learns” in-context: Disentangling task recognition and task learning. *arXiv [cs.CL]*.
- Park, C. F., Lubana, E. S., Pres, I., and Tanaka, H. (2024). Competition dynamics shape algorithmic phases of in-context learning. *arXiv [cs.LG]*.
- Saglam, B., Yang, Z., Kalogerias, D., and Karbasi, A. (2025). Learning task representations from in-context learning. *arXiv [cs.CL]*.
- Saxe, R., Brett, M., and Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4):1088–96; discussion 1097–9.
- Scarl, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2023). Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting.
- Shao, N., Cai, Z., Xu, H., Liao, C., Zheng, Y., and Yang, Z. (2022). Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.
- Srivastava, A., Rastogi, A., Rao, A., Shueb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., and Nushi, B. (2024). Improving instruction-following in language models through activation steering. *arXiv [cs.CL]*.
- Su, J., Zhang, J., Ullrich, K., Bottou, L., and Ibrahim, M. (2025). A single character can make or break your LLM evals. *arXiv [cs.CL]*.
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., , and Wei, J. (2022). Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).

- Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. (2024). Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardaş, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*.
- Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2022a). Self-instruct: Aligning language models with self-generated instructions. *arXiv [cs.CL]*.
- Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. (2022b). Super-naturalinstructions: Generalization via declarative instructions on 1600+ tasks. In *EMNLP*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2019). HuggingFace’s transformers: State-of-the-art natural language processing. *arXiv [cs.CL]*.
- Wu, X., Yao, W., Chen, J., Pan, X., Wang, X., Liu, N., and Yu, D. (2024). From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2341–2369, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. *arXiv [cs.CL]*.
- Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.*, 22(2):297–306.
- Yin, K. and Steinhardt, J. (2025). Which attention heads matter for in-context learning? *arXiv [cs.LG]*.
- Zhang, R., Frei, S., and Bartlett, P. (2023). Trained transformers learn linear models in-context. *J. Mach. Learn. Res.*, 25(49):49:1–49:55.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023a). LIMA: Less is more for alignment. *arXiv [cs.CL]*.
- Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. (2023b). Instruction-following evaluation for large language models. *arXiv [cs.CL]*.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: We are careful to limit the scope of claims in the abstract and introduction to the experiments and evidence in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of our work in subsection 5.1.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not provide any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide thorough descriptions of the methods we use section 2 and Appendix B, example prompt templates in Appendix A, a list of the models we evaluate in Table 4, and a list of the datasets we use in Appendix D. We will also release the code, and all the data we use is available as we reuse the datasets provided by Todd et al. (2024).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: We provide a zip archive of our code with instructions for how to set up the environment and example commands to launch our experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.



- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide experimental details in Appendix E and the code to run our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We report standard errors of the mean (SEMs) on all figures in the paper that estimate average performance in some condition across a set of conditions. We do not report statistical significance results as it did not seem material for the comparisons we make, but are open to doing so should the reviewers request.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We attempt to ballpark the number of experiment runs required to generate our results (and hence, reproduce our work) in Appendix E.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our work conforms with the NeurIPS Code of Ethics in its entirety.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We do not foresee direct societal impacts of the work we performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We provide no data or models that have a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and refer to the original creators of code, data, and models we used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We document the code we provide in the supplementary materials, and will further improve the documentation before we open-source code when the work is de-anonymized.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We report no crowdsourcing or human experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We report no human subjects research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our work describes an evaluation of LLMs across a variety of tasks in attempt to extract task representations. We also used LLMs to generate tests for researcher-written code and autocomplete using GitHub Copilot.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Prompt generation prompt templates

```
<|start_header_id|>user<|end_header_id|>
Instructions
You are powerful model helping write prompts to help smaller models perform tasks better.
Below, you will be given a set of input-output pairs for a particular undescribed task. First,
please study the examples to deduce what the task is, and describe your thinking under the
header "# Task Deduction". Next, please write 10 prompts that might help a smaller model
perform this task. The prompts should be:
    1. Short, up to 10 words.
    2. Informative about what the task is.
    3. Not repetitive with each other.
Please write your prompts under the header "# Task Prompts".

Task examples
{task examples}

Now, think step by step and follow the instructions above.
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

**Figure 5: Prompt for generating short task instructions given ICL examples.** The prompted model (in our experiments, Llama-3.1-405B) must analyze the examples given under the header “task examples” and generate instructions for another model. We provide no information beyond the in-context examples, and repeat this procedure 20 times to generate a set of approximately 200 instructions for each task, which we then deduplicate. In practice, the vast majority of these short instructions tokenized to 16 or fewer tokens, so we set that as the limit for what we considered a short instruction.

```
<|start_header_id|>user<|end_header_id|>
Instructions
You are powerful model helping write prompts to help smaller models perform tasks better.
Below, you will be given a set of input-output pairs for a particular undescribed task. First,
please study the examples to deduce what the task is, and describe your thinking under the
header "# Task Deduction". Next, please write 10 prompts that might help a smaller model
perform this task. The prompts should be:
    1. As long as necessary to be helpful for the smaller model.
    2. Informative about what the task is.
    3. Not repetitive with each other.
    4. Not including any examples of the task.
Please write your prompts under the header "# Task Prompts".

Task examples
{task examples}

Now, think step by step and follow the instructions above.
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

**Figure 6: Prompt for generating longer task instructions given ICL examples.** The prompted model (in our experiments, Llama-3.1-405B) must analyze the examples given under the header “task examples” and generate instructions for another model. We provide no information beyond the in-context examples, and repeat this procedure 20 times to generate a set of approximately 200 instructions for each task, which we then deduplicate. In practice, the vast majority of these longer instructions tokenized to 64 or fewer tokens, so we set that as the limit for what we considered a long instruction.

### A.1 Example generated instruction prompts

**Table 2: Example short instructions for five tasks.**

<b>Task</b>	<b>Short Instructions</b>
<b>antonym</b>	Create an opposing term Identify the antithesis of this word Create a counter-term Reverse the semantic meaning Provide a word that is the semantic opposite
<b>country-capital</b>	Country to capital city correlation Learn country-capital associations Map country names to their capitals Identify the administrative center Provide the capital city for the given country
<b>concept_v_object_5</b>	Select the word that is not a noun "Find the word that is not a concrete object." Select the word that tells us more about something Which word has a distinct semantic meaning? Identify the adverb or adjective in the list
<b>english-spanish</b>	Spanish equivalent for this English term Translate everyday English words to Spanish Spanish translation of English word Find Spanish counterpart for English word Find Spanish translation
<b>product-company</b>	"Associate product name with company name" Which company created this software? "Classify product by owner company" Identify the company that developed this technology "Link this device to its manufacturer."

**Table 3: Example long instructions for five tasks.**

<b>Task</b>	<b>Long Instructions</b>
<b>antonym</b>	<p>Find a word that, when compared to the input word, presents a contrasting meaning. This word should highlight the differences and serve as an antonym</p> <p>Generate a word that cancels out the meaning of the input word</p> <p><b>**Meaning reversal**</b>: Reverse the meaning of the input word by generating a word that represents its opposite. Ensure that the generated word is semantically accurate and contextually relevant</p> <p>This task tests the ability to navigate the vocabulary of a language to find and generate antonyms. Please focus on producing words that are directly opposite or clearly contrasting</p> <p><b>**Find a word that contrasts with the input word in meaning.**</b> This could involve finding a word that is the opposite of the input word or one that describes a different extreme or end of a spectrum</p>
<b>country-capital</b>	<p>What is the name of the city where a country’s president or monarch typically resides and conducts official business?</p> <p>Determine the capital city of a country by identifying the city where the national government is seated and where major political decisions are made</p> <p>Provide the name of the city that is generally accepted as the capital of a particular country</p> <p>What city is recognized as the center of administration and governance for a given country?</p> <p>"Countries around the world each have a capital city where their government is based. Your task is to know what these cities are for any country you are asked about."</p>
<b>concept_v_object_5</b>	<p>Determine the word in the list that is a verb or an action</p> <p>Identify the word in the list that describes a quality, property, or characteristic of something</p> <p>Identify the word in the list that describes a quality or property of something</p> <p><b>**Determine the Quality Word**</b>: Determine which word from the list describes a quality, state, or condition. This word should tell us about the nature or attributes of something</p> <p>Find the word that can be used in a sentence to describe an action, event, or situation</p>
<b>english-spanish</b>	<p>Translate the English word into Spanish, making sure to use the most appropriate and commonly used term in Spanish-speaking contexts</p> <p>Provide a Spanish translation of the input word that is both accurate and fluent</p> <p>Translate the input word from English to Spanish, considering any relevant context or connotations</p> <p>Translate the given English word into its equivalent in Spanish, ensuring to maintain the original meaning and word type (noun, verb, adjective, etc.)</p> <p>Identify the Spanish equivalent of the provided English term, ensuring the translation is accurate and suitable for the context</p>
<b>product-company</b>	<p>Identify the developer of a given operating system, platform, or tool</p> <p>Given the name of a product, technology, or format, find the company that owns or developed it. Use your knowledge of industry leaders and their offerings</p> <p>Identify the company or organization that developed or owns the product, technology, or format specified in the input</p> <p>Identify the company that created this file format</p> <p>Determine the company that is associated with the specified brand, product, or format</p>

## B Uninformative instruction baselines

### B.1 Additional uninformative instruction details

**Equiprobable token sequences.** Given task-informative instructions  $q_m^t$ , our goal is to sample an uninformative instruction  $\tilde{q}_m^t$ . We encode  $q_m^t$  as a token sequence  $q_m^t = [w_1, w_2, \dots, w_L]$  whose probability under the model  $f$  is  $P(q_m^t) = \prod_{l \leq L} P(w_l | w_{<l}) = f(w_{<l})[w_l]$ . We begin from the BOS token, and sequentially sample  $\tilde{w}_l$  to create  $\tilde{q}_m^t = [\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_L]$ . We do so by approximately matching the conditional probabilities  $f(\tilde{w}_{<l})[\tilde{w}_l] \approx f(w_{<l})[w_l]$  using the following logic. First, we mask out any non-text (“added vocabulary”) tokens. At each step, we compute the (log-) probability of the  $l$ ’th token,  $\log P(w_l | w_{<l})$ . We compute the distribution over next tokens of the uninformative instructions,  $\log P(\cdot | \tilde{w}_{<l})$ . We set an initial threshold and a threshold increment (in our experiments, both were set to 0.1). We increase the threshold by the increment until at least one token in  $\log P(\cdot | \tilde{w}_{<l})$  falls within the incremented threshold of the log-probability of the current instruction token. Denote by  $t$  our initial threshold,  $\Delta t$  our increment, and  $k \in \mathbb{Z}^*$  the number of increments required, we find the following:  $\min_k : \left( \sum_{\tilde{w}_l'} \mathbb{1}[t - k\Delta t \leq \log |P(w_l | w_{<l}) - \log P(\tilde{w}_l' | \tilde{w}_{<l})| \leq t + k\Delta t] \right) > 0$ , where  $\mathbb{1}[\cdot]$  denotes the indicator function. We then sample uniformly between all  $\tilde{w}_l'$  satisfying the previous relation and append the sampled  $\tilde{w}_l$  to the  $\tilde{q}_m^t$  we are constructing, continuing until we reach the same length as the informative  $q_m^t$ .

**Real texts.** Ahead of time, we precompute the log-probability of texts from WikiText-103-v1 (Merity et al., 2016) with each model  $f$ . We randomly sample entries from the WikiText-103-v1 dataset. For each sampled entry, we extract overlapping prefixes that end with whitespace (excluding the terminal whitespaces), tokenize these prefixes, and keep those with a length of 64 or fewer tokens. We extract prefixes to create strings that might coherently appear at the beginning of a text. We then compute and cache the (log-) probability of each such prefix and its token sequence. We cache approximately  $2^{16} = 65536$  token sequences with each model (approximately as we stop after the Wikitext entry that brought us over the threshold, but do not discard sequences beyond the  $2^{16}$ ’th one).

To sample an uninformative  $\tilde{q}_m^t$  for some  $q_m^t$ , we tokenize  $q_m^t$  to arrive at its length  $L(q_m^t)$  and compute its log-probability  $\log f(q_m^t)$ . We now create a representative sample of  $N = 100$  cached texts with approximately the same length as  $L(q_m^t)$ . We begin by only consider texts of with a length of precisely  $L(q_m^t)$  tokens. If there are over  $N$  of those, we stop; otherwise, for  $k \in \mathbb{Z}^+$ , we also consider texts of length  $L(q_m^t) \pm k$ , increasing  $k$  by one until we have a set of at least  $N$  candidates. Once we have attained this set of candidates, we select the  $\tilde{q}_m^t$  with the closest log-probability under  $f$  as  $q_m^t$  has, and remove it from the set. As we require five uninformative baselines for each instruction  $q_m^t$  (see Appendix E), we will use the five real texts with the closest log-probability to  $q_m^t$  subject to being within a small number of tokens from  $L(q_m^t)$ .

**Other task instructions.** We follow a conceptually similar procedure to the one described above for real texts, but using instructions generated for other tasks. As this is a smaller set, we do not cache these log probabilities in advance. We exclude instructions generated for the same task, but otherwise follow an identical procedure. Denote by  $L(q_m^t)$  the length of the instructions  $q_m^t$ . We identify a set of candidate alternative instructions with a length of approximately  $L(q_m^t)$ , increasing the acceptable difference in length until we reach  $N = 100$  candidates. We then return the one with the closest log-probability to  $q_m^t$  under the model and remove it from the set, so as before, we will use the five other task instructions with the closest log-probability to  $q_m^t$  subject to being within a small number of tokens from  $L(q_m^t)$ .

### B.2 Prompt baseline examples





**Table 4: Models studied.** We use the Huggingface Transformers (Wolf et al., 2019) model implementations.

Model	Citation	Huggingface ID	$ L $	$ a_t $
Llama-3.2-1B	Llama Team (2024)	meta-llama/Llama-3.2-1B	16	32
Llama-3.2-1B-Instruct	Llama Team (2024)	meta-llama/Llama-3.2-1B-Instruct	16	32
Llama-3.2-3B	Llama Team (2024)	meta-llama/Llama-3.2-3B	28	24
Llama-3.2-3B-Instruct	Llama Team (2024)	meta-llama/Llama-3.2-3B-Instruct	28	24
Llama-3.1-8B	Llama Team (2024)	meta-llama/Llama-3.1-8B	32	32
Llama-3.1-8B-Instruct	Llama Team (2024)	meta-llama/Llama-3.1-8B-Instruct	32	32
OLMo-2-1124-7B	OLMo et al. (2024)	allenai/OLMo-2-1124-7B	32	32
OLMo-2-1124-7B-SFT	OLMo et al. (2024)	allenai/OLMo-2-1124-7B-SFT	32	32
OLMo-2-1124-7B-DPO	OLMo et al. (2024)	allenai/OLMo-2-1124-7B-DPO	32	32
OLMo-2-1124-7B-Instruct	OLMo et al. (2024)	allenai/OLMo-2-1124-7B-Instruct	32	32
Llama-2-7b	Touvron et al. (2023)	meta-llama/Llama-2-7b-hf	32	32
Llama-2-7b-chat	Touvron et al. (2023)	meta-llama/Llama-2-7b-chat-hf	32	32

## D Full list of tasks

**Table 5: Tasks used.**

Task	Citation
adjective_v_verb_3	Todd et al. (2024)
adjective_v_verb_5	Todd et al. (2024)
alphabetically_first_3	Todd et al. (2024)
alphabetically_first_5	Todd et al. (2024)
alphabetically_last_3	Todd et al. (2024)
alphabetically_last_5	Todd et al. (2024)
animal_v_object_3	Todd et al. (2024)
animal_v_object_5	Todd et al. (2024)
antonym	Nguyen et al. (2017)
capitalize	Todd et al. (2024)
capitalize_first_letter	Todd et al. (2024)
capitalize_last_letter	Yin and Steinhardt (2025)
capitalize_second_letter	Yin and Steinhardt (2025)
choose_first_of_3	Todd et al. (2024)
choose_first_of_5	Todd et al. (2024)
choose_last_of_3	Todd et al. (2024)
choose_last_of_5	Todd et al. (2024)
choose_middle_of_3	Todd et al. (2024)
choose_middle_of_5	Todd et al. (2024)
color_v_animal_3	Todd et al. (2024)
color_v_animal_5	Todd et al. (2024)
concept_v_object_3	Todd et al. (2024)
concept_v_object_5	Todd et al. (2024)
conll2003_location	Tjong Kim Sang and De Meulder (2003)
conll2003_organization	Tjong Kim Sang and De Meulder (2003)
conll2003_person	Tjong Kim Sang and De Meulder (2003)
country-capital	Todd et al. (2024)
country-currency	Todd et al. (2024)
english-french	Conneau et al. (2017)
english-german	Conneau et al. (2017)
english-spanish	Conneau et al. (2017)
fruit_v_animal_3	Todd et al. (2024)
fruit_v_animal_5	Todd et al. (2024)
landmark-country	Hernandez et al. (2023)
lowercase_first_letter	Todd et al. (2024)
lowercase_last_letter	Todd et al. (2024)
national_parks	Todd et al. (2024)
next_capital_letter	Todd et al. (2024)
next_item	Todd et al. (2024)
object_v_concept_3	Todd et al. (2024)
object_v_concept_5	Todd et al. (2024)
park-country	Todd et al. (2024)
present-past	Todd et al. (2024)
prev_item	Todd et al. (2024)
product-company	Hernandez et al. (2023)
singular-plural	Todd et al. (2024)
synonym	Nguyen et al. (2017)
verb_v_adjective_3	Todd et al. (2024)
verb_v_adjective_5	Todd et al. (2024)
word_length	Todd et al. (2024)

All tasks used were sourced from Todd et al.’s (2024) repository:  
[https://github.com/ericwtodd/function\\_vectors](https://github.com/ericwtodd/function_vectors).

We omitted the following tasks, as they are classification tasks with specific output formats that the model-generated instructions often did not specify: ag\_news, commonsense\_qa, person-instrument, person-occupation, person-sport, sentiment

## E Full experimental settings

We detail our experimental settings to aid reproducibility. As a guideline we strive to match or minimally adapt decisions made by Todd et al. (2024):

- For each model and each task, we use the  $J = 5$  instructions with the highest accuracy over the training split.
- We compute the mean activations over 100 total prompts, 20 with each of the 5 best instructions.
- We compute the causal indirect effects over 25 total uninformative prompts, 5 generated for each of the best instructions.
- We batch our results with a batch size that depends on the model and task, but does not exceed 5 for any model or task (see code for batch size computation logic).
- We split each dataset 70% to train and 30% to test. Where we require a validation set, we split it again from the training set.
- We load all models in full precision.
- We use the  $|\mathcal{A}| = 20$  top heads in all experiments we report.
- We evaluate the FV interventions at every possible depth (that is, after every layer of the model).
  - In all main manuscript figures, we report the accuracy intervening after the  $\lfloor L/3 \rfloor$  layer (layer 9 for the Llama-3.2-3B models and layer 11 for the Llama-3.1-8B and OLMo-2-1123-7B ones).
  - In appendix figures that report the empirically optimal intervention layer, we compute the accuracy using the layer that would result in the highest mean accuracy for each model, averaging over both the 0-shot and shuffled 10-shot evaluations, separately for demonstrations and instructions.
- When we intervene with two function vectors (either demonstrations and instructions, or twice with one), we sweep over the range  $[\lfloor L/4 \rfloor, \lceil L/2 \rceil]$  (as the optimal intervention depths for all models fell in this range).
  - In the main manuscript figure, we report the accuracy intervening with both additively at the  $\lfloor L/3 \rfloor$  layer.
  - In appendix figures showcasing the empirically optimal layer(s), we follow the same process described above.
- Error bars we report in all figures are standard errors of the mean, averaged within each model.
- We use the following query template in all of our instruction-based experiments:

```
<instructions>
Q: < $x_{iq}$ >
A:
```

**Figure 9: Instruction query template.** We use the query template proposed by Todd et al. (2024) and prepend the instructions to it.

- And this query template in our demonstration-based ones:

Q: $\langle x_{i1} \rangle$
A: $\langle y_{i1} \rangle$
Q: $\langle x_{i2} \rangle$
A: $\langle y_{i2} \rangle$
...
Q: $\langle x_{i10} \rangle$
A: $\langle y_{i10} \rangle$
Q: $\langle x_{iq} \rangle$
A:

**Figure 10: Demonstration query template.** We use the query template proposed by Todd et al. (2024).

## E.1 Experiment compute resources

Before running the main set of experiments, we run a few preliminary steps. We generate instructions (Appendix A) 20 times for each of the  $\approx 50$  datasets, generating shorter and longer instructions separately, resulting in approximately 2000 invocations of the stronger model used for instruction generation (in our experiments, Llama-3.1-405B). We also cache the log-probabilities of texts from WikiText-103-v1, which takes an hour or two with each model (Appendix B)

For each of the 12 models we consider (Table 4), and each of the 50 tasks, we begin by running the ‘training’ job that computes mean task-conditioned activations and estimates head causal effects. We do so with shorter and longer instructions, using each of our three uninformative baselines, yielding roughly  $12 \text{ (models)} \times 50 \text{ (tasks)} \times 2 \text{ (instruction lengths)} \times 3 \text{ (baselines)} = 3600$  jobs. With these jobs behind us, we can compute the overall top instruction FV heads for each model. With those, we can evaluate each model with the function vector constructed using the overall top heads, evaluating an intervention at every layer in both the zero-shot and shuffled 10-shot evaluations. We run each evaluation separately using the mean activations with the best performing long and short prompts, but with the heads identified averaged over the causal scores from both, and average the final results over these two mean activations.

Most of our additional experiments only require these final evaluation jobs:

- Intervening with both function vectors (§3.2; Figure 2B).
- Intervening with the same function vector twice (a control for §3.2; Figure 18).
- Constructing ‘incongruent’ FVs with top heads identified by demonstrations and mean activations from instructions (or vice versa; §3.4; Figure 3G-H).
- Constructing function vectors using the least important overall heads and the bottom heads (a control for §3.4; Figure 30)
- Steering base models with post-trained model instruction FVs (§3.5; Figure 4A).

Many of these we only run for the four models we focus our investigation on (Llama-3.2-3B, Llama-3.2-3B-Instruct, Llama-3.1-8B, Llama-3.1-8B-Instruct). However, all of these run for approximately 50 datasets, using both short and long instruction mean activations. We conservatively estimate these evaluations required another  $50 \text{ (tasks)} \times 2 \text{ (instruction lengths)} \times 4 \text{ (models)} \times 10 \text{ (additional experiments)} = 4000$  experiments.

In addition, we run the demonstration ICL extraction procedure and evaluation on all models we report. As these only have one variant, they contribute only another  $12 \text{ (models)} \times 50 \text{ (tasks)} = 600$  experiments or so.

We run all of our experiments on Volta and Pascal-series GPUs, with a single GPU sufficing for every experiment we launch. Experiment wall-clock time varied drastically by the model, the size of each

task’s dataset, and the lengths of the data points in each tasks; however, all were on the order of hours, not days.

## F Baseline and ‘skyline’ accuracies

Table 6: Baseline and ‘skyline’ accuracies by model.

Model	10-shot	Shuffled 10-shot	Best instruction	Top-5 instructions	0-shot
Llama-3.2-3B	0.7531 $\pm$ 0.0205	0.1536 $\pm$ 0.0159	0.7654 $\pm$ 0.0225	0.7105 $\pm$ 0.0227	0.1530 $\pm$ 0.0163
Llama-3.2-3B-Instruct	0.7895 $\pm$ 0.0173	0.1858 $\pm$ 0.0154	0.8638 $\pm$ 0.0172	0.8330 $\pm$ 0.0184	0.1066 $\pm$ 0.0088
Llama-3.1-8B	0.8207 $\pm$ 0.0179	0.1991 $\pm$ 0.0148	0.8200 $\pm$ 0.0208	0.7668 $\pm$ 0.0225	0.1283 $\pm$ 0.0122
Llama-3.1-8B-Instruct	0.8456 $\pm$ 0.0171	0.1793 $\pm$ 0.0160	0.8874 $\pm$ 0.0159	0.8507 $\pm$ 0.0185	0.0772 $\pm$ 0.0071
Llama-3.2-1B	0.6562 $\pm$ 0.0211	0.1300 $\pm$ 0.0141	0.6281 $\pm$ 0.0246	0.5484 $\pm$ 0.0241	0.1779 $\pm$ 0.0169
Llama-3.2-1B-Instruct	0.6930 $\pm$ 0.0191	0.1674 $\pm$ 0.0174	0.7164 $\pm$ 0.0226	0.6598 $\pm$ 0.0225	0.1566 $\pm$ 0.0136
Llama-2-7b-hf	0.7403 $\pm$ 0.0186	0.1405 $\pm$ 0.0150	0.6589 $\pm$ 0.0230	0.5816 $\pm$ 0.0232	0.1284 $\pm$ 0.0144
Llama-2-7b-chat-hf	0.8040 $\pm$ 0.0168	0.1813 $\pm$ 0.0161	0.8133 $\pm$ 0.0188	0.7715 $\pm$ 0.0202	0.0693 $\pm$ 0.0071
OLMo-2-1124-7B	0.7288 $\pm$ 0.0185	0.1713 $\pm$ 0.0157	0.8567 $\pm$ 0.0192	0.8244 $\pm$ 0.0207	0.1686 $\pm$ 0.0128
OLMo-2-1124-7B-SFT	0.7743 $\pm$ 0.0158	0.1754 $\pm$ 0.0149	0.8698 $\pm$ 0.0176	0.8390 $\pm$ 0.0197	0.1478 $\pm$ 0.0099
OLMo-2-1124-7B-DPO	0.7694 $\pm$ 0.0167	0.1663 $\pm$ 0.0159	0.8665 $\pm$ 0.0179	0.8319 $\pm$ 0.0205	0.1400 $\pm$ 0.0096
OLMo-2-1124-7B-Instruct	0.7741 $\pm$ 0.0167	0.1635 $\pm$ 0.0158	0.8699 $\pm$ 0.0175	0.8360 $\pm$ 0.0201	0.1468 $\pm$ 0.0100

We report mean model accuracies on the evaluation conditions (without interventions) and on their corresponding informative conditions, averaged over the full set of tasks.

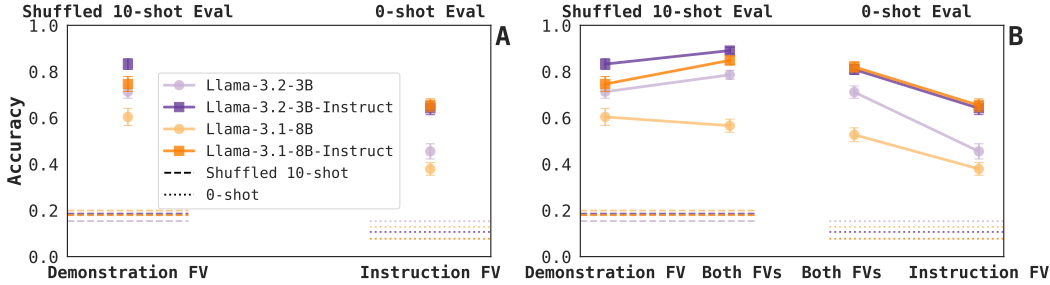
For the shuffled 10-shot evaluation condition, its informative condition is 10-shot (without label shuffling).

For the 0-shot evaluation, we report both each model’s accuracy with the best instruction for it for that task, and the mean accuracy with the five best instructions (which were used to compute the causal indirect effects, §2).

Errors reflect the standard error of the mean.

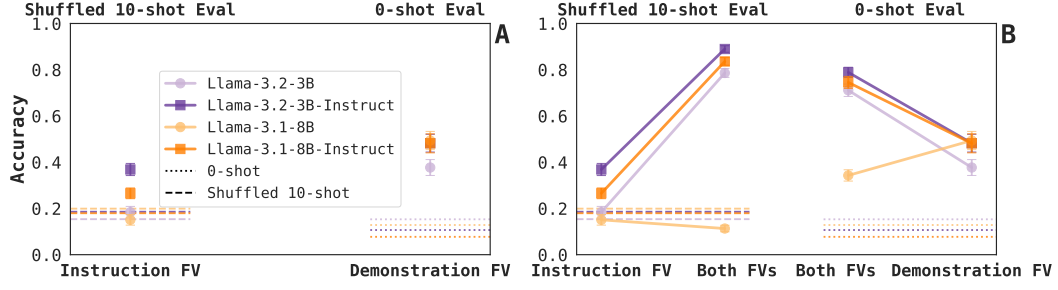
## G Additional results for findings 1 and 2

### G.1 Findings 1 and 2 with empirically optimal layer



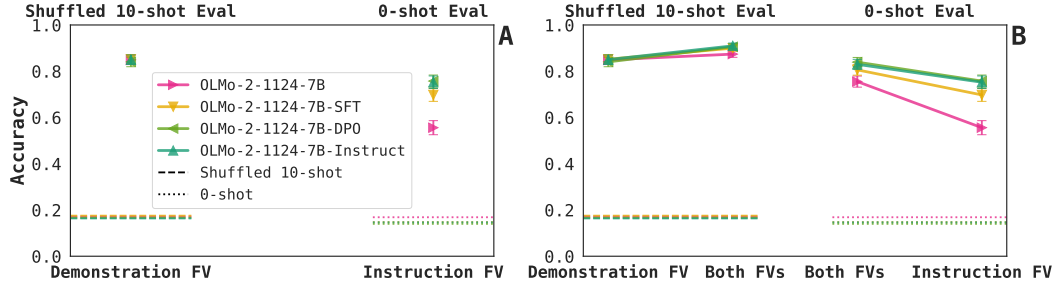
**Figure 11: Empirically optimal intervention depth version of Figure 2** This figure matches Figure 2, but using the empirically optimal intervention layer for each model and intervention, rather than fixing to the  $\lfloor L/3 \rfloor$  layer. This showcases the ceiling potential of function vectors interventions, above what might be lost by selecting intervention depth using a fixed rule. **(A)** In the post-trained models, the same intervention depth is optimal for both function vectors:  $11/28 = 0.3929$  for Llama-3.2-3B-Instruct and  $14/32 = 0.4375$  for Llama-3.1-8B-Instruct; in the base models, it varies by the choice of FV. For Llama-3.2-3B, it is  $13/28$  for demonstrations and  $9/28$  for instructions; for Llama-3.1-8B, it is  $8/32$  for demonstrations and  $15/32$  for instructions. We qualitatively match previous observations. **(B)** For three of the four models, *adding both vectors to the same layer performs best*: Llama-3.2-3B (layer  $9/28$ ), Llama-3.2-3B-Instruct (layer  $11/28$ ), and Llama-3.1-8B-Instruct (layer  $13/32$ ). Only for the base Llama-3.1-8B model does the highest accuracy arise from different intervention depths: adding the demonstration FV at layer  $10/32$  and the instruction FV at layer  $8/32$ .

### G.2 Findings 1 and 2 with FV-incongruent evaluations

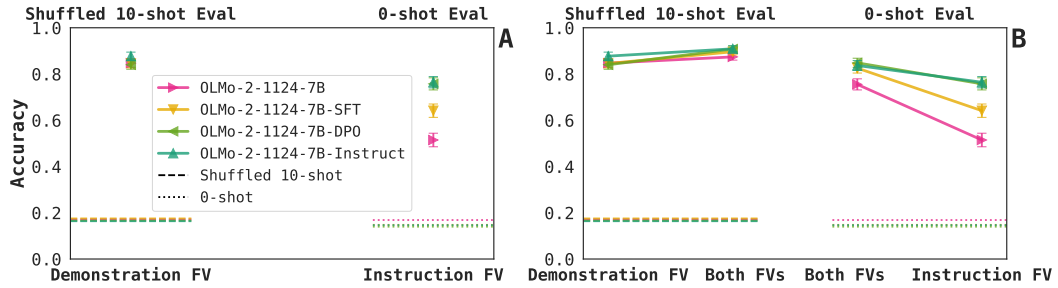


**Figure 12: Mistmatched evaluation version of Figure 2.** We report evaluation results from the evaluations that are incongruent with the function vector extraction settings— 0-shot for demonstration FVs and shuffled 10-shot for instruction FVs. Both FVs perform worse in the incongruent setting, but instruction FVs more so.

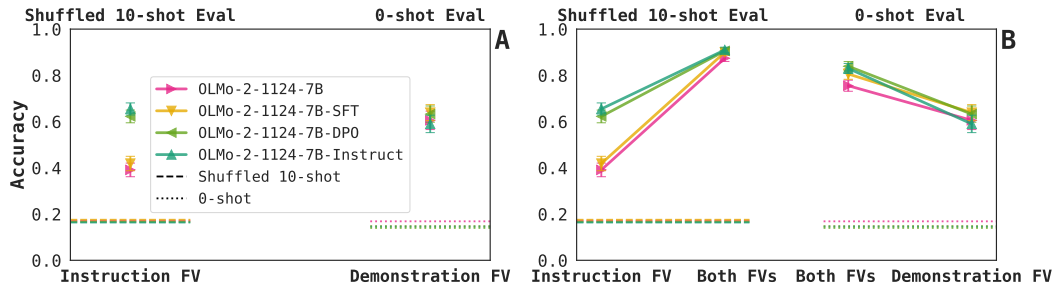
### G.3 Findings 1 and 2 with OLMo models



**Figure 13: A replication of Figure 2, with the OLMo models explored in subsection 3.5.**

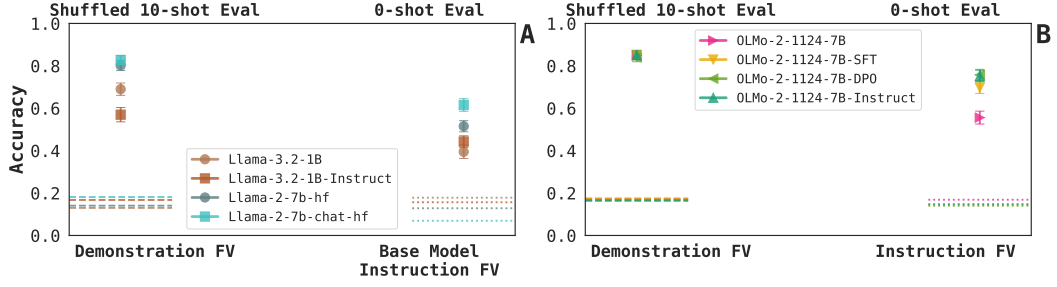


**Figure 14: A replication of Figure 2, with the OLMo models explored in subsection 3.5, using the empirically-optimal intervention depth; i.e., Figure 11, but with the OLMo models.**

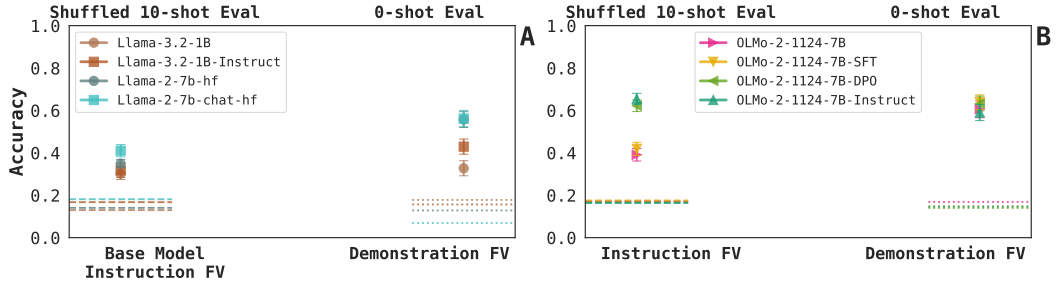


**Figure 15: A replication of Figure 2, with the OLMo models explored in subsection 3.5, using the mismatched evaluation version; i.e., Figure 12, but with the OLMo models.**

#### G.4 Findings 1 additional model results

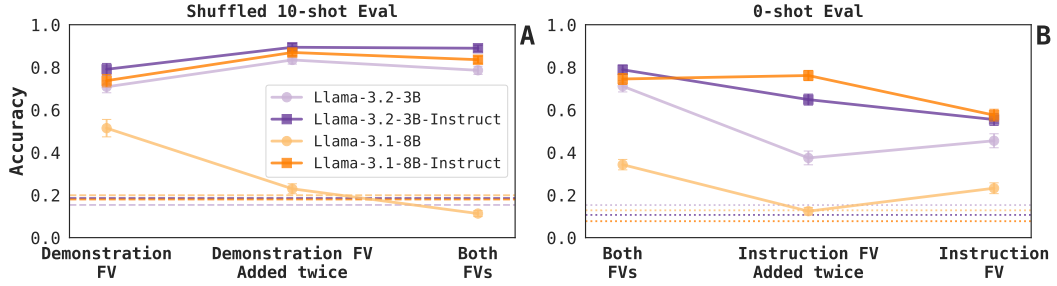


**Figure 16: Additional model results matching Figure 2A.** We report evaluation results for the rest of the models we compare. Panel B in this model also appears in the main manuscript as Figure 4B.



**Figure 17: Additional model results with mismatched evaluations.** This is an unmatched evaluations version of Figure 16.

#### G.5 Adding an FV twice control



**Figure 18: Adding the same function vector twice control condition.** We observe that in some cases, adding the same function vector twice is close to, if not better than adding both function vectors. This surprising effect happens less often when we examine the incongruent evaluations (Figure 19), suggesting that adding both function vectors confers advantages in both task presentations. This effect is also weaker when we examine the empirically optimal intervention depths (Figure 20).



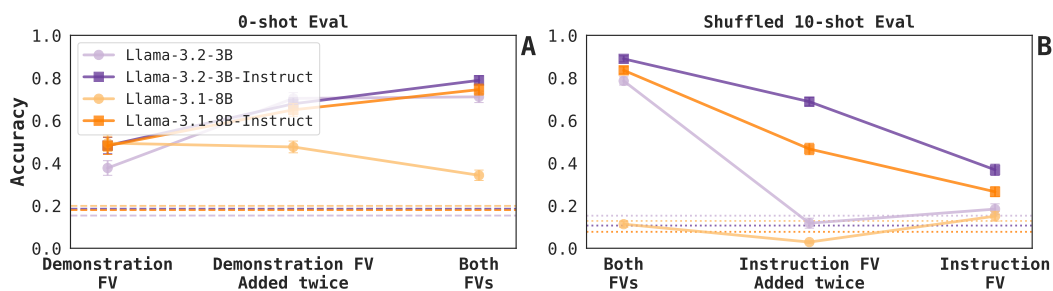


Figure 19: Adding the same function vector twice control condition, mismatched evaluations. See Figure 18.

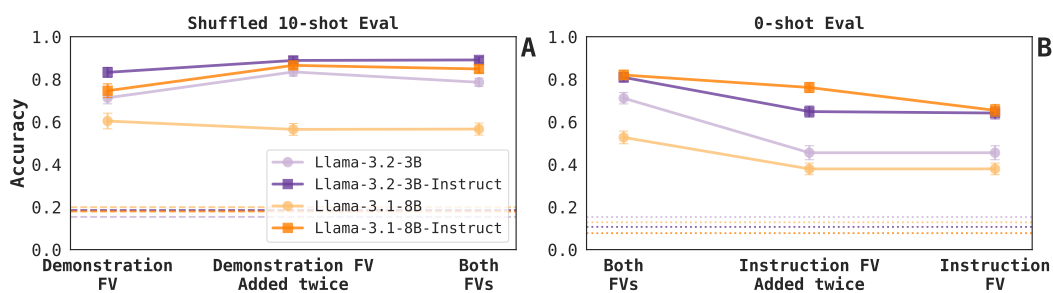
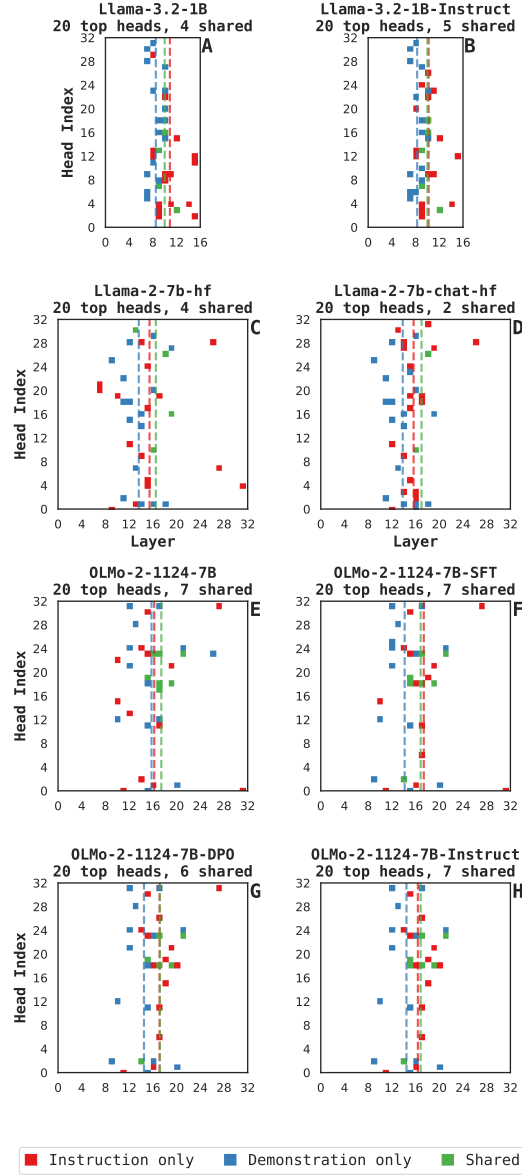


Figure 20: Adding the same function vector twice control condition, with empirically optimal layers. See Figure 18.

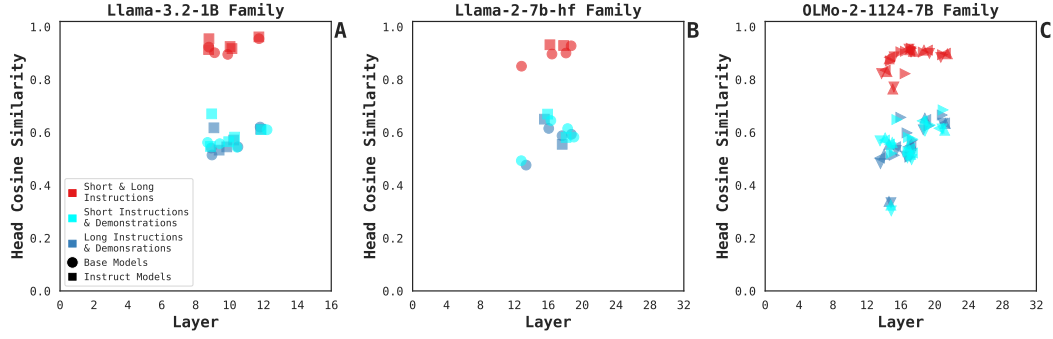
## H Additional results for finding 3

### H.1 Demonstration and instruction top heads for additional models



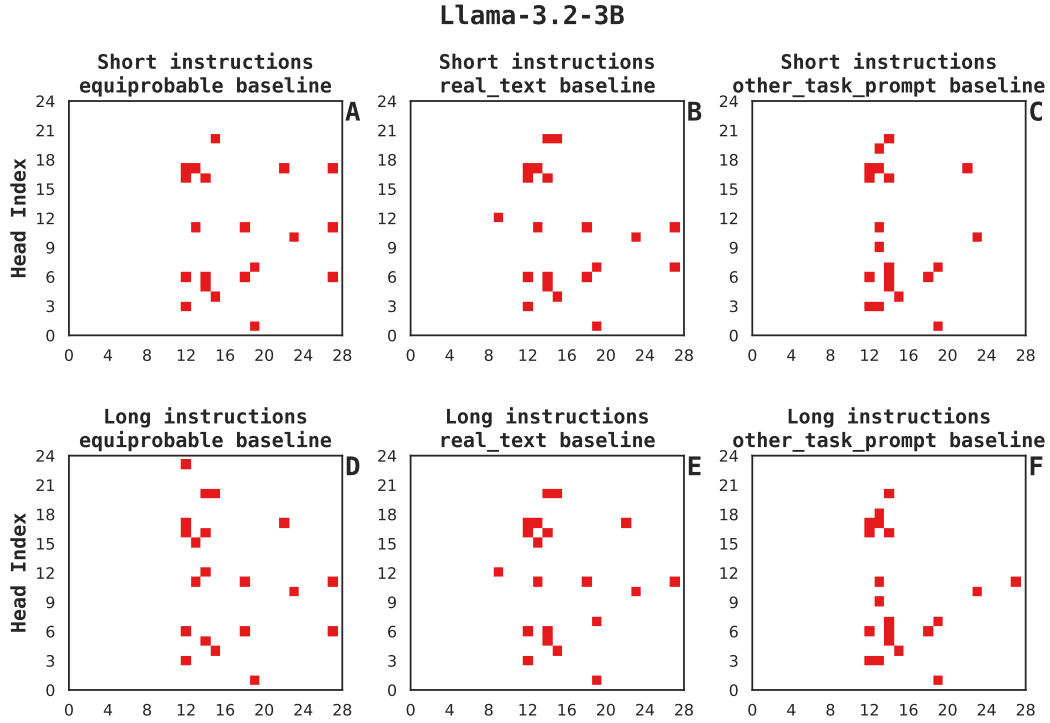
**Figure 21: Shared top heads results for additional models.** This panel follows Figure 3A-D for the rest of the models we evaluated.

### H.2 Shared attention head mean activation similarity for additional models

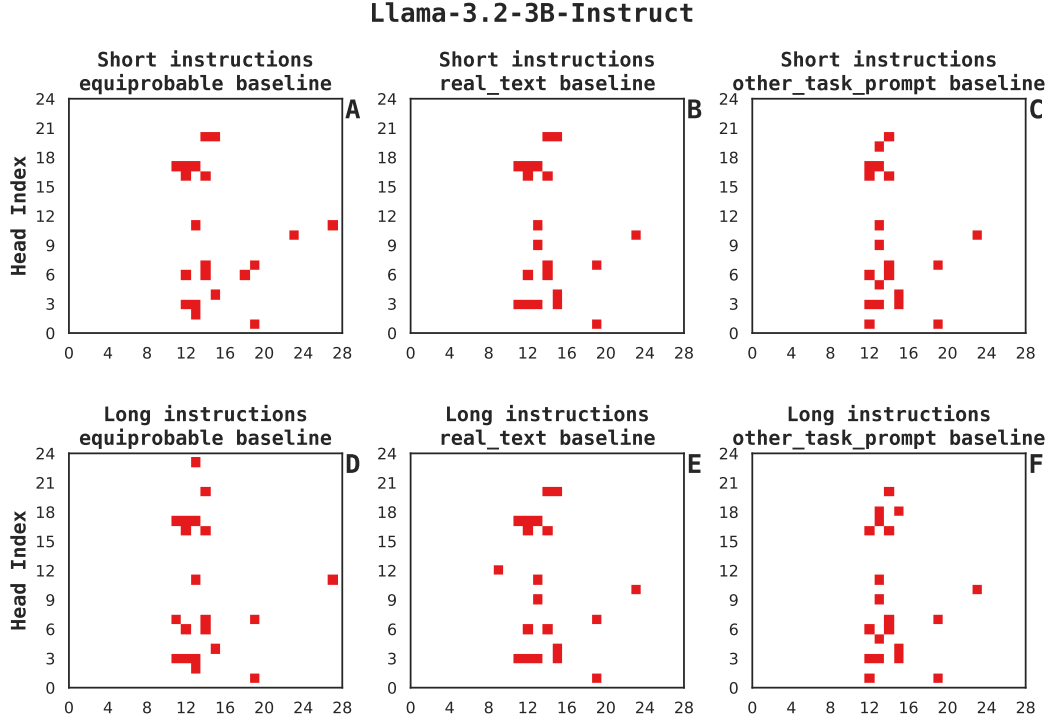


**Figure 22: Shared top heads similarities for additional models.** This panel follows Figure 3F-G for the rest of the models we evaluated. For the OLMo family of models, the triangular markers follow the ones used in Figure 4B.

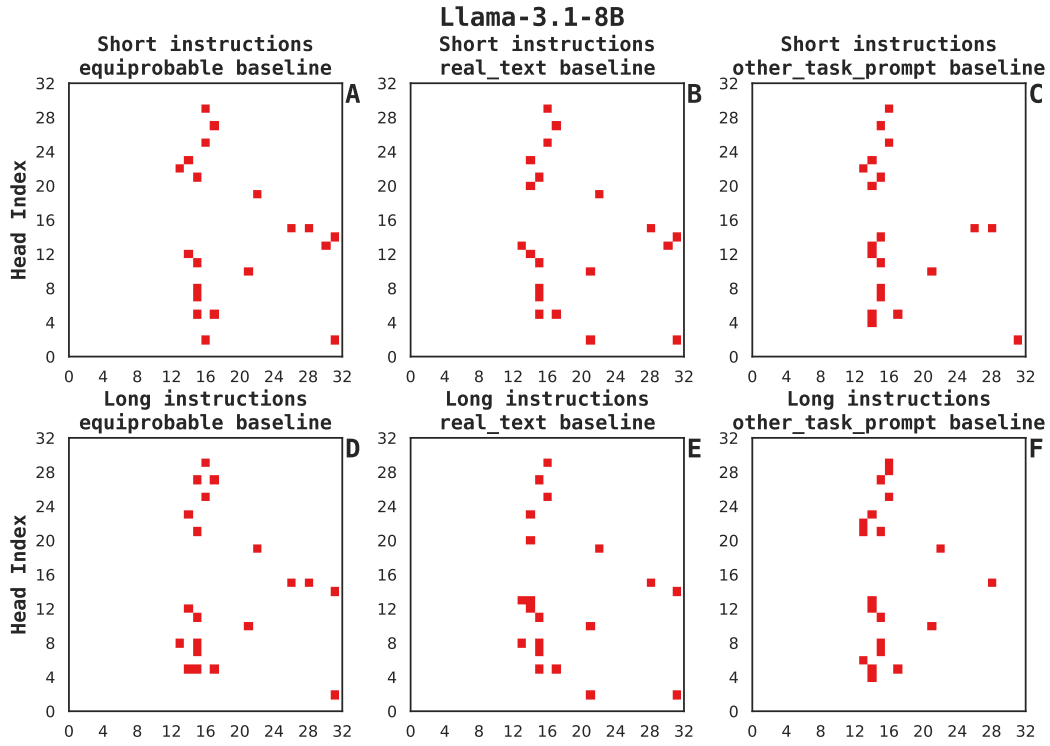
### H.3 Attention head set similarity by prompt length / baseline type



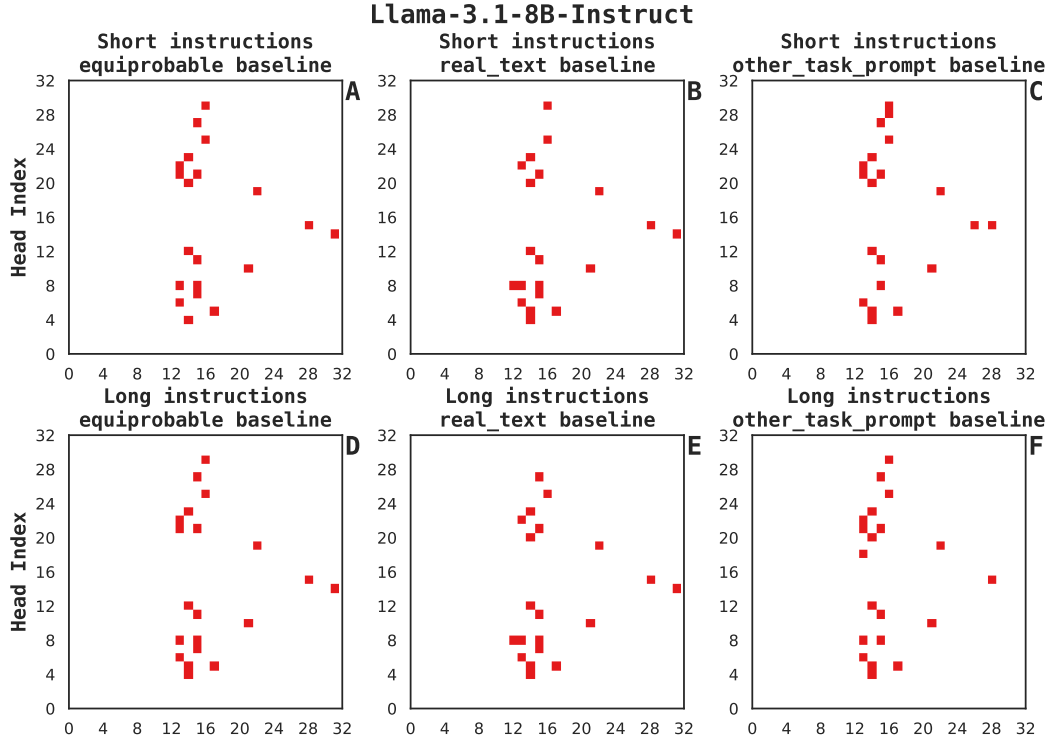
**Figure 23: Top heads split by instruction length and uninformative baseline.** This figure follows Figure 3A-D, but breaks it down by individual conditions, for the base Llama-3.2-3B model.



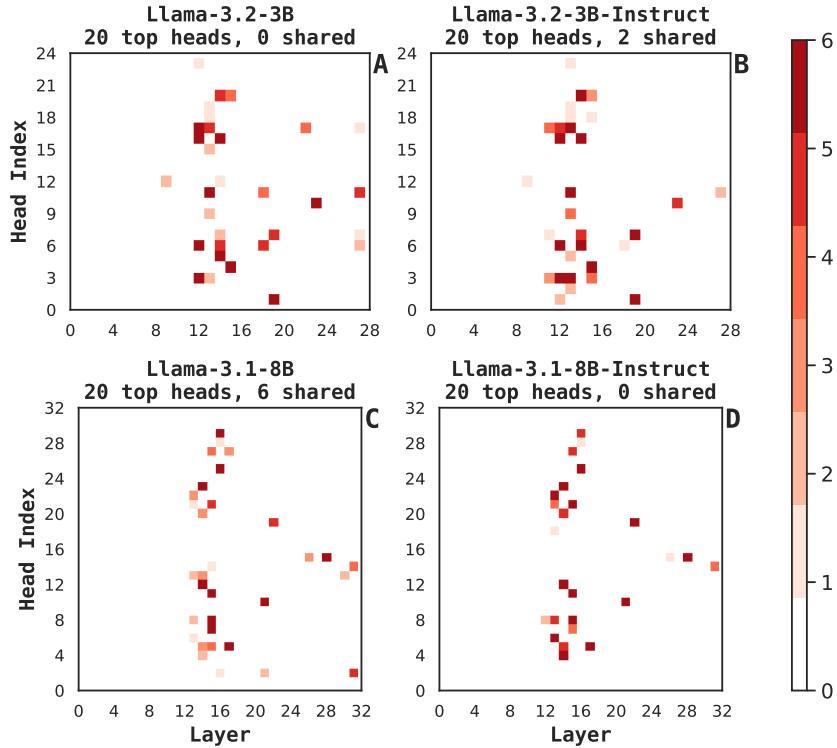
**Figure 24: Top heads split by instruction length and uninformative baseline.** This figure follows Figure 3A-D, but breaks it down by individual conditions, for the post-trained Llama-3.2-3B-Instruct model.



**Figure 25: Top heads split by instruction length and uninformative baseline.** This figure follows Figure 3A-D, but breaks it down by individual conditions, for the base Llama-3.1-8B model.



**Figure 26: Top heads split by instruction length and uninformative baseline.** This figure follows Figure 3A-D, but breaks it down by individual conditions, for the post-trained Llama-3.1-8B-Instruct model.



**Figure 27: Count of occurrence in top heads over different instruction lengths and baselines.** This figure summarizes Figure 23–Figure 26. For each attention head, we count how many times it appears in the top  $|\mathcal{A}| = 20$  heads over the two instruction lengths and three uninformative baselines.

## I Finding 4 appendix results

### I.1 Finding 4 with empirically optimal layer

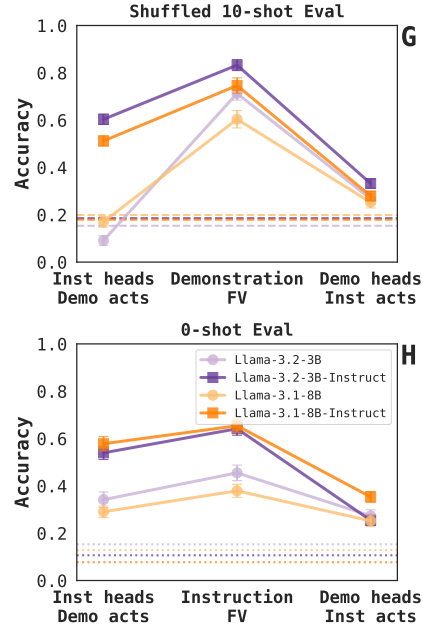


Figure 28: Empirically optimal intervention depth version of Figure 3G-H

### I.2 Finding 4 with OLMo models

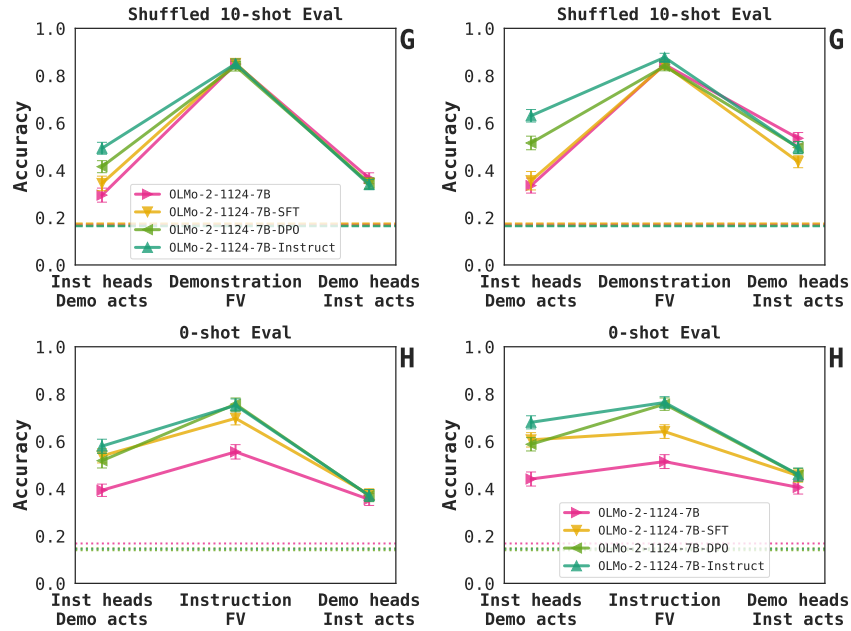


Figure 29: (Left) A replication of the joint intervention results of Figure 3G-H, using the OLMo models explored in subsection 3.5. (Right) Same as the left panel, but using the empirically optimal intervention depths; a version of Figure 28, but with the OLMo models.

Table 7: Overall head causal indirect effect ratio between demonstrations and instructions

Heads	CIE Ratio	Llama-3.2-3B	Llama-3.2-3B-Instruct	Llama-3.1-8B	Llama-3.1-8B-Instruct
10	Mean	5.438	4.688	6.498	2.749
	Median	2.730	1.801	5.187	2.299
20	Mean	3.901	3.570	4.794	2.181
	Median	1.482	1.359	2.894	1.337
100	Mean	2.566	2.477	3.384	1.769
	Median	1.036	1.027	1.850	1.273

Heads	CIE Ratio	Llama-3.2-1B	Llama-3.2-1B-Instruct	Llama-2-7b-hf	Llama-2-7b-chat-hf
10	Mean	5.427	4.562	4.807	4.613
	Median	4.616	4.338	4.110	5.914
20	Mean	4.128	3.463	3.788	3.809
	Median	1.846	1.356	2.676	2.660
100	Mean	2.531	2.435	2.153	2.740
	Median	1.110	1.264	0.957	1.711

Heads	CIE Ratio	OLMo-2-1124-7B	OLMo-2-1124-7B-SFT	OLMo-2-1124-7B-DPO	OLMo-2-1124-7B-Instruct
10	Mean	2.734	3.439	3.415	3.534
	Median	1.685	2.675	2.511	2.658
20	Mean	2.067	2.522	2.613	2.682
	Median	1.112	1.356	1.425	1.409
100	Mean	1.242	1.577	1.724	1.763
	Median	0.490	0.717	0.813	0.836

For each model, for the top  $N \in \{10, 20, 100\}$  heads identified by either demonstrations or instructions, we compute the mean and median causal indirect effect (CIE) scores, and report the ratio between the two, demonstration scores divided by instruction scores. We observe the following trends:

- (1) Mean CIE ratios are consistently higher than the median CIE ratios. This suggests the distribution of demonstration CIEs has a heavier positive tail than that of instruction CIEs.
- (2) As the number of heads  $N$  examined increases, both ratios drop closer to 1 (and in some cases, the medians drop below 1). This offers further evidence to the heavy-tailed nature of the demonstration CIEs, as compared to the instruction CIEs. This also suggest that instruction task representations are more diffuse in the models, as at high numbers of heads, the median contribution is higher for instructions than it is for demonstrations.

## J Finding 4 analyses and controls

Table 8: Localizer experiment causal indirect effects.

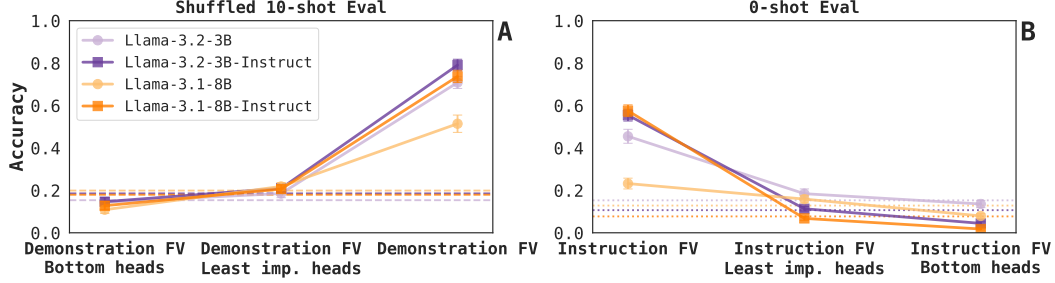
	Llama-3.2-3B	Llama-3.2-3B-Instruct	Llama-3.1-8B	Llama-3.1-8B-Instruct
<b>Overall median demonstration CIE</b>	4.1926e-07	4.3410e-06	1.1681e-06	2.7066e-06
<b>Demonstration heads / demonstration CIE</b>	1.3882e-02	2.5550e-02	1.3635e-02	1.4652e-02
<b>Instruction heads / demonstration CIE</b>	2.8993e-03	4.1367e-03	3.0959e-03	3.6646e-03
<b>Localizer difference</b>	1.1313e-03	1.4519e-03	2.1435e-03	1.7800e-03
<b>Demonstration heads / instruction CIE</b>	1.7680e-03	2.6849e-03	9.5237e-04	1.8846e-03
<b>Instruction heads / instruction CIE</b>	3.5586e-03	7.1606e-03	2.8476e-03	6.7488e-03
<b>Overall median instruction CIE</b>	2.0577e-06	8.1827e-06	7.1886e-06	1.1005e-05

We observe consistently higher causal indirect scores for instruction FV heads in the demonstration setting, compared to using demonstration FV heads in the instruction setting (“localizer difference,” highlighted).

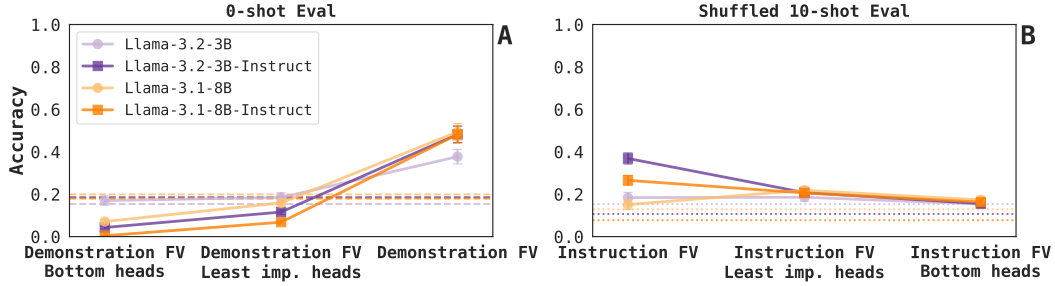
This effect is not merely due causal effects being higher in the demonstration setting; the first and last row provide the median CIE in each condition, and we observe that the median CIE in the instruction setting is higher in every case, by as much as an order of magnitude for some models.

### J.1 Finding 4 control conditions

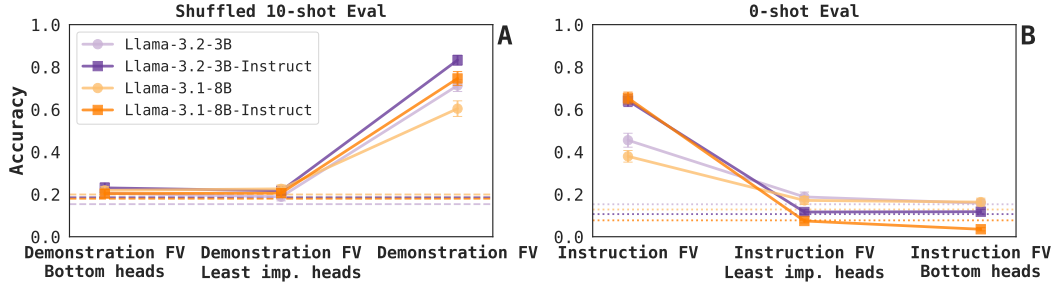
We perform two control experiments to ensure that this observed asymmetry is meaningful (and not an artifact of selecting arbitrary sets of heads). In one, we select the set of heads with the lowest absolute causal scores across both instructions and demonstrations, which we consider to be the heads most unrelated to inducing task representations from either presentation. In another, we select the bottom heads — that is, the ones with the largest negative causal scores for either instructions or prompts. We report both in Appendix J.1. Performance with the least important heads is indistinguishable from the baselines, and performance with the bottom heads is often below the baselines. The existence of heads with negative causal scores is, itself, curious — these are attention heads whose mean task-conditioned activations *lower* the probability assigned to the correct token, rather than raise it.



**Figure 30: Localizer control conditions: evaluating the least important and bottom heads.** To validate the effects we observe in Figure 3G-H and Figure 28 are not a function of selecting any arbitrary set of attention heads, we report these two control conditions. In both, accuracy is at or below chance, as expected. We observe similar, though weaker results when using the mismatched evaluations (Figure 31), and similar, equally strong results when using the empirically optimal layers (Figure 32).



**Figure 31: Localizer control conditions: evaluating the least important and bottom heads, with mismatched evaluations.** See Figure 30.

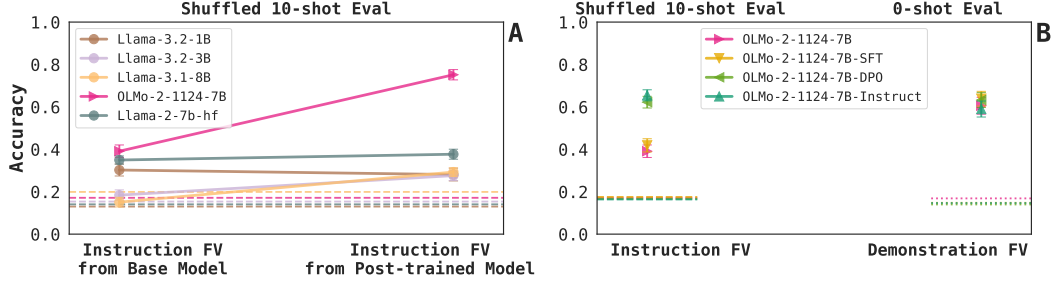


**Figure 32: Localizer control conditions: evaluating the least important and bottom heads, with empirically optimal layer.** See Figure 30.

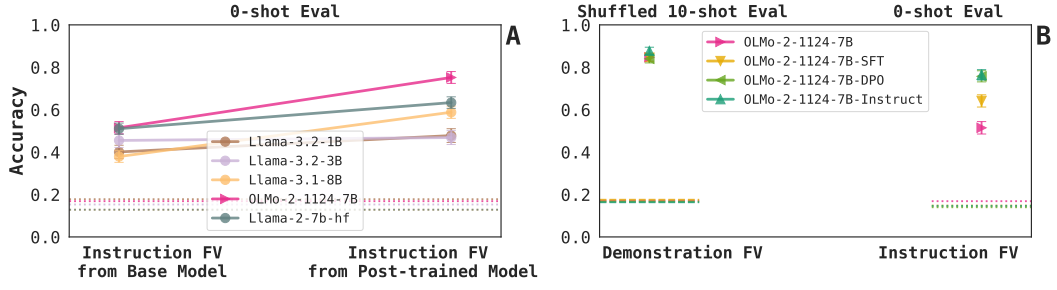


## K Finding 5 appendix results

### K.1 Post-training result variations



**Figure 33: Figure 4 with mismatched evaluations.** (A) We observe that instruction FVs steer base models in the mismatched evaluation setting as well. (B) In the OLMo model family, mismatched evaluation performance is roughly equal between the two FV types.



**Figure 34: Figure 4 with empirically optimal intervention layer.** (A) We observe that instruction FVs transfer beneficially steer base models when evaluated at the empirically optimal intervention layer. (B) We observe the same qualitative effect here as we did with the fixed  $|L/3|$  intervention depth—two accuracy jumps, one in the SFT model and the second in the DPO model.