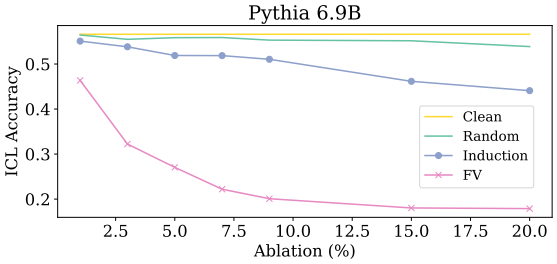
Which Attention Heads Matter for In-Context Learning?

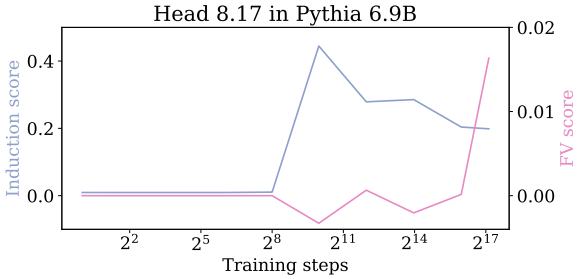
Kayo Yin Jacob Steinhardt
UC Berkeley

Abstract

Large language models (LLMs) exhibit impressive in-context learning (ICL) capability, enabling them to perform new tasks using only a few demonstrations in the prompt. Two different mechanisms have been proposed to explain ICL: induction heads that find and copy relevant tokens, and function vector (FV) heads whose activations compute a latent encoding of the ICL task. To better understand which of the two distinct mechanisms drives ICL, we study and compare induction heads and FV heads in 12 language models. Through detailed ablations, we discover that few-shot ICL performance depends primarily on FV heads, especially in larger models. In addition, we uncover that FV and induction heads are connected: many FV heads start as induction heads during training before transitioning to the FV mechanism. This leads us to speculate that induction facilitates learning the more complex FV mechanism that ultimately drives few-shot ICL[1].



(a) ICL accuracy of Pythia 6.9B across different percentages of heads ablated.



(b) Induction score (blue) and FV score (pink) of an FV head during training.

Figure 1. (a) Ablating function vector (FV) heads significantly degrades few-shot in-context learning (ICL) accuracy, while ablating induction heads has minimal impact beyond ablating random heads. (b) Evolution of an FV head during training, demonstrating high induction scores earlier in training that decrease as FV score emerges. This pattern suggests induction may serve as a precursor for FV mechanism.

1. Introduction

One of the most remarkable features of large language models (LLM) is their ability to perform in-context learning (ICL), where they can adapt to various new tasks using only context given in the prompt at inference time. This capability has become crucial to adapt pre-trained LLMs to specific tasks, sparking significant research into its underlying mechanisms (Olsson et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023).

To date, two key mechanisms have been associated with ICL, each supported by different lines of evidence. First, *induction circuits* (Elhage et al., 2021) were hypothesized to be the primary mechanism behind ICL in LLMs (Olsson et al., 2022; Singh et al., 2024; Crosbie & Shutova, 2024; Dong et al., 2023). Induction circuits operate by identifying previous occurrences of the current token in the prompt and copying the subsequent token. More recently, Todd et al. (2024) and Hendel et al. (2023) propose the existence of *function vectors* (FV). FVs are a compact representation of a task extracted from specific attention heads, and they can be added to a model's computation to recover ICL behavior without in-context demonstrations.

To resolve whether one or both mechanisms drive ICL in transformer LLMs, we conduct a comprehensive study of

. Correspondence to: Kayo Yin <kayoyin@berkeley.edu>, Jacob Steinhardt <jsteinhardt@berkeley.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]Code and data: `https://github.com/kayoyin/icl-heads`.

Table 1. Summary of findings in this work, where ✓ represents findings with evidence directly shown by our experiments and ∼ represents conjectures that our results suggest.

Findings	Evidence	Section	Contribution
Induction heads and FV heads are distinct.	✓	3	Context-setting
Induction scores and FV scores are correlated.	✓	3	Context-setting
Ablating FV heads hurts few-shot ICL accuracy more than ablating induction heads.	✓	4	Main finding
Some FV heads evolve from induction heads during training.	✓	5	Main finding
FV heads implement more complex or abstract computations than induction heads.	∼	5	Speculation
Induction heads serve as a stepping stone for models to develop FV heads	∼	5	Speculation

the attention heads implementing these mechanisms (termed *induction heads* and *FV heads*) across 12 decoder-only transformer models ranging from 70M to 7B parameters (Table 2) and 45 natural language ICL tasks (listed in Appendix A.8). Our analysis reveals several key findings.

First, we verify that there is a difference to explain, i.e. that induction and FV heads are indeed distinct (§3). Across models, there is a low or zero overlap between induction and FV heads. These heads also have distinct characteristics: induction heads generally appear in slightly earlier layers than FV heads, and emerge significantly earlier during training. On the other hand, there are correlations in behavior: FV heads behave more similarly to induction heads than a random head from the same network, and vice versa.

Second, through ablation studies (§4), we demonstrate that FV heads are the primary drivers of ICL performance. Removing FV heads substantially degrades ICL task accuracy, while removing induction heads has a limited effect (Figure 1a). This effect is consistent across all 12 models we studied, and is more pronounced in larger models (§4). Interestingly, this challenges the prevailing view of induction heads as the key mechanism for few-shot ICL (Olsson et al., 2022; Crosbie & Shutova, 2024; Dong et al., 2023).

Third, we reconcile our findings with previous work by identifying three key methodological differences (§7): earlier studies used a different metric for ICL that does not strongly track few-shot performance, did not account for correlations between FV and induction heads, and sometimes focused on small models. We find the choice of metric is the most significant factor, as discussed in §7 and demonstrated by experiments in §4.

Finally, by analyzing training dynamics (§5), we uncover a surprising developmental relationship: many induction heads *evolve* into FV heads during training, but the reverse never occurs (Figure 1b). This leads us to speculate that induction heads facilitate learning the more complex FV heads for ICL – the FV mechanism is more effective at performing ICL, and therefore eventually replaces the simpler induction mechanism §6.

Aside from clarifying the drivers of few-shot ICL, our findings offer broader lessons for model interpretability research. They highlight how correlations between related mechanisms can lead to illusory explanations (e.g. the confounding effect of the correlation between induction and FV heads), and the choice of definitions may lead to different conclusions (e.g. different metrics to measure ICL). Additionally, our results challenge strong versions of universality – the difference between the importance of FV heads and induction heads shifts significantly with model scale (§4).

Table 2. Models studied in this work. We use huggingface implementations (Wolf et al., 2020) for all models. We report the number of parameters, number of layers $|L|$, and number of attention heads $|a|$ for each model.

Model	Param.	$\|L\|$	$\|a\|$
Pythia (Biderman et al., 2023)	70M	6	48
	160M	12	144
	410M	24	384
	1B	16	128
	1.4B	24	384
	2.8B	32	1024
	6.9B	32	1024
GPT-2 (Radford et al., 2019)	117M	12	144
	345M	24	384
	774M	36	720
	1.6B	48	1200
Llama 2 (Touvron et al., 2023)	7B	32	1024

2. Background

We present a comparative analysis of two mechanisms proposed to explain ICL: induction heads (Elhage et al., 2021; Olsson et al., 2022) and FV heads (Todd et al., 2024; Hendel et al., 2023). In this section, we first present the different conceptualizations of ICL. Then, we provide the definitions of induction heads and FV heads.

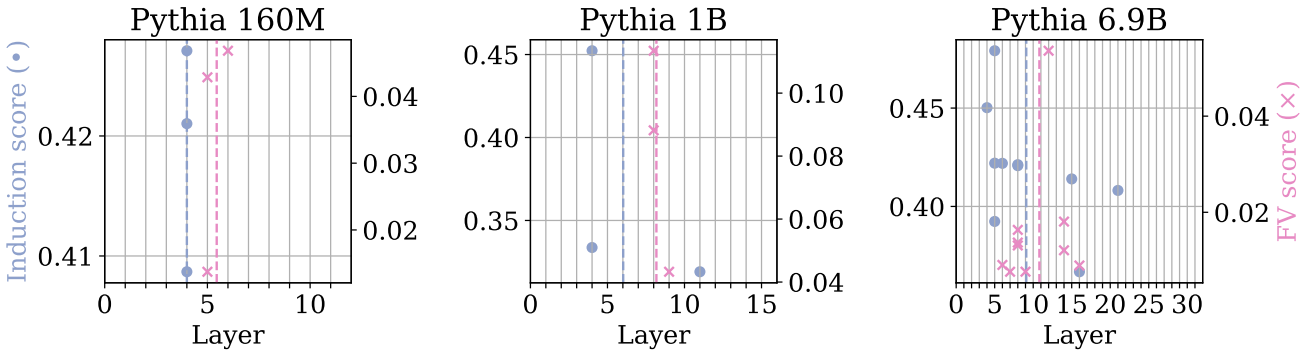


Figure 2. Location of induction heads (blue) and FV heads (pink) in model layers. The average layer of induction and FV heads are shown in blue and pink dotted lines respectively. Most induction heads appear in early-middle layers, FV heads appear at layers slightly deeper than induction heads.

2.1. In-context learning

ICL broadly refers to the ability of a machine learning model, especially LLMs, to adapt its behavior and perform new tasks by leveraging contextual information provided with the input prompt, without any changes to its underlying parameters. This allows the model to dynamically interpret and respond to a wide range of tasks solely based on the information given at inference time, such as instructions or examples, making it highly adaptable. Two distinct conceptualizations of ICL have emerged in the literature, reflecting different operationalizations of the phenomenon:

ICL as few-shot task accuracy. This conception, popularized by Brown et al. (2020), defines ICL through few-shot performance – the model's accuracy on novel tasks when provided with a few demonstrations in the prompt. The majority of existing works studying ICL focuses on this conception of ICL (Dong et al., 2023; Akyürek et al., 2023; Wei et al., 2023; Bansal et al., 2023; Crosbie & Shutova, 2024), and ICL is often used synonymously with this few-shot ability at inference time. We adopt this definition of ICL in this paper as well, as it is currently the most standard in the literature. Unless specified otherwise, "in-context learning" in this paper will refer to this definition.

ICL as context-dependent loss reduction. On the other hand, few-shot learning could be interpreted as a subset of ICL, where ICL is defined more broadly as the model's ability to reduce loss at later tokens by using the context of earlier observations (Kaplan et al., 2020; Olsson et al., 2022; Lampinen et al., 2024). In this conception, ICL is measured by the difference between the model loss on earlier tokens and later tokens in the sequence. This difference was previously called "ICL score" in Olsson et al. (2022), to avoid confusion we call this **"token-loss difference"**.

While these two conceptualizations of ICL have often been conflated, our experiments demonstrate these metrics can diverge significantly (§4). Models may maintain high few-shot ICL accuracy while showing reduced token-loss difference, and vice versa. This distinction helps explain conflicting conclusions about ICL mechanisms in prior studies.

2.2. Induction heads

Induction heads were first identified by Elhage et al. (2021) and extensively studied by Olsson et al. (2022) as the mechanism behind ICL. They are attention heads that identify repeated patterns in the input: when processing a token, they attend to the token that followed a previous occurrence of the same token, predicting it will appear next.

The initial evidence for induction heads' role in ICL came from Olsson et al. (2022), who studied small attention-only models (1-3 layers). They observed that the emergence of induction heads during training coincided with improvements in ICL ability – measured as the difference between the loss at the 500th versus 50th token in the context. Their ablation studies showed that removing induction heads impaired this metric.

To identify and analyze induction heads, we measure their **induction scores** using the TransformerLens framework (Nanda & Bloom, 2022). For each attention head a, we compute its induction score on a synthetic sequence constructed by repeating a uniformly sampled random token sequence: $r = r_1 r_2 ... r_{50} r'_1 r'_2 ... r'_{50}$. The induction score is defined as:

$$S_I(a, r) = \sum_{i=1}^{50} a_{r'_i \to r_{i+1}}$$

where $a_{r'_i \to r_{i+1}}$ represents the attention weight that head a places on token r_{i+1} when processing token r'_i. For each attention head in each model, we take the mean induction score over 1000 samples of random sequences r, normalized by total attention mass to obtain a value between 0 and 1.

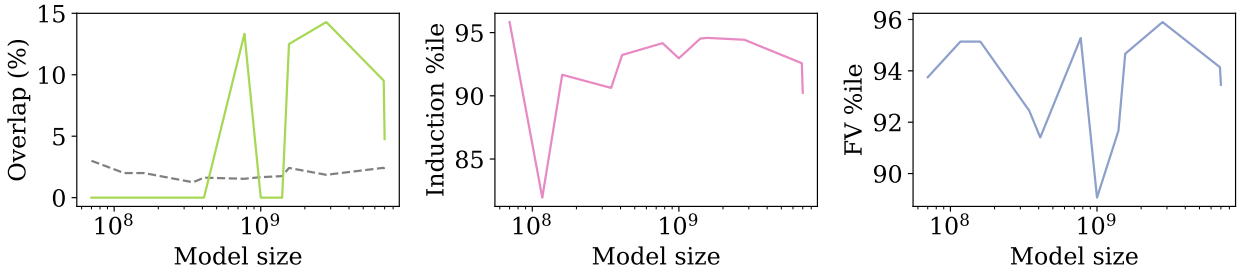


Figure 3. Percentage of head overlap between induction and FV heads (left) in green, and between induction and randomly sampled heads in gray. Percentile of induction score of FV heads (center). Percentile of FV score of induction heads (right). There is little overlap between induction and FV heads, but FV heads have relatively high induction scores and vice versa.

2.3. FV heads

Function vectors (FV) were concurrently discovered by Todd et al. (2024) and Hendel et al. (2023). They represent a different mechanism for ICL: FVs are compact vector representations of ICL tasks that can be extracted from specific attention heads and added back into the language model's computations to reproduce ICL behavior. We refer to the attention heads that encode and transport these function vectors as **FV heads**.

To identify FV heads, we employ the casual mediation analysis framework from Todd et al. (2024). For each ICL task t in our task set $\mathcal{T}$, where t is defined by a dataset P_t of in-context prompts $p_i^t \in P_t$ consisting of input-output pairs (x_i, y_i), we:

1. Compute the mean activation of an attention head a over prompts in P_t: $\bar{a}^t = \frac{1}{P_t} \sum_{p_i^t \in P_t} a(p_i^t)$
2. Create corrupted ICL prompts $\tilde{p}_i^t \in \tilde{P}_t$ by randomly shuffling the output labels $\tilde{y}_i$ while maintaining the same inputs x_i
3. Measure each head's **function vector score** (FV score) as its causal contribution to recovering the correct output y for the input x given corrupted examples $(x_i, \tilde{y}_i)$ when its activation pattern is replaced with the mean task-conditioned activation $\bar{a}^t$:

$$S_{FV}(a|\tilde{p}_i^t) = f(\tilde{p}_i^t|a := \bar{a}^t)[y] - f(\tilde{p}_i^t)[y].$$

For each attention head, we take the mean FV score across 37 natural language ICL tasks from (Todd et al., 2024) (Appendix A.8), using 100 prompts per task. Each prompt contains 10 input-output demonstration pairs followed by a single test instance.

3. Induction heads and function vector heads are distinct but correlated

Before analyzing the relative contributions of induction and FV heads to ICL performance, we first establish that these represent distinct mechanisms, while noting important correlations between them.

3.1. Head locations

We begin by examining the location of the top induction and FV heads within the models. Figure 2 shows the layers where the top 2% induction heads and FV heads appear in three representative Pythia models (see Appendix A.10 for all 12 models). In certain experiments such as this one, we need to decide on a threshold to differentiate between meaningful induction or FV heads and the long tail of attention heads that perform neither induction nor the FV mechanism. The 2% threshold was carefully chosen following previous work by Todd et al. (2024), and verified with the shape of the distribution of induction and FV scores (Appendix A.2) that this threshold indeed meaningfully distinguish the important heads from others.

In general, induction heads seem to appear in early-middle layers and FV heads appear in slightly deeper layers than induction heads, although this trend is not statistically significant in all models (Appendix A.10). This suggests that induction and FV heads do not fully overlap and are indeed distinct mechanisms. Moreover, the deeper locations of FV heads may indicate they implement more abstract computations than induction heads, though this interpretation remains speculative.

3.2. Overlap between induction and FV heads

To further examine how distinct induction and FV heads are, we analyze the extent of the overlap between the two types of heads in two ways.

First, we measure direct overlap - the percentage of heads that rank in the top 2% for both mechanisms: $100 \times \frac{|IH \cap FV|}{|IH|}$ where IH and FV represent the sets of top induction and FV heads respectively. The results show minimal overlap: seven of our twelve models show zero overlap, with the remaining models showing only 5-15% overlap (Figure 3 left). This leads us to conclude that **induction heads and**

FV heads are mostly distinct. This motivates our subsequent experiments that study induction and FV heads as two separate phenomena.

However, a more nuanced pattern emerges when we compute the percentile of the induction score of the top 2% FV heads (Figure 3 center) and the percentile of the FV score of the top 2% induction heads (Figure 3 right). In most models, FV heads are at around the 90-95th percentile of induction scores, and vice versa. Therefore, although there is little overlap between the sets of induction and FV heads, **induction and FV scores are correlated**: FV heads have high induction scores relative to other attention heads, and induction heads have relatively high FV scores [2].

4. Function vector heads drive in-context learning

Having established that induction and FV heads represent distinct mechanisms, we now investigate their relative causal importance for ICL through systematic ablation studies. Our analysis focuses primarily on few-shot ICL accuracy while also examining effects on metrics used in previous work for comparison.

4.1. Method

Ablation. To assess the causal contribution of different attention heads, we measure how ICL performance changes when specific heads are disabled. We use mean ablation, where we replace each target head's output with its average output across our task dataset (described in later sections). This approach avoids the out-of-distribution effects associated with zero ablation (Hase et al., 2021; Wang et al., 2023; Zhang & Nanda, 2024), though our findings remain robust across different ablation methods (Appendix A.4).

To control for the correlation between induction and FV heads identified in Section 3, we introduce ablation with "exclusion": when ablating n FV heads, we select the top n heads by FV score that are not in the top 2% by induction score, and vice versa. This helps isolate the unique contributions of each mechanism.

Few-shot ICL accuracy. We primarily evaluate ICL performance on a series of few-shot ICL tasks. Each ICL task is defined by a set of input-output pairs (x_i, y_i). The model is prompted with 10 input-output exemplar pairs that demonstrate this task, and one query input x_q that corresponds to a target output y_q that is not part of the model's prompt. We compute the model's accuracy in predicting the correct output y_q. We summarize the full set of ICL tasks we study in Appendix A.8.

To avoid leakage between ICL tasks used to identify FV heads and those used to evaluate FV head ablations, we randomly split the 37 ICL tasks from Todd et al. (2024) into 26 tasks used to measure FV scores of heads, and 11 tasks to evaluate ICL performance. We also add 8 new tasks for ICL evaluation: 4 tasks are variations of tasks in Todd et al. (2024), and 4 are binding tasks from Feng & Steinhardt (2024). In total, we evaluate ICL accuracy on 19 natural language tasks, with 100 prompts per task.

Token-loss difference. In Olsson et al. (2022), ICL performance is measured by the difference between the model loss of a token appearing early in the context (e.g., the 50th token) and later in the context (e.g., the 500th token). We also report results using this metric to compare with previous work.

We measure token-loss difference by taking the loss of the 50th token in the input prompt minus the loss of the 500th token in the prompt[3], averaged over 10,000 randomly sampled examples from the Pile dataset (Gao et al., 2021).

4.2. Results

We evaluate the impact of ablating different proportions (1-20%) of the top attention heads based on induction or FV score, across all models. We compare against two baselines: model performance with no ablation ("clean") and with ablations of randomly sampled heads ("random"). Figure 4 shows results for three representative models, where few-shot ICL accuracy is averaged over the 19 evaluation tasks. We provide comprehensive results across all models and ICL accuracy broken down by task in Appendix A.9. We plot the average induction/FV scores and percentiles of the heads preserved during ablations in Appendix A.11.

Our initial ablation experiments, shown in the top row of Figure 4, removed heads based on their scores without considering the potential overlap between induction and FV heads. These results revealed that ablating FV heads caused greater degradation in few-shot ICL performance compared to ablating induction heads, with this disparity becoming more pronounced in larger models. We also find that ablating induction heads has more effect on ICL performance than random. The effect of ablating induction heads converges to the effect of ablating FV heads as we increase the number of heads ablated.

However, the convergence noted above may be due to an increasing overlap in the set of heads ablated in the induction head and FV head ablations (Appendix A.12). To address

[2] In our main analysis, we do not rely on the correlation between the distribution of induction scores and FV scores across the full set of attention heads because there is a long tail of attention heads with low scores on both induction and FV. For completeness, we plot the induction and FV scores of all heads in Appendix A.2.

[3] We invert the difference used in Olsson et al. (2022) so higher scores indicate better ICL performance.

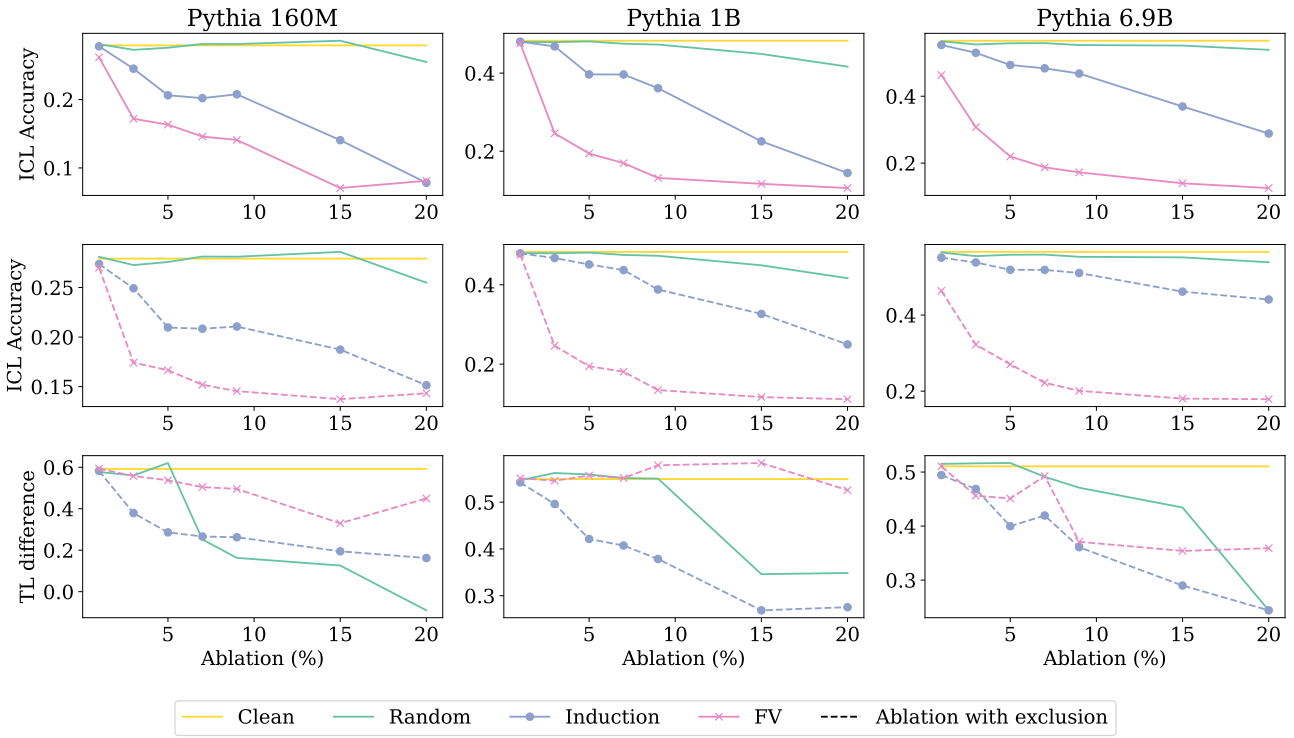


Figure 4. Top: Few-shot ICL accuracy after ablating induction and FV heads. Center: Few-shot ICL accuracy after ablating non-FV induction and non-induction FV heads. Bottom: Token-loss difference after ablating non-FV induction and non-induction FV heads. Ablating FV heads lead to a bigger drop in ICL accuracy, especially in larger models. Ablating induction heads with low FV scores does not significantly affect ICL accuracy. ICL accuracy and token-loss difference behave differently.

this, we conduct a second set of experiments using ablation with *exclusion*, shown in the center row of Figure 4.

When ablating induction heads while preserving the top 2% FV heads, we observe minimal impact on few-shot ICL performance – comparable to random ablations in models exceeding 1B parameters. Conversely, ablating FV heads while preserving induction heads continues to significantly impair ICL performance. The performance gap between FV and induction head ablations widens with model scale, suggesting that the observed effects of induction head ablations without exclusion are primarily due to heads exhibiting both induction and FV properties.

These ablations suggest that the contributions of induction heads to ICL in the top row of Figure 4 mostly come from heads that are both induction and FV heads, and that **FV heads matter the most for few-shot ICL**: as long as the model preserves its top 2% FV heads, it can perform ICL with reasonable accuracy even if we ablate induction heads.

The bottom row of Figure 4 presents the effects of ablations with exclusion on token-loss difference. In smaller models (below 160M parameters), neither ablating induction nor FV heads shows significant impact to token-loss difference compared to random ablations. However, in models with over 345M parameters, induction head ablations affect token-loss difference more than FV head ablations. This experiment primarily demonstrates that few-shot ICL accuracy and token-loss difference measure two very different things. In particular, we find instances where ICL accuracy and token-loss difference diverge: when we ablate random heads and induction heads, the model has high ICL accuracy but low token-loss difference. We speculate that this is due to the model preserving its ICL abilities (thus the high ICL accuracy), but loses other abilities that are also associated with learning important signals from context (thus the low token-loss difference). These contrasting results between the two metrics help reconcile apparently contradictory findings in existing literature.

5. FV heads evolve from induction heads

Finally, to further understand how these two families of attention heads develop, we analyze their evolution during model training. We examine attention heads across 8 intermediate training checkpoints in 7 Pythia models.

5.1. Induction and FV strength during training

To measure the general strength of induction and FV mechanisms during training, we plot the mean induction and FV scores of the top 2% induction and FV heads at each model

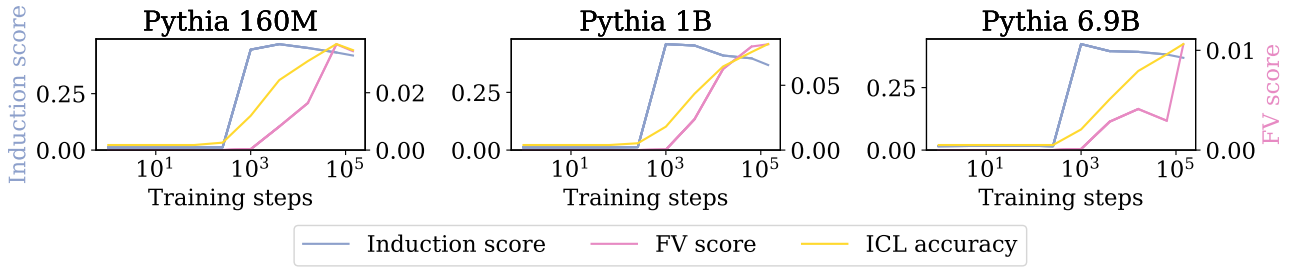


Figure 5. Evolution of induction and FV score averaged over top 2% heads across training. Induction score rises sharply, then plateaus. FV score rises slightly later and gradually increases. ICL accuracy rises around the same time as induction and gradually increases.

checkpoint, along with few-shot ICL accuracy (Figure 5). We include plots for all Pythia models in Appendix A.13.

Our analysis reveals a consistent pattern across all Pythia models: induction heads emerge early in training, at around step 1,000 out of 143,000, while FV heads appear substantially later at around step 16,000. The development of these heads shows distinct characteristics as well – induction scores exhibit a sharp initial rise followed by a plateau or slight decline, whereas FV scores demonstrate a gradual but sustained increase from step 16,000 through the end of training. This temporal asymmetry suggests that induction heads represent a simpler mechanism that models can acquire earlier, while FV heads embody a more complex mechanism that requires extended training. In addition, we observe that in all models, few-shot ICL accuracy begins to improve around the same time as when induction heads appear, and continues to gradually increase throughout training.

5.2. Evolution of individual heads during training

To gain more granular insights into head development, we investigate the evolution of individual attention heads throughout training. Figure 6 the induction scores (top row) and FV scores (bottom row) of the top 2% induction and FV heads across training steps. Individual heads are represented by continuous lines, with line opacity corresponding to their final induction or FV scores.

A striking pattern emerges across all models: many heads that ultimately become strong FV heads initially exhibit high induction scores, emerging around the same time as dedicated induction heads. These proto-FV heads initially achieve induction scores comparable to those of specialized induction heads. However, as training progresses, their induction scores gradually decline while their FV scores increase. Importantly, this pattern is unidirectional; we found no instances of induction heads that develop significant FV capabilities during training, as evidenced by their consistently low FV scores throughout training. This suggests **many FV heads evolve from induction heads during training**, but not vice versa.

6. Interpretation and discussion

Our investigation revealed several key insights about the relationship between induction and FV heads in transformer models and their effect on ICL. While these mechanisms are distinct, they show notable correlation (§3). FV heads consistently appear in deeper layers and emerge later in training compared to induction heads (§3,5.1). Our ablation studies demonstrated that FV heads are crucial for few-shot ICL performance, particularly in larger models, while induction heads have comparatively minimal impact (§4). Furthermore, we observed multiple instances of heads transitioning from induction to FV functionality during training, but never the reverse (§5.1). We propose two working conjectures to explain these empirical findings more broadly, and consider arguments for and against them.

Our first conjecture (C1) posits that **induction heads serve as precursots to FV heads**. Under this interpretation, induction heads serve as a stepping stone for models to develop the more sophisticated FV mechanism. The mechanism underlying induction heads is simpler and easier to learn, but this method does not fully solve more complex ICL problems. As FV heads emerge and prove to be more effective at difficult ICL tasks, they gradually supersede the simpler induction mechanism.

Several lines of evidence support this conjecture. First, we observed multiple FV heads that initially displayed strong induction behavior before transitioning to FV functionality. The subsequent decline in induction scores suggests these heads abandon the simpler mechanism once the more effective FV capability develops. The unidirectional nature of this transition - we never observe induction heads with initially high FV scores - further supports this interpretation. Additionally, the minimal effect of ablating pure induction heads (those with low FV scores) on few-shot ICL performance suggests their role becomes less critical once FV heads develop. To further verify this, future work could explore how removing induction heads during training could impact the development of FV heads. However, C1 does not fully explain the existence of FV heads with low induction

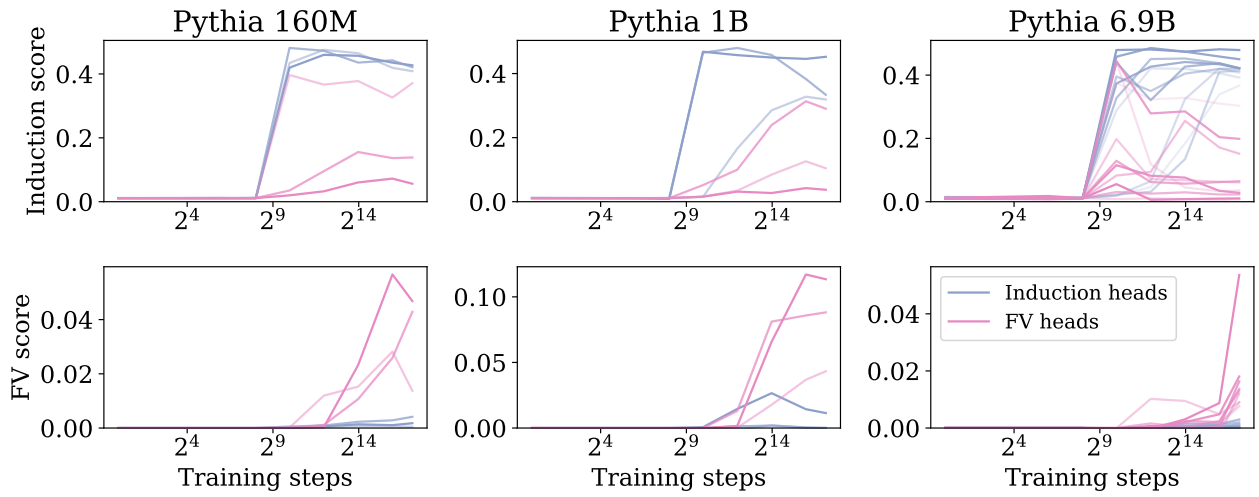


Figure 6. Evolution of induction scores (top) and FV scores (bottom) of individual induction and FV heads across training. Certain FV heads have a high induction score earlier in training; the reverse is not true for induction heads.

scores throughout training.

C1 also aligns with our observations about architectural and training dynamics. FV heads consistently emerge later in training and appear in deeper layers, consistent with them implementing a more complex computation. Furthermore, we observe that in all models, few-shot ICL accuracy begins to rise around the same time as when induction heads appear, and continues to gradually increase until the end of training. Since ICL accuracy continues to improve even after the formation of induction heads, we speculate that the sharp emergence of induction heads contributes to an initial rise in ICL performance, but the gradual formation of FV heads drives the further improvements in ICL. This reinforces our conjecture that induction heads may serve as a stepping stone for FV

An alternative conjecture (C2) suggests that **FV heads are a combination of induction and another mechanism**. This conjecture proposes that heads that appear to "transition" from induction to FV functionality may actually be polysemanticheads, implementing both mechanisms and possibly others. Their measured induction scores decline as their attention patterns diversify to support multiple mechanisms.

While C2 explains the correlation between induction and FV mechanisms as arising from shared underlying mechanisms, it faces a significant challenge: our ablation studies show that removing monosemantic FV heads (those without significant induction scores) substantially hurts ICL performance. This is difficult to reconcile with C2's prediction that pure FV heads should be less critical if the key functionality depends on combined induction-FV mechanisms.

7. Related work

In this paper, we compared two mechanisms that have been previously proposed to explain ICL: induction heads (Elhage et al., 2021; Olsson et al., 2022) and FV heads (Todd et al., 2024; Hendel et al., 2023). A line of previous work have attributed the mechanism behind ICL to induction heads. Olsson et al. (2022) studied small attention-only models with 1-3 layers and showed that the emergence of induction heads co-occur with the emergence of ICL performance. We observe a similar co-occurence in §5.1. They also showed that ablating induction heads decreases token-loss difference, similarly to our results in §4.2. Singh et al. (2024) expanded on this by introducing a causal framework to manipulate activations in toy models, and identified three subcircuits that drives induction head formation, and consequently, the phase transition in ICL.

Crosbie & Shutova (2024) performed ablation studies on LLMs of size 8B and 20B parameters, and Bansal et al. (2023) on an LLM of size 66B, to demonstrate that induction heads are important for ICL in large models as well. We observe similar results in the top row of Figure 4, where we performed ablations without accounting for the overlap between induction and FV heads.

Further works explored the ICL capabilities and internal mechanisms of transformer models through multiple complementary lenses. Edelman et al. (2024) demonstrated how models develop statistical induction heads for Markov chain prediction. Xie et al. (2022) frames ICL as implicit Bayesian inference when pretrainng data exhibits latent coherence. Garg et al. (2022) empirically proved transformer models can learn linear functions and complex classes in-context.

Guo et al. (2024) extended this to compositional tasks and observed mechanisms such as copying behaviors within transformers.

Reconciling divergent findings

While induction heads and FV heads have been proposed by their respective works as the mechanism behind ICL, our side-by-side analysis of induction and FV heads reveals that FV heads seem to primarily contribute to ICL performance. We believe that the main reason for the divergence between our result and previous work lies in several intuitively related concepts in the literature that are assumed to be the same.

First, ICL is used synonymously with few-shot learning in most of the literature, while other works conceptualize ICL as loss reduction in tokens appearing later in a sequence. The two distinct definitions of ICL have often been conflated to refer to the same phenomenon. However, our experiments reveal that the two ICL metrics capture different phenomena: FV heads strongly influence few-shot ICL accuracy but not token-loss difference, while induction heads show the opposite pattern (§4). This divergence of token-loss difference from few-shot ICL performance accounts for much of the apparent contradiction with previous findings.

Second, we find that induction heads and FV heads are correlated (§3), a confound not previously controlled for. Initial ablation studies, including ours (§4), showed that removing induction heads significantly degrades few-shot ICL accuracy (Crosbie & Shutova, 2024; Bansal et al., 2023). However, when we control for this correlation by only ablating induction heads with low FV scores, their impact becomes comparable to random ablation. In contrast, ablating FV heads with low induction scores still significantly degrades ICL performance. This suggests that previous studies may have attributed to induction heads effects that actually stemmed from FV-like behavior in a subset of induction heads.

Third, scale matters. Previous work establishing induction heads as the key ICL mechanism (Olsson et al., 2022; Singh et al., 2024) focused on small models to enable detailed mechanistic analysis. However, we find that the relative importance of FV heads increases with model scale. In our smallest model (70M parameters), induction and FV heads have similar causal effects on few-shot ICL (§A.3), but this does not hold true in larger models, which highlights the importance of studying these phenomena across different model scales.

8. Conclusion

Our research challenges the prevailing understanding of in-context learning mechanisms in transformer models. While induction heads have been widely considered the primary driver of ICL, our evidence demonstrates that function vector (FV) heads play a more crucial causal role in few-shot ICL performance. We attribute previous misconceptions to two key factors: the conflation of few-shot ICL with token-loss difference, and not accounting for the overlap between induction and FV heads.

Remarkably, although induction and FV mechanisms appear to implement two distinct processes, we also observe an interesting interplay between the two types of heads: induction and FV scores are correlated, and many FV heads are "former" induction heads with high induction scores earlier in training. This observation supports the conjecture that induction heads serve as precursors to FV heads: the simpler induction mechanism provides an initial foundation for ICL, from which the more sophisticated FV mechanism eventually emerges.

Our investigation also yields important methodological insights for the broader field of model interpretability. First, seemingly equivalent definitions of model capabilities (such as few-shot accuracy versus token-loss difference) can lead to substantially different conclusions. Second, studying mechanistic components in isolation may produce misleading results when these components share overlapping behaviors, as demonstrated by the confounding effects of ablating heads that exhibit both high induction and FV scores. We thereby recommend future works to carefully define the type of ICL studied, and consider the interactions between different circuits.

Furthermore, our results challenge strong versions of the universality hypothesis in interpretability. While both induction and FV heads contribute meaningfully to few-shot ICL in smaller models, their relative importance diverges with scale – FV heads become increasingly crucial while induction heads' impact approaches that of random ablations. This scale-dependent behavior suggests that mechanisms may vary across model architectures.

These findings prompt several important questions for future research. If induction heads indeed serve as precursors to FV heads, what makes this necessary? What role do the remaining induction heads serve in fully trained models? Are there additional mechanisms that provide an even more complete explanation of ICL capabilities? We also leave practical applications of our findings to future work – evidence of the higher importance of FV suggests that ICL could be optimized in smaller models by using training methods or architectures that promote the formation of FV heads.

Acknowledgements

We thank Jiahai Feng, Neel Nanda, Robert Kirk, and Anish Kachinthaya for their helpful feedback, and anonymous

reviewers for useful comments. KY is supported by the Vitalik Buterin Ph.D. Fellowship in AI Existential Safety.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Bansal, H., Gopalakrishnan, K., Dingliwal, S., Bodapati, S., Kirchhoff, K., and Roth, D. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, 2023.

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 Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Crosbie, J. and Shutova, E. Induction heads as an essential mechanism for pattern matching in in-context learning. *ArXiv preprint*, 2024.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey on in-context learning. *ArXiv preprint*, 2023.

Edelman, E., Tsilivis, N., Edelman, B. L., Malach, E., and Goel, S. The evolution of statistical induction heads: In-context learning markov chains. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Feng, J. and Steinhardt, J. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800gb dataset of diverse text for language modeling. *ArXiv preprint*, 2021.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Guo, T., Hu, W., Mei, S., Wang, H., Xiong, C., Savarese, S., and Bai, Y. How do transformers learn in-context beyond simple functions? A case study on learning with representations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Hase, P., Xie, H., and Bansal, M. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *ArXiv preprint*, 2020.

Lampinen, A. K., Chan, S. C. Y., Singh, A. K., and Shanahan, M. The broader spectrum of in-context learning, 2024.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

McGrath, T., Rahtz, M., Kramar, J., Mikulik, V., and Legg, S. The hydra effect: Emergent self-repair in language model computations, 2023.

Nanda, N. and Bloom, J. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`, 2022.

Nguyen, K. A., Schulte im Walde, S., and Vu, N. T. Distinguishing antonyms and synonyms in a pattern-based neural network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *ArXiv preprint*, 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C. Y., and Saxe, A. M. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Tjong Kim Sang, E. F. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, 2023.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, 2023.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. Larger language models do in-context learning differently. *ArXiv preprint*, 2023.

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., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015.

A. Appendix

A.1. Limitations of our work

Our findings are limited to models under 7B parameters, and we leave scaling our analysis to larger models to future work. The range of models we study also limits the complexity of the ICL tasks we studied, as we selected tasks where models have reasonable few-shot performance. To ensure that our findings apply to real-world tasks, we included realistic tasks such as translation, commonsense question answering, news topic classification. Our ablation studies also do not account for the hydra effect (McGrath et al., 2023), as measuring the extent of self-repair when ablating specific heads remains an open problem (the method in McGrath et al. (2023) applies to ablating all heads in a single layer). If the same magnitude of the hydra effect is present in our ablations of different head types, our findings on the relative importance of heads would still hold, however this is not yet proven. We leave to future work to measure the extent of second-order effects in ablations.

A.2. Induction scores vs. FV scores

In Figure 7, we plot the induction score and FV score of each attention head. We also color the heads that are determined as an induction head, an FV heads, or both, using the top 2% for each score as the threshold. The plots show that the 2% threshold meaningfully select important heads from the cluster of other low-scoring heads across models. Taking a higher percentage for the threshold would select heads with FV scores close to 0.

A.3. Ablations

In Figure 8, we plot model accuracy averaged over ICL tasks across different quantities of heads ablated in each head type. In Figure 9, we plot the token-loss difference of models across different quantities of heads ablated.

A.4. Random and zero ablations

In Figure 10, we plot model accuracy averaged over ICL tasks across different quantities of heads ablated with random ablation or zero ablation. For random ablations, we replace the head's output vector with the output vector of a randomly sampled different head. For zero ablations. we replace the head's output vector with a zero vector.

A.5. Ablating random heads at specific layers

In Figure 11, we ablate heads randomly sampled from specific layers of the model. Let L be the number of heads in each layer, A be the number of heads we're ablating, and ℓ be the layer we're targeting. Then, if $A < L$, we sample A heads from layer ℓ. If $A \geq L$, we ablate all L heads in layer ℓ and we sample $A - L$ heads from other layers to ablate.

A.6. Induction and function vector scores across models

Our ablation studies reveal a consistent trend where FV heads are increasingly important relative to induction heads for ICL performance as model scale increases. To further explore this trend, we examine how induction scores and FV scores vary with model scale, and whether these scores follow similar trends to our ablation experiments.

In Figure 12, we plot the maximum and mean induction and FV scores across all heads, and mean scores of top 2% heads, for each model. The left plot in Figure 12 shows that induction scores are relatively similar across model size, with a small increase in maximum induction score and a decrease in the top 2% mean induction score with model scale.

In the right plot of Figure 12, there is no clear trend between FV score and model scale, however, Pythia 1B and 1.4B models have markedly higher maximum FV scores. One possible explanation is that models with high head dimensionality relative to total parameter count have stronger FV heads: Pythia 1B and 1.4B have head dimensionality of 256 and 128 respectively (Table 2) whereas other models with similar parameter count have only 64-80 attention head dimensions.

We also find very low FV scores in Pythia 70M and Llama 2 models. FV scores may be low in Pythia 70M because it is too small in parameter size for FV heads to emerge. Low scores in Llama 2 compared to other models may be due to differences in architecture, and additional experiments can help confirm this. Overall, we do not recover the same trend in induction/FV scores as the trend in our ablation studies.

For reference, we also provide box plots of the full distribution of induction and FV scores in Figure 13.

A.7. Evaluating function vectors on task execution

To further inspect the prevalence of the FV mechanism in different models, we evaluate the efficacy of FVs for ICL task execution. A successful FV triggers the model to execute the particular task the FV encodes, even when the model sees no useful in-context demonstrations of the task. First, to extract FVs, for each model we gather the top 2% attention heads with highest FV scores as the set $\mathcal{A}$. Then, for each ICL task $t \in \mathcal{T}$, we sum the average outputs of heads in $\mathcal{A}$ over prompts from t and obtain the FV for the task t: $FV_t = \sum_{a \in \mathcal{A}} \bar{a}^t$.

In Figure 14, we report model accuracy averaged over 40 ICL tasks where the model performs inference on uncorrupted prompts (clean), prompts with shuffled labels (shuffled), shuffled prompts with FV_t added to hidden states at

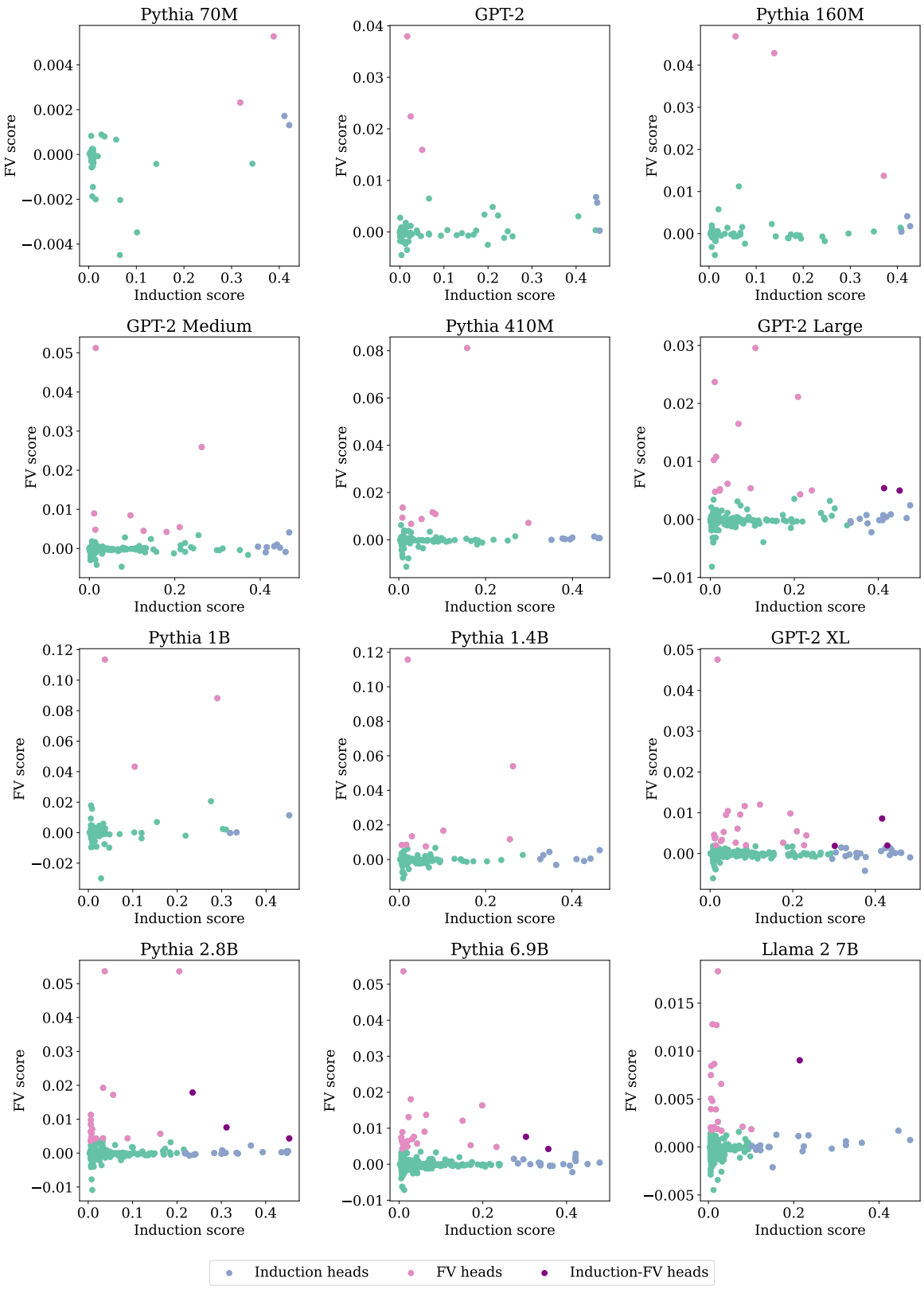


Figure 7. Induction and FV scores of attention heads.

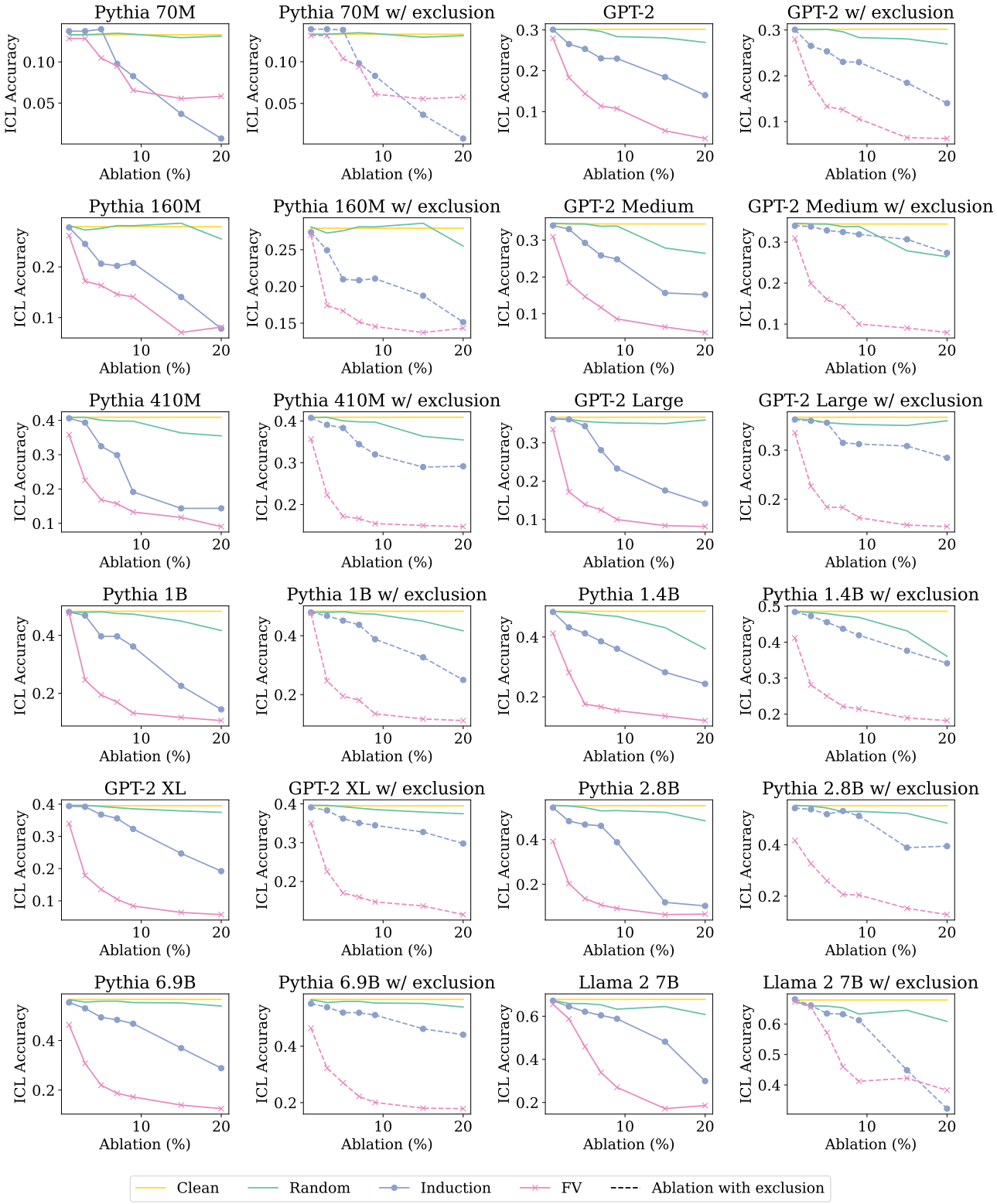


Figure 8. ICL accuracy after ablating induction and FV heads.

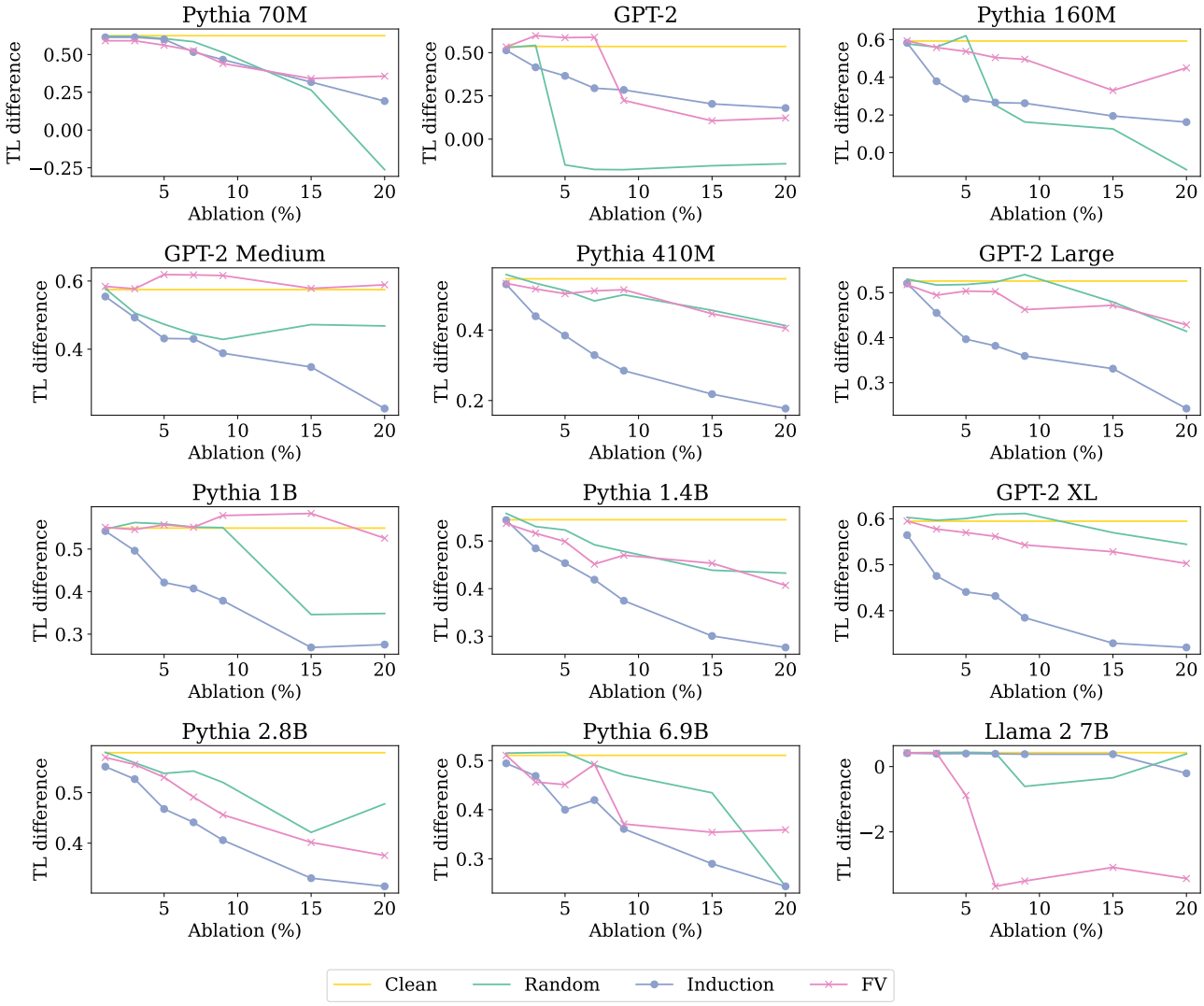


Figure 9. Token-loss difference after ablating induction heads with low FV scores and FV heads with low induction scores.

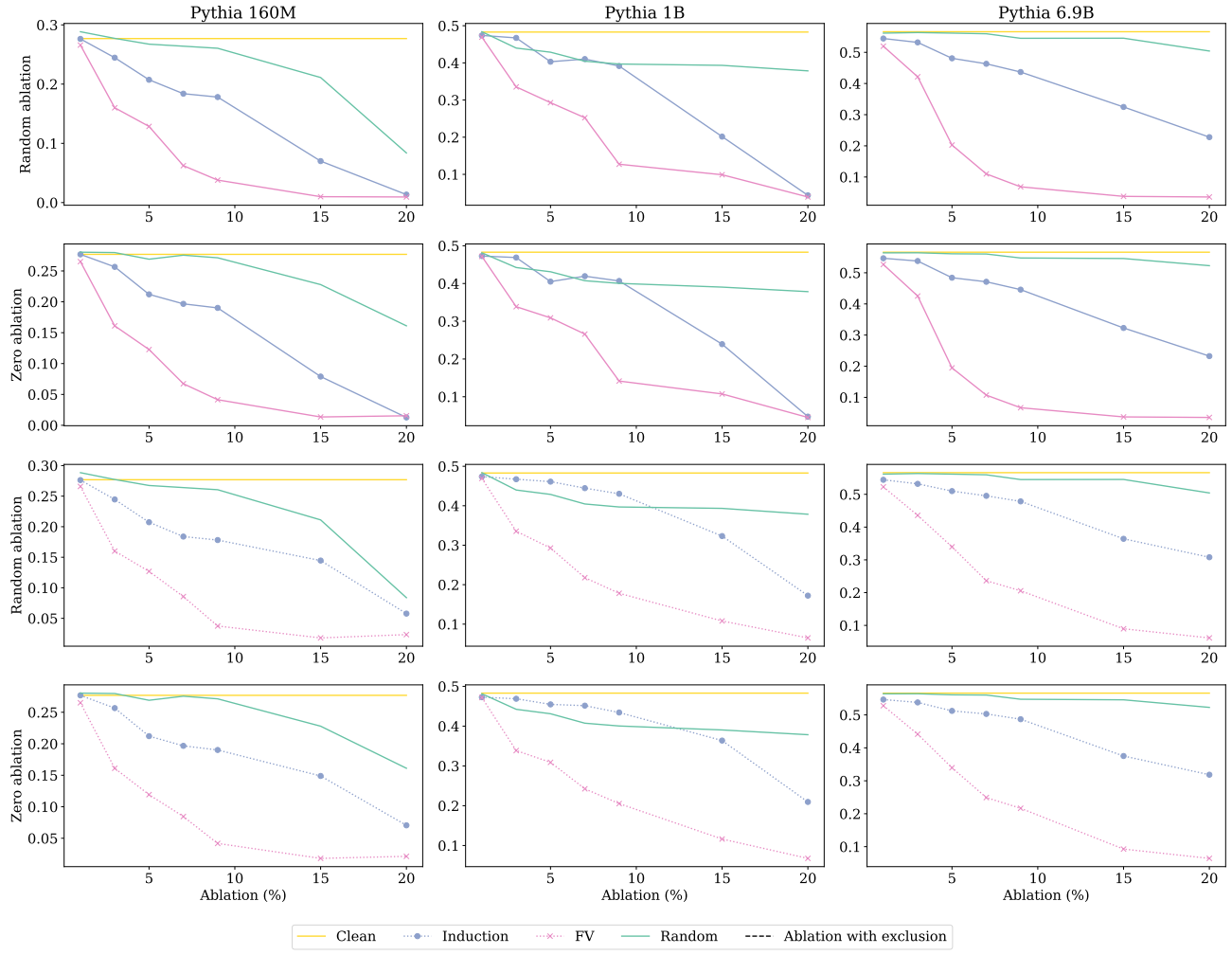


Figure 10. Few-shot ICL accuracy after ablating induction and FV heads, using either random ablation method (rows 1 and 3) or zero ablation method (rows 2 and 4). Overall, the observation that ablating FV heads decreases ICL accuracy more than induction heads, is robust against different methods of ablation.

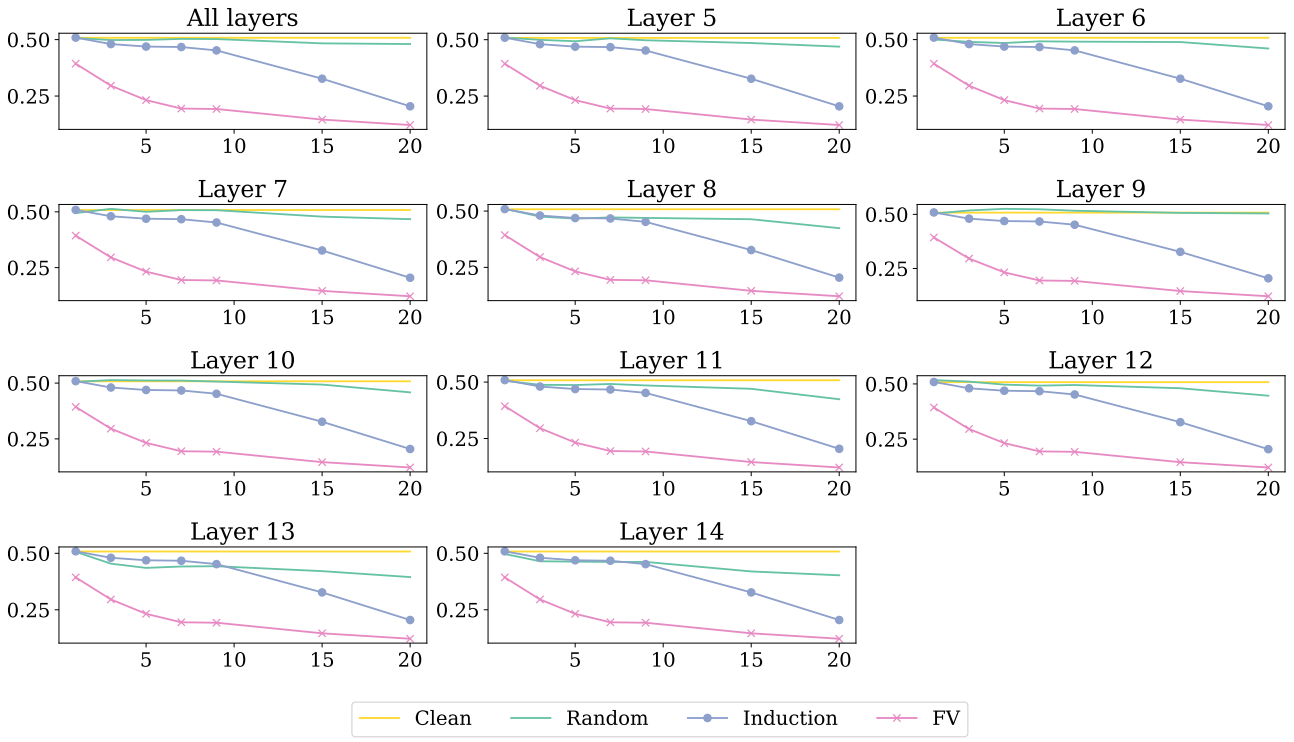


Figure 11. ICL accuracy after ablating randomly sampled heads from specific layers. The clean ICL accuracy, induction ablations and FV ablations are also plotted for comparison but only the random ablations (green curve) are affected by the choice of target layer.

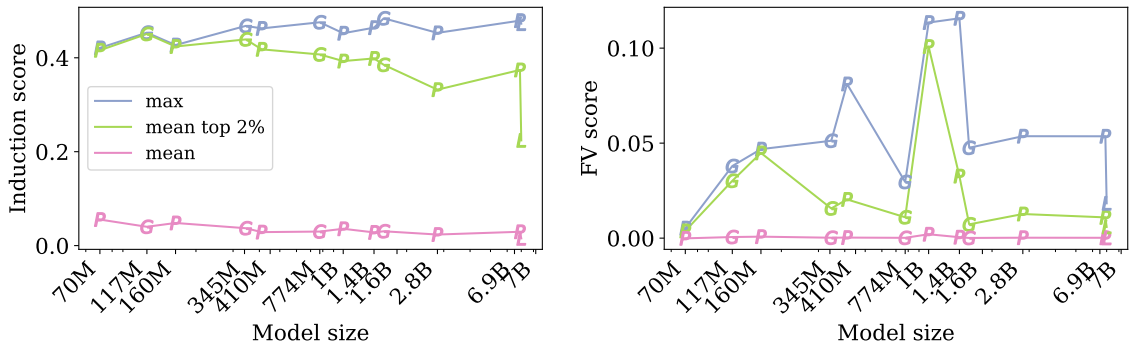


Figure 12. Induction score (left) and FV score (right) of attention heads across model size. We plot the maximum score of all heads, mean of the top 2% scores, and mean score of all heads. Overall, induction scores are similar across models. Pythia 70M and Llama 2 have relatively low FV scores, Pythia 1B and 1.4B have relatively high FV scores.

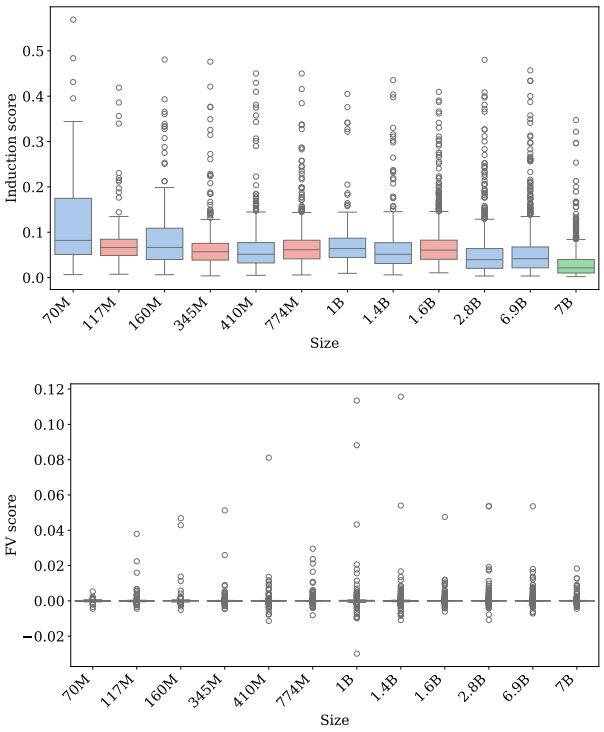


Figure 13. Distribution of induction scores (top) and FV scores (bottom) across model size.

layer $|L|/3$, and shuffled prompts with FV extracted from random heads added to hidden states at layer $|L|/3$. We take 1000 examples per task that are previously unseen during FV score computation.

In most models, adding the FV recovers model performance on uncorrupted prompts, with the exception of Pythia 2.8B. One possible explanation for this is again due to head dimensionality: Pythia 2.8B has head dimension 80, which is significantly smaller than other models with similar parameter size that have head dimensions of 128. Together with our experiments in §A.6, results provide preliminary evidence that **high head dimensionality relative to model size is a predictor of FV strength (H6)**.

A.8. ICL tasks

In Table 3, we list the ICL tasks used in this study. We refer to Todd et al. (2024) and Feng & Steinhardt (2024) for a detailed description of each task.

A.9. Ablations by task

In Figures 15-18, we plot the ICL accuracy after ablating induction heads and FV heads for each task in the evaluation set. We also compute the random baseline for each task,

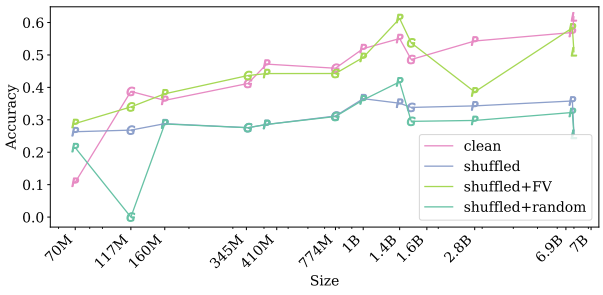


Figure 14. Model ICL accuracy on prompts with 10 in-context examples (clean), on uninformative shuffled prompts, on shuffled prompts with FV, and on shuffled prompts with random head outputs. Adding FV recovers most of the model accuracy on a clean run, with the exception of Pythia 2.8B.

where we randomly sample outputs seen during training and compare these random outputs to the ground truth. The random baselines are shown in red horizontal lines.

A.10. Head locations

In Figure 19, we plot the locations of induction heads and FV heads across model layers.

A.11. Induction and FV scores of heads after ablation

In Figures 20 – 27, we plot the induction scores, percentile of induction scores, FV scores, and percentile of FV scores of the heads ablated in §4, for both ablation and ablation with exclusion.

A.12. Overlap between ablated induction and FV heads

In Figure 28, we plot the percentage of attention heads that overlap between the set of induction heads and FV heads we ablate. We find that as the number of ablated heads increases, the overlap between the two sets of ablated heads also increases. This demonstrates the importance of performing ablations with exclusion to control for overlap.

A.13. Scores across training

In Figure 29, we plot the evolution of induction and FV scores averaged over top 2% heads across model training, along with the few-shot accuracy of the model checkpoints averaged over the evaluation tasks. In Figure 30, we plot the evolution of induction and FV scores of individual heads across training.

Table 3. Summary of ICL tasks used in our study. Tasks in **bold** are new tasks that were not used in (Todd et al., 2024).

Abstractive Tasks	**Extractive Tasks**
Abstract clf (Ours.)	Adjective vs. verb (Todd et al., 2024)
Antonym (Nguyen et al., 2017)	Animal vs. object (Todd et al., 2024)
Binding capital (Feng & Steinhardt, 2024)	Choose first of list (Todd et al., 2024)
Binding capital parallel (Feng & Steinhardt, 2024)	Choose middle of list (Todd et al., 2024)
Binding fruit (Feng & Steinhardt, 2024)	Choose last of list (Todd et al., 2024)
Binding shape (Feng & Steinhardt, 2024)	Color vs. animal (Todd et al., 2024)
Capitalize first letter (Nguyen et al., 2017)	Concept vs. object (Todd et al., 2024)
Capitalize index (Ours.)	Fruit vs. animal (Todd et al., 2024)
Capitalize second letter (Ours.)	Object vs. concept (Todd et al., 2024)
Capitalize (Nguyen et al., 2017)	Verb vs. adjective (Todd et al., 2024)
Country-capital (Todd et al., 2024)	CoNLL-2003, NER-person (Tjong Kim Sang, 2002)
Country-currency (Todd et al., 2024)	CoNLL-2003, NER-location (Tjong Kim Sang, 2002)
English-French (Lample et al., 2018)	CoNLL-2003, NER-organization (Tjong Kim Sang, 2002)
English-German (Lample et al., 2018)	
English-Spanish (Lample et al., 2018)	
French-English (Lample et al., 2018)	
Landmark-Country (Hernandez et al., 2024)	
Lowercase first letter (Todd et al., 2024)	
National parks (Todd et al., 2024)	
Next-item (Todd et al., 2024)	
Previous-item (Todd et al., 2024)	
Park-country (Todd et al., 2024)	
Person-instrument (Hernandez et al., 2024)	
Person-occupation (Hernandez et al., 2024)	
Person-sport (Hernandez et al., 2024)	
Present-past (Todd et al., 2024)	
Product-company (Hernandez et al., 2024)	
Singular-plural (Todd et al., 2024)	
Synonym (Nguyen et al., 2017)	
CommonsenseQA (MC-QA) (Talmor et al., 2019)	
Sentiment analysis (SST-2) (Socher et al., 2013)	
AG News (Zhang et al., 2015)	

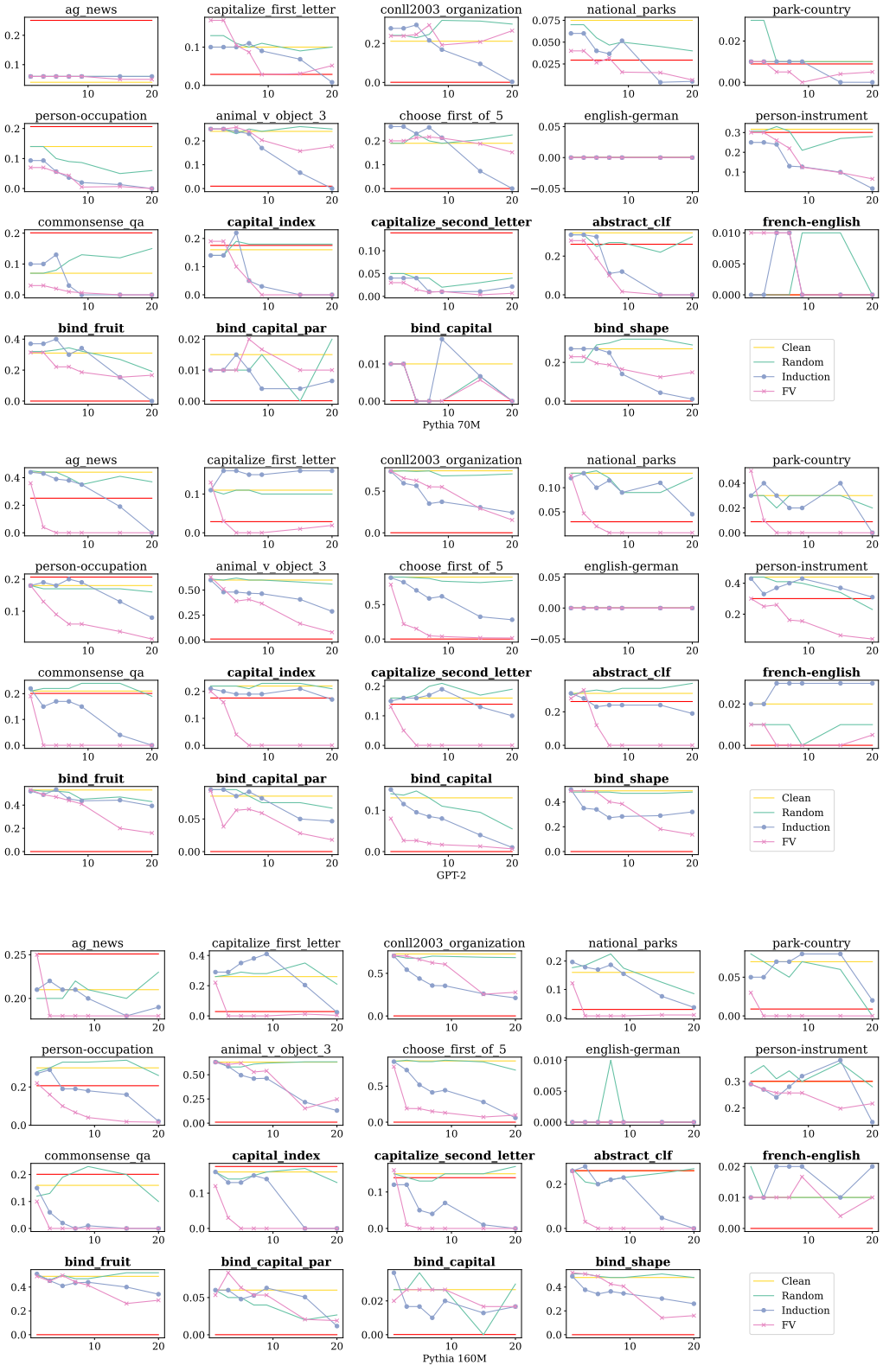


Figure 15. ICL accuracy after ablations by task. The red horizontal line represents the random baseline.

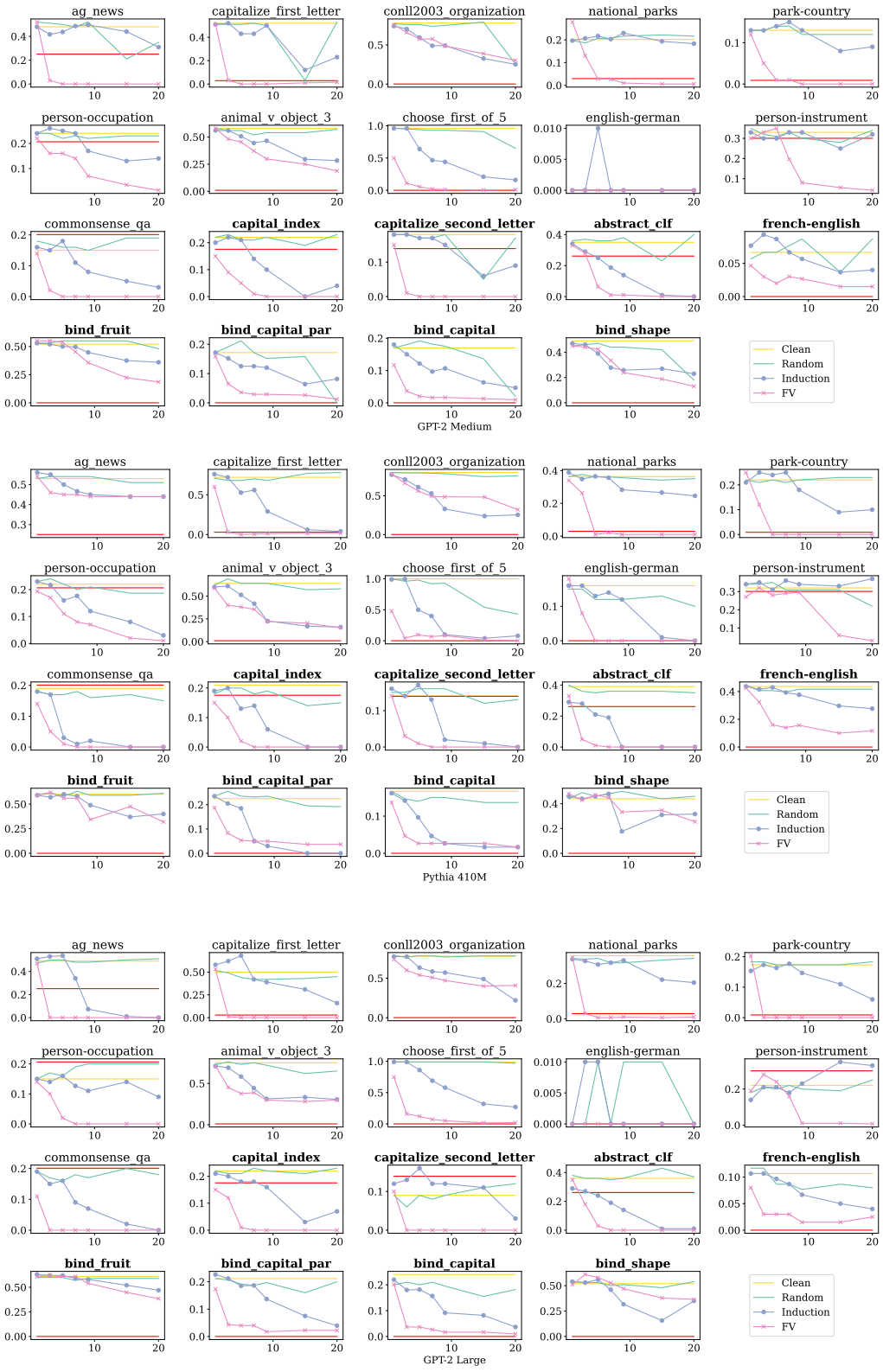


Figure 16. ICL accuracy after ablations by task. The red horizontal line represents the random baseline.

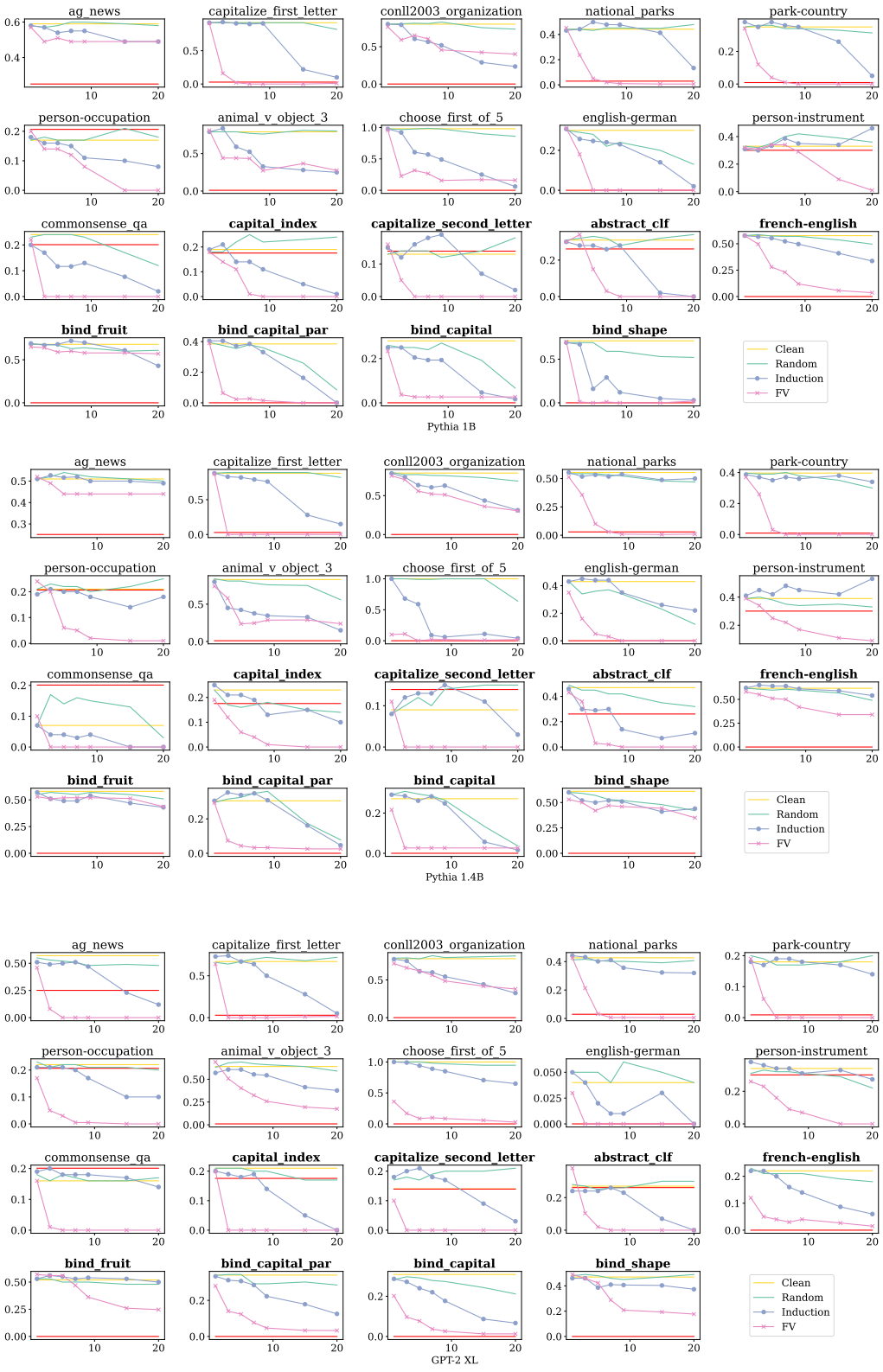


Figure 17. ICL accuracy after ablations by task. The red horizontal line represents the random baseline.

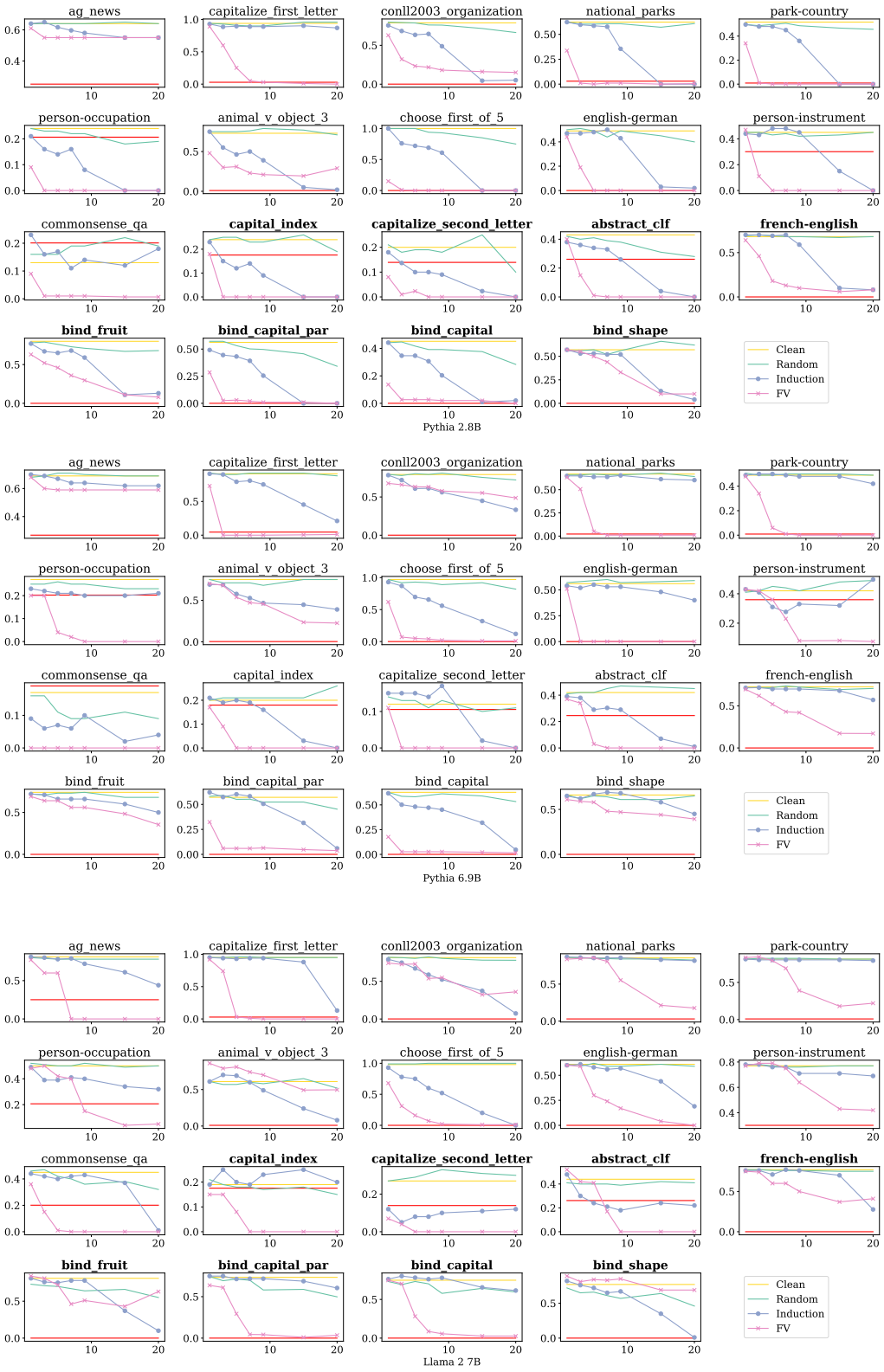


Figure 18. ICL accuracy after ablations by task. The red horizontal line represents the random baseline.

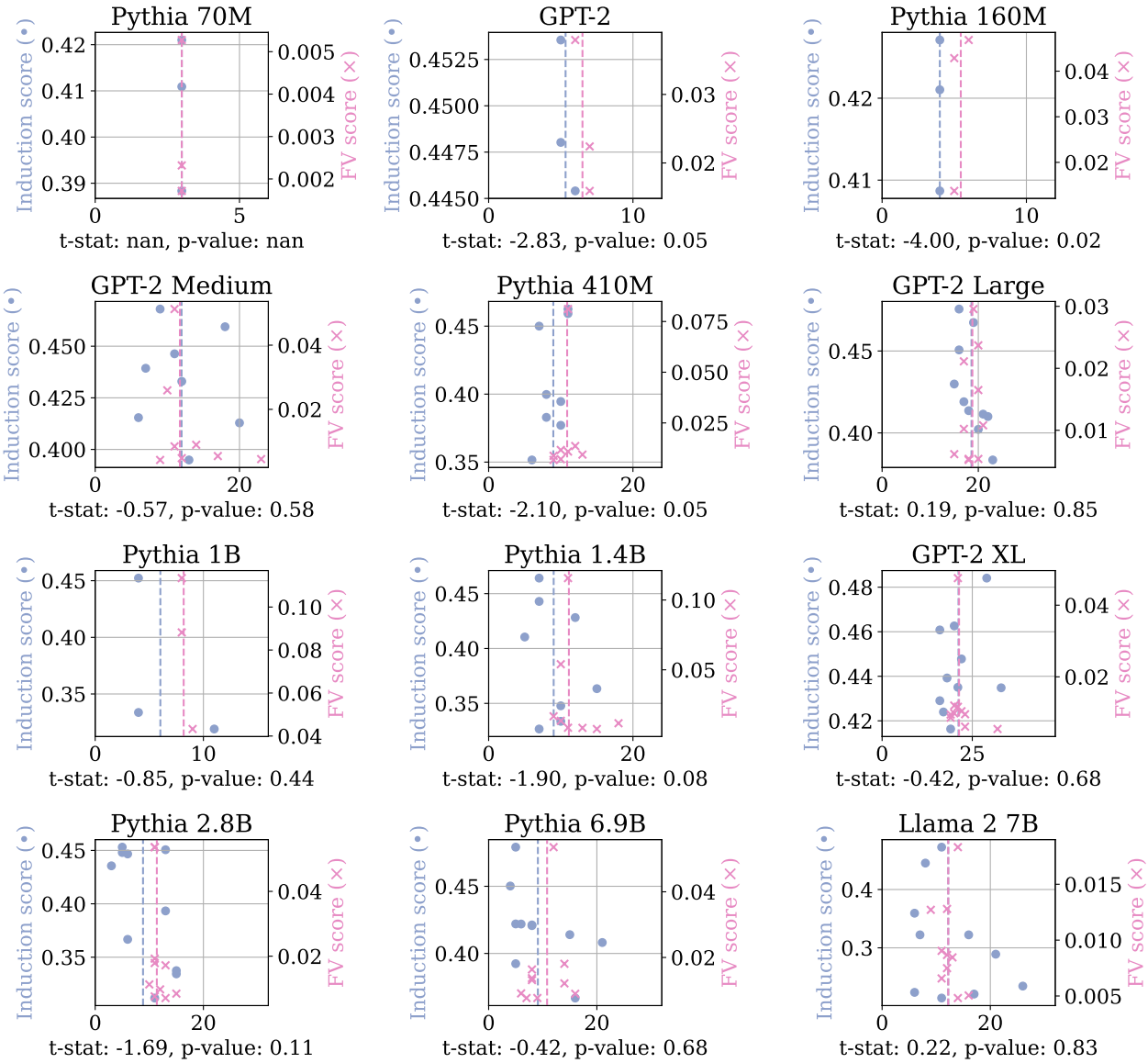


Figure 19. Location of induction heads (blue) and FV heads (pink) in model layers

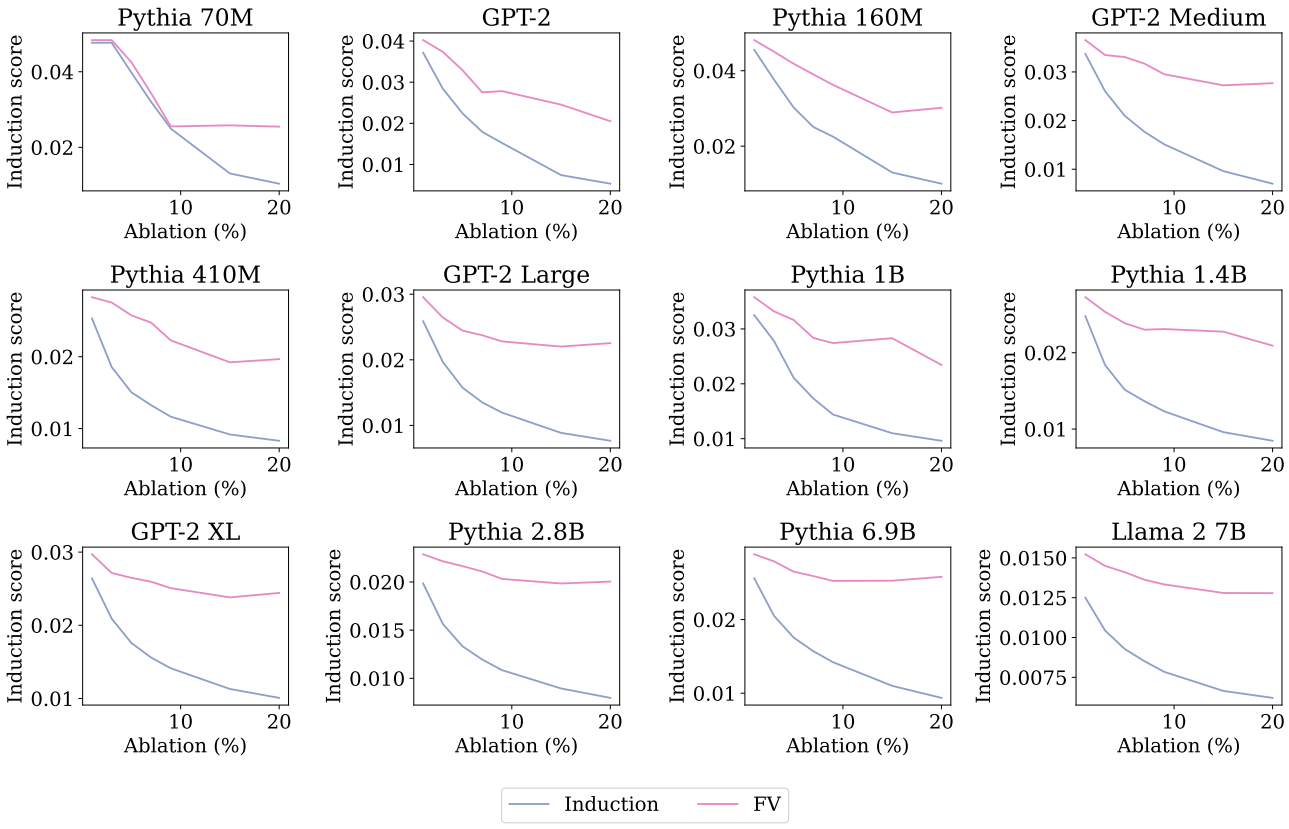


Figure 20. Induction scores of the remaining heads after ablation.

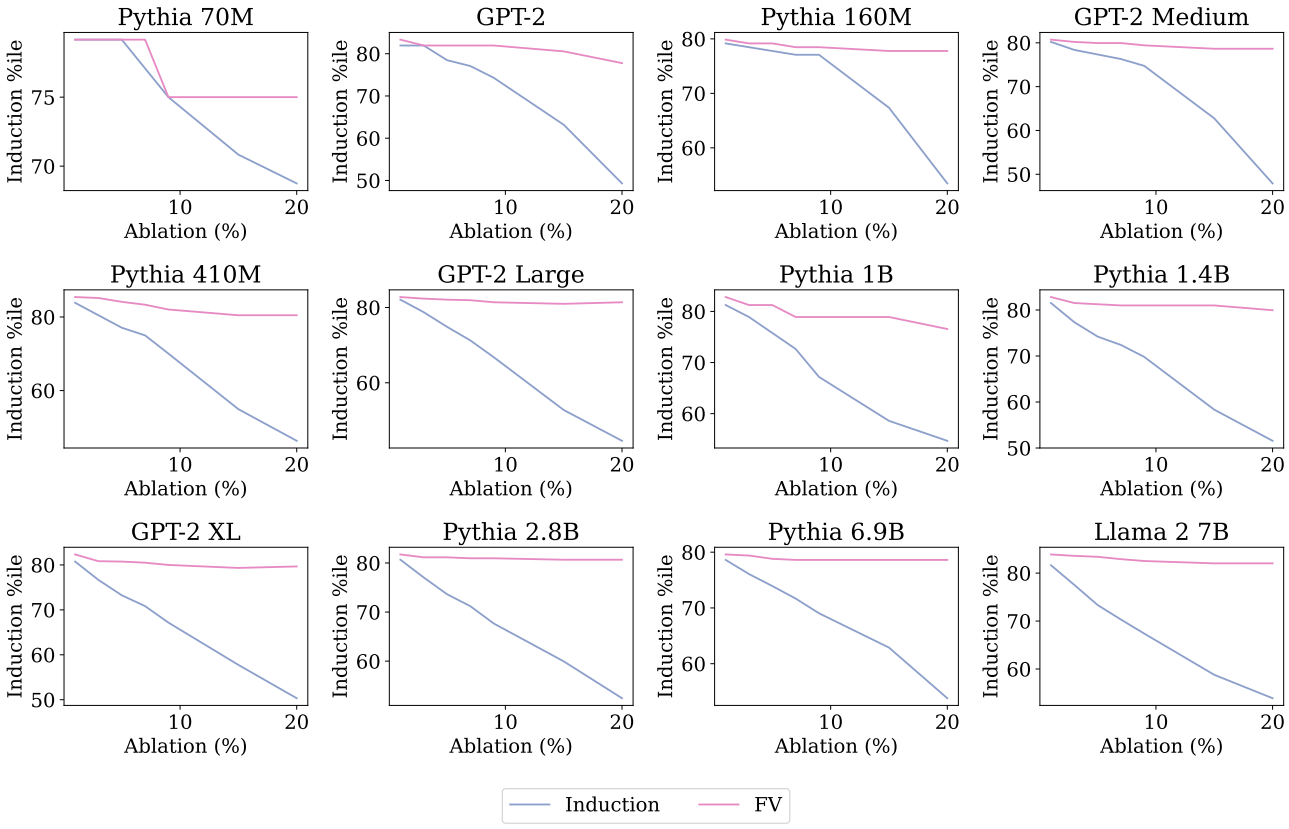


Figure 21. Percentile of induction scores of the remaining heads after ablation.

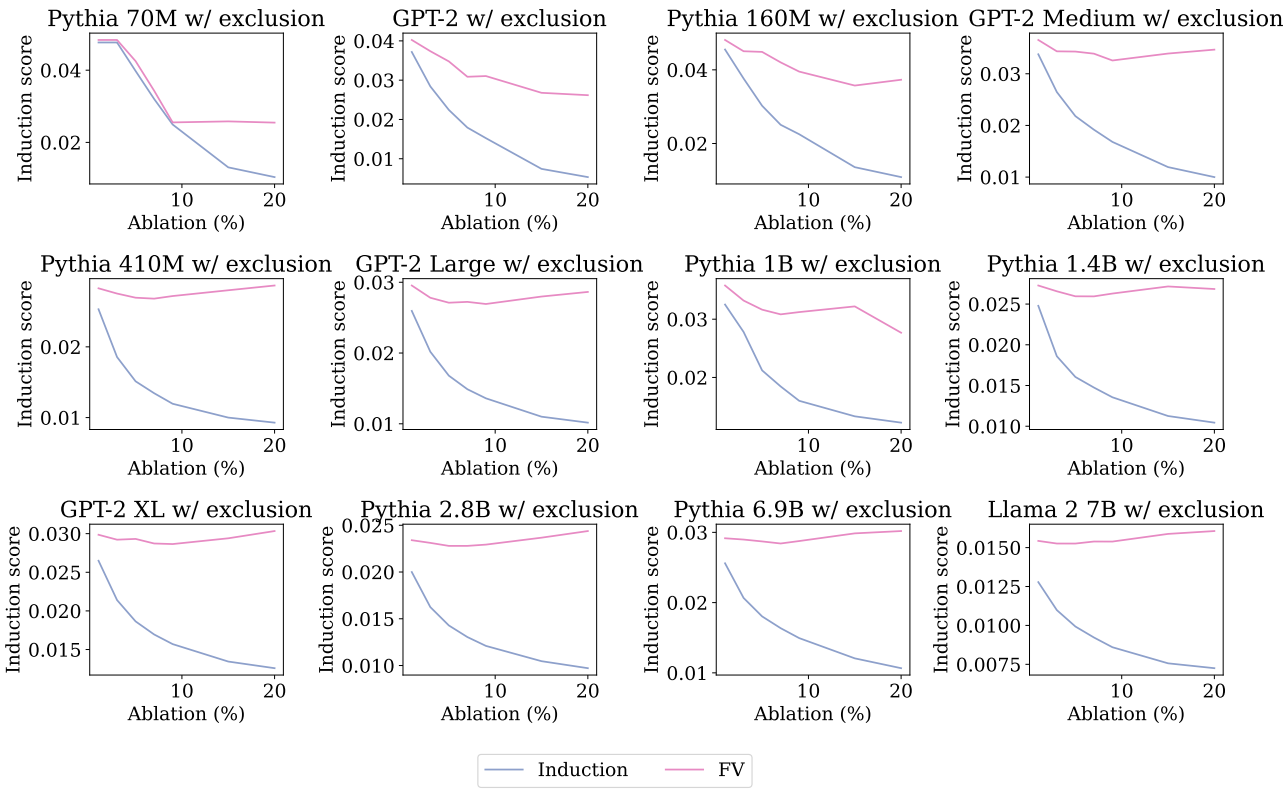


Figure 22. Induction scores of the remaining heads after ablation with exclusion.

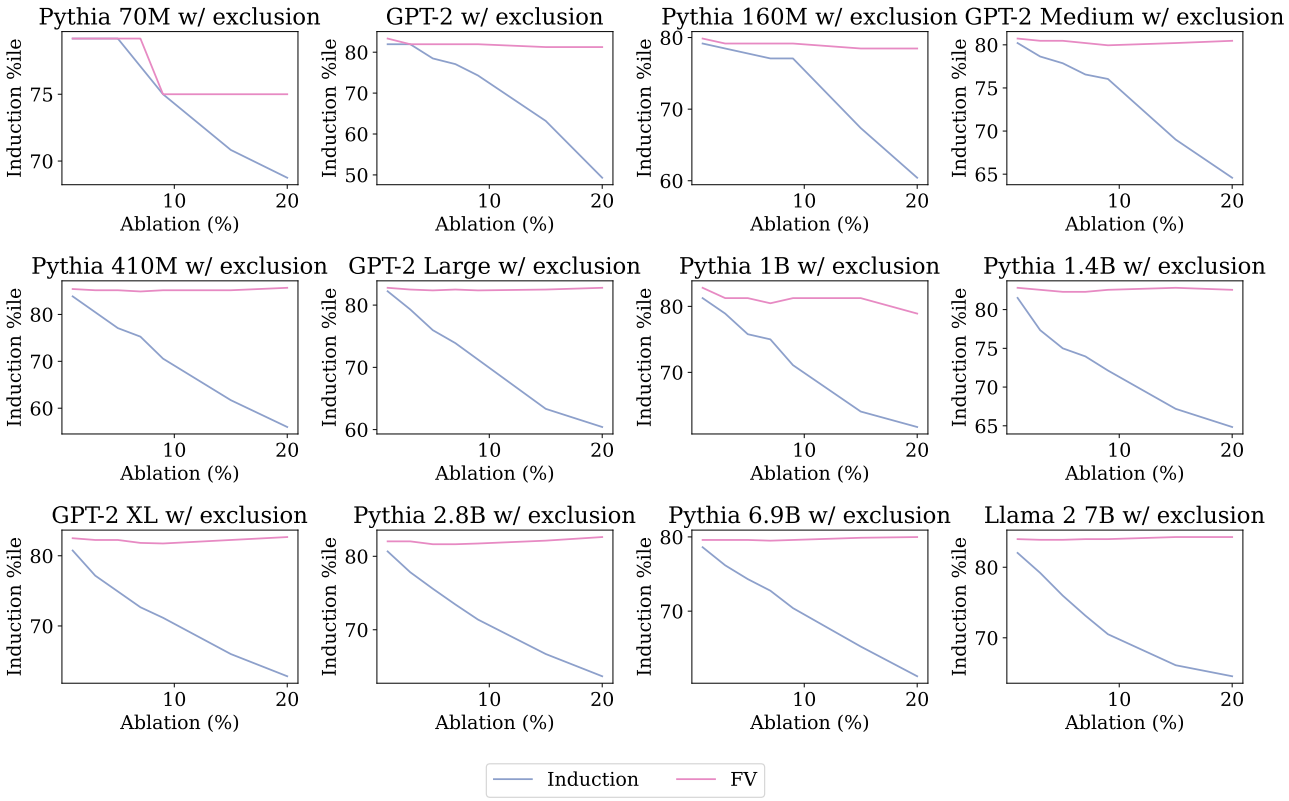


Figure 23. Percentile of induction scores of the remaining heads after ablation with exclusion.

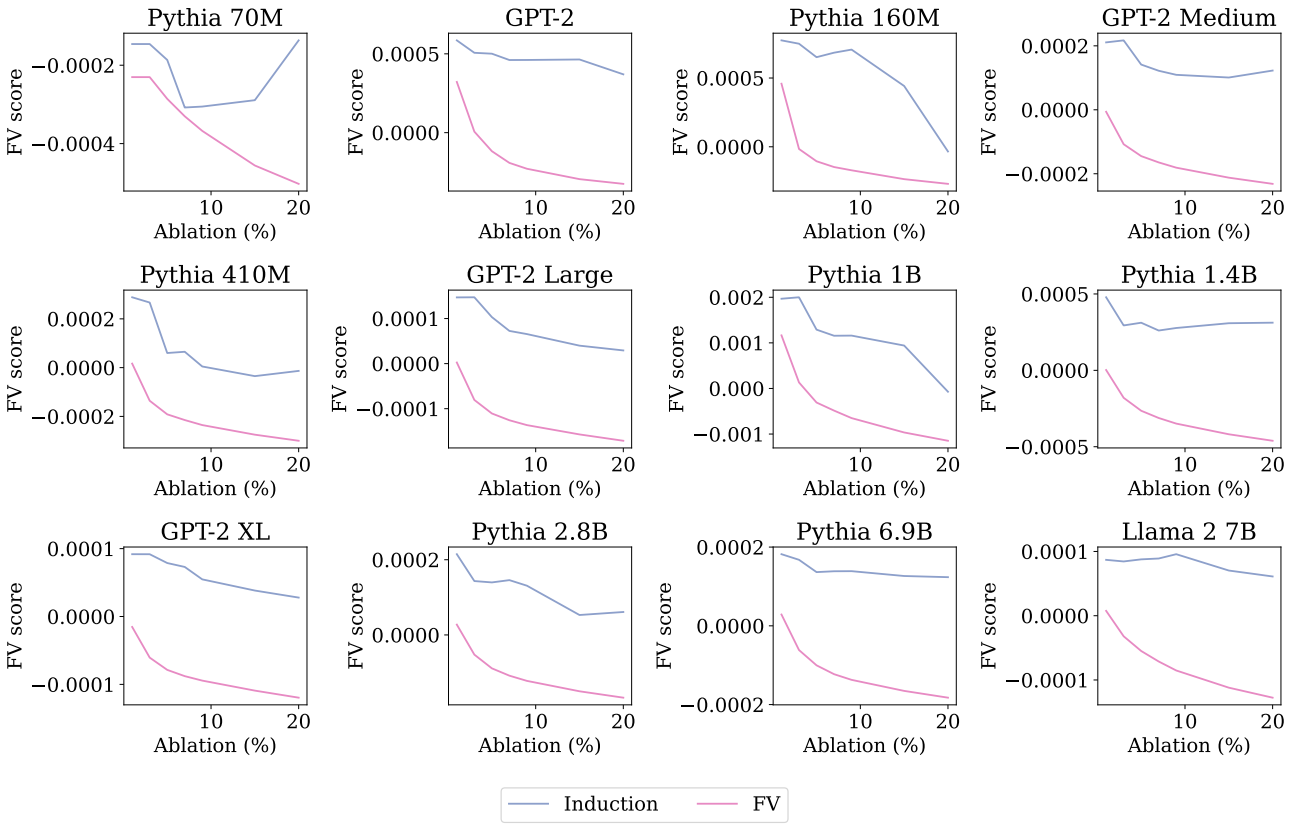


Figure 24. FV scores of the remaining heads after ablation.

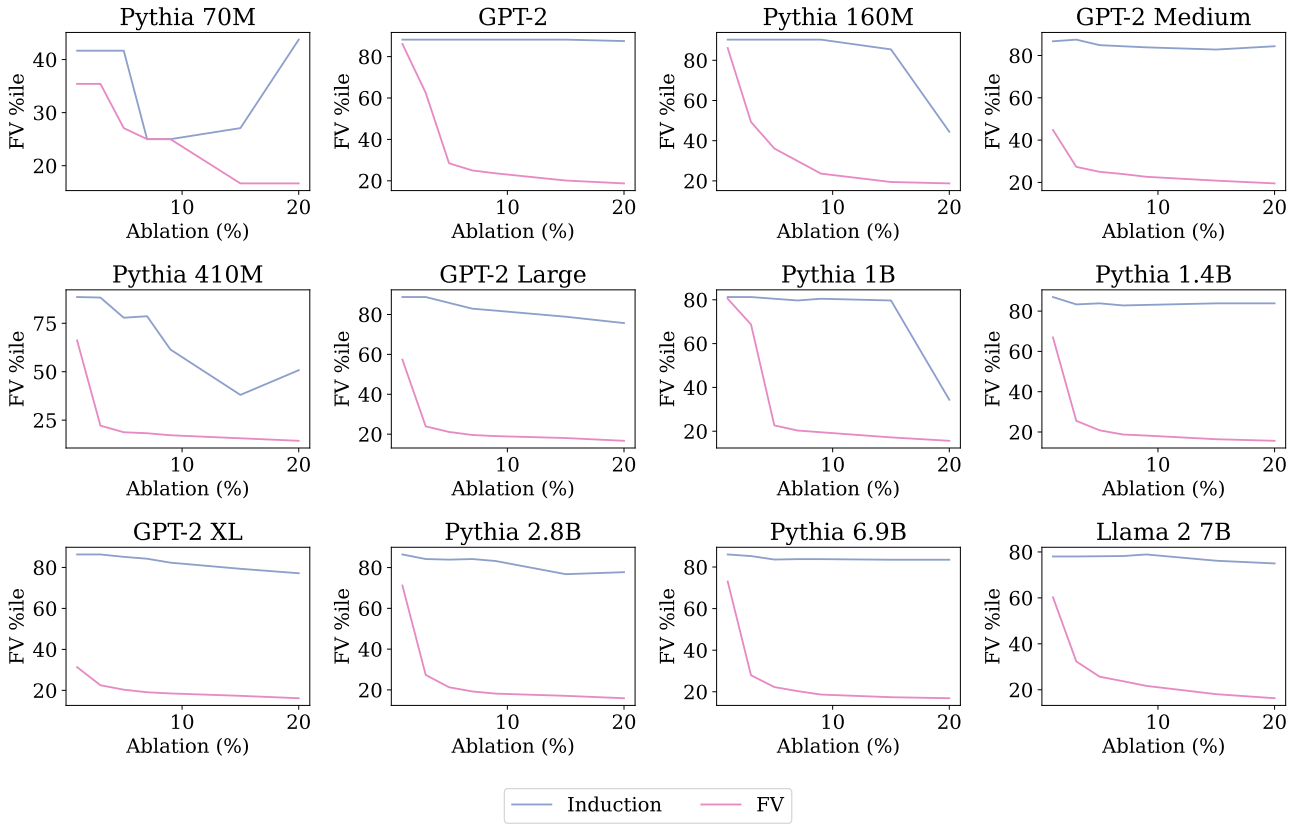


Figure 25. Percentile of FV scores of the remaining heads after ablation.

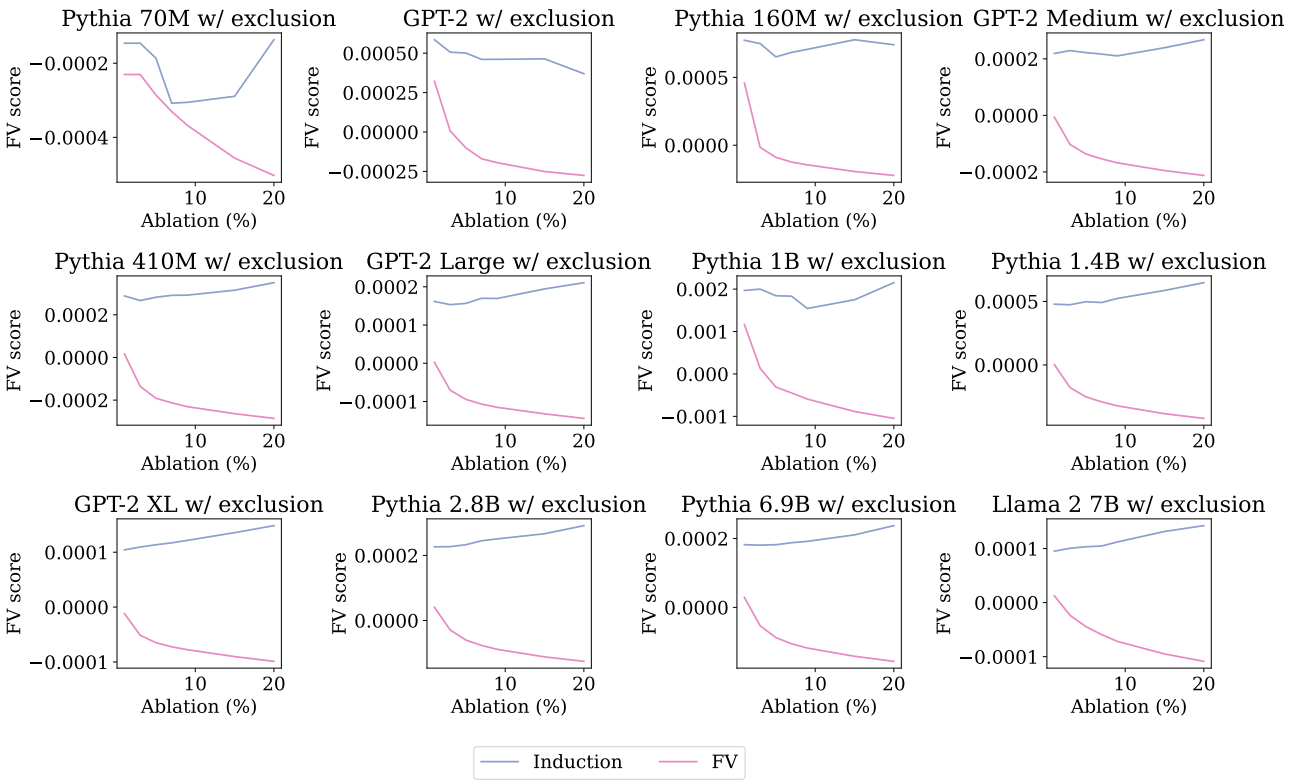


Figure 26. FV scores of the remaining heads after ablation with exclusion.

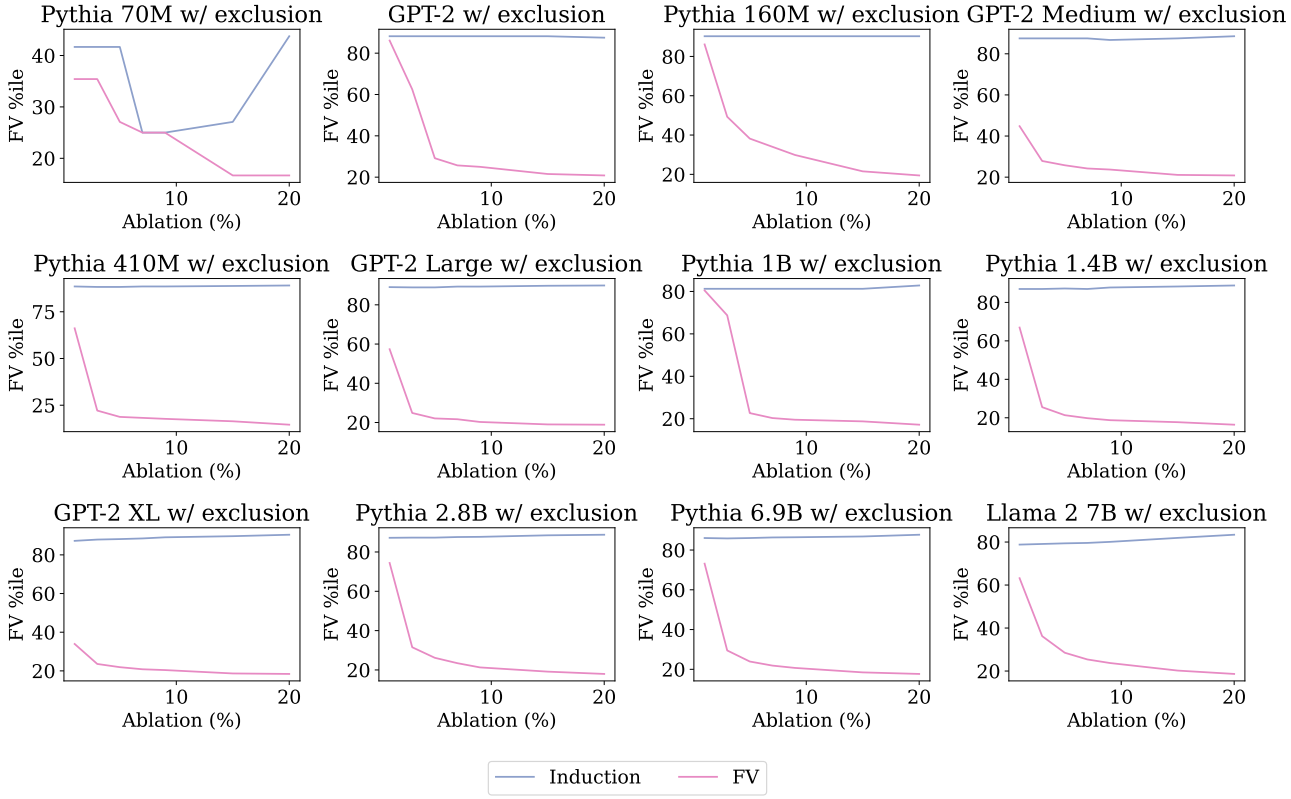


Figure 27. Percentile of FV scores of the remaining heads after ablation with exclusion.

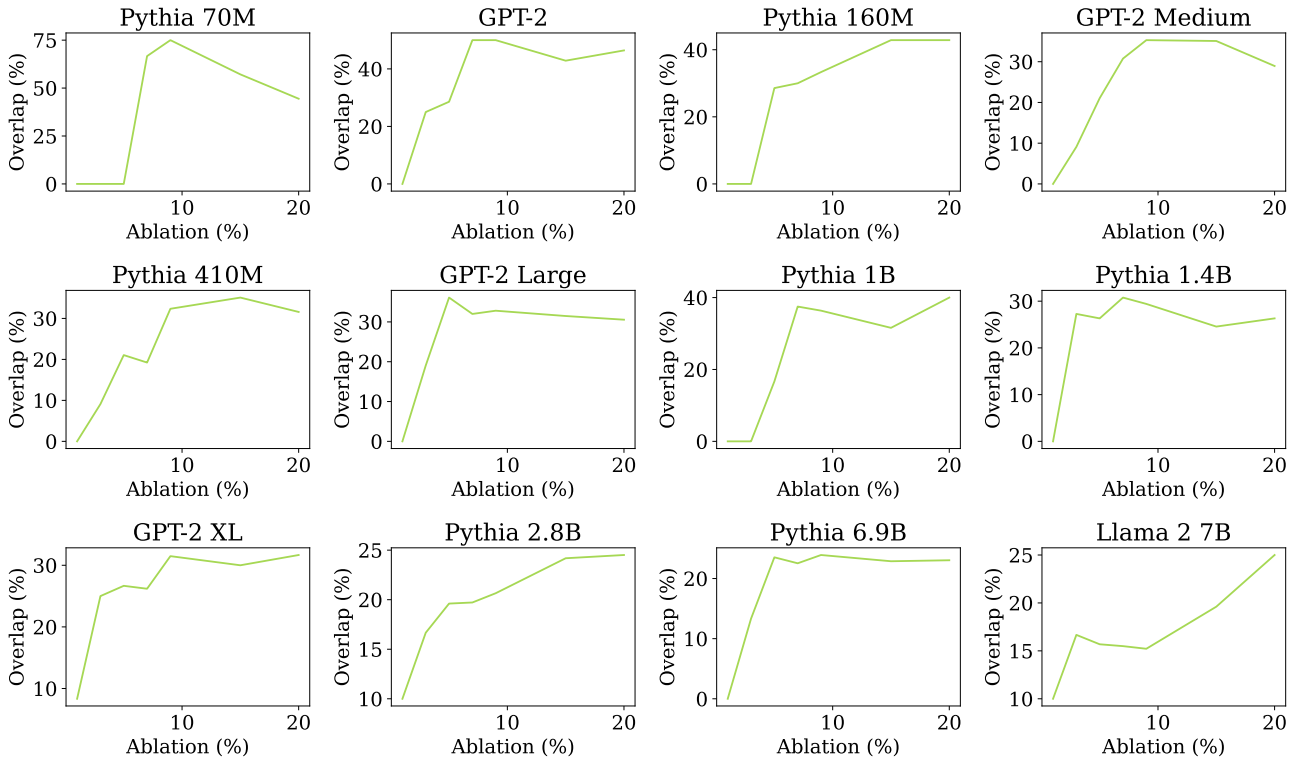


Figure 28. Overlap between set of induction heads and FV heads ablated.

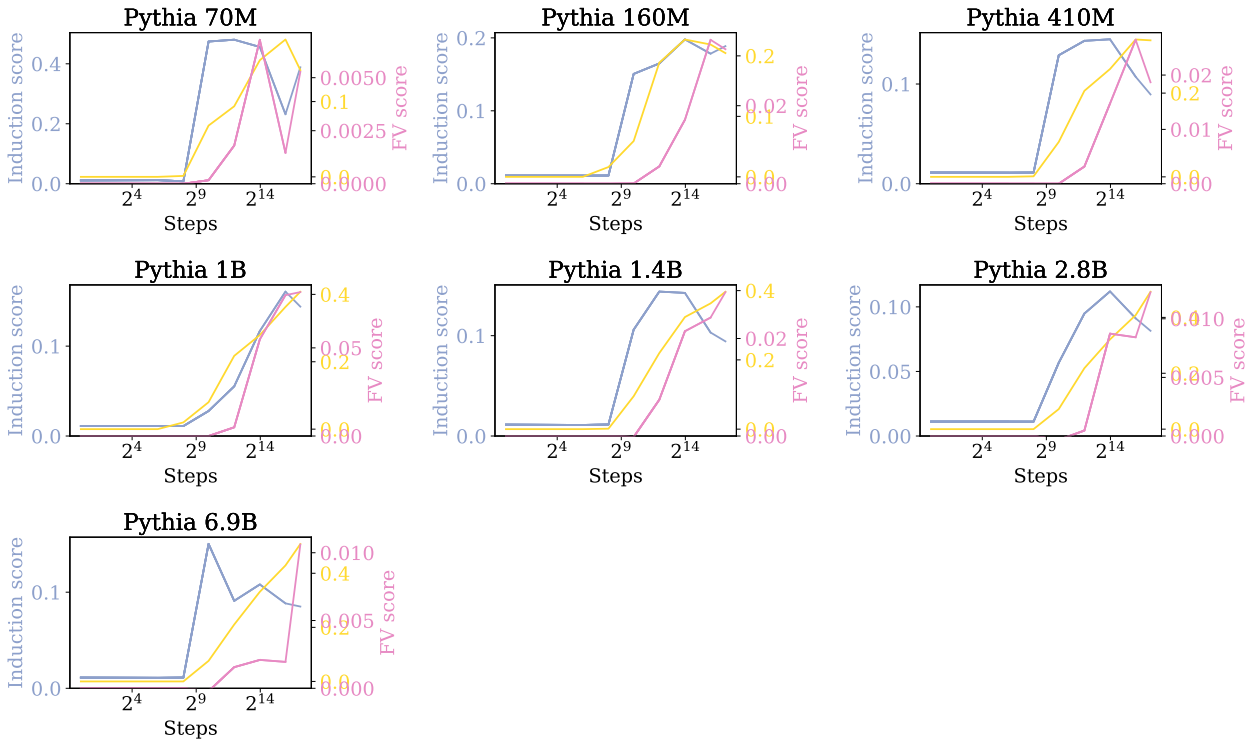


Figure 29. Evolution of induction score and FV score averaged over top 2% heads across training.

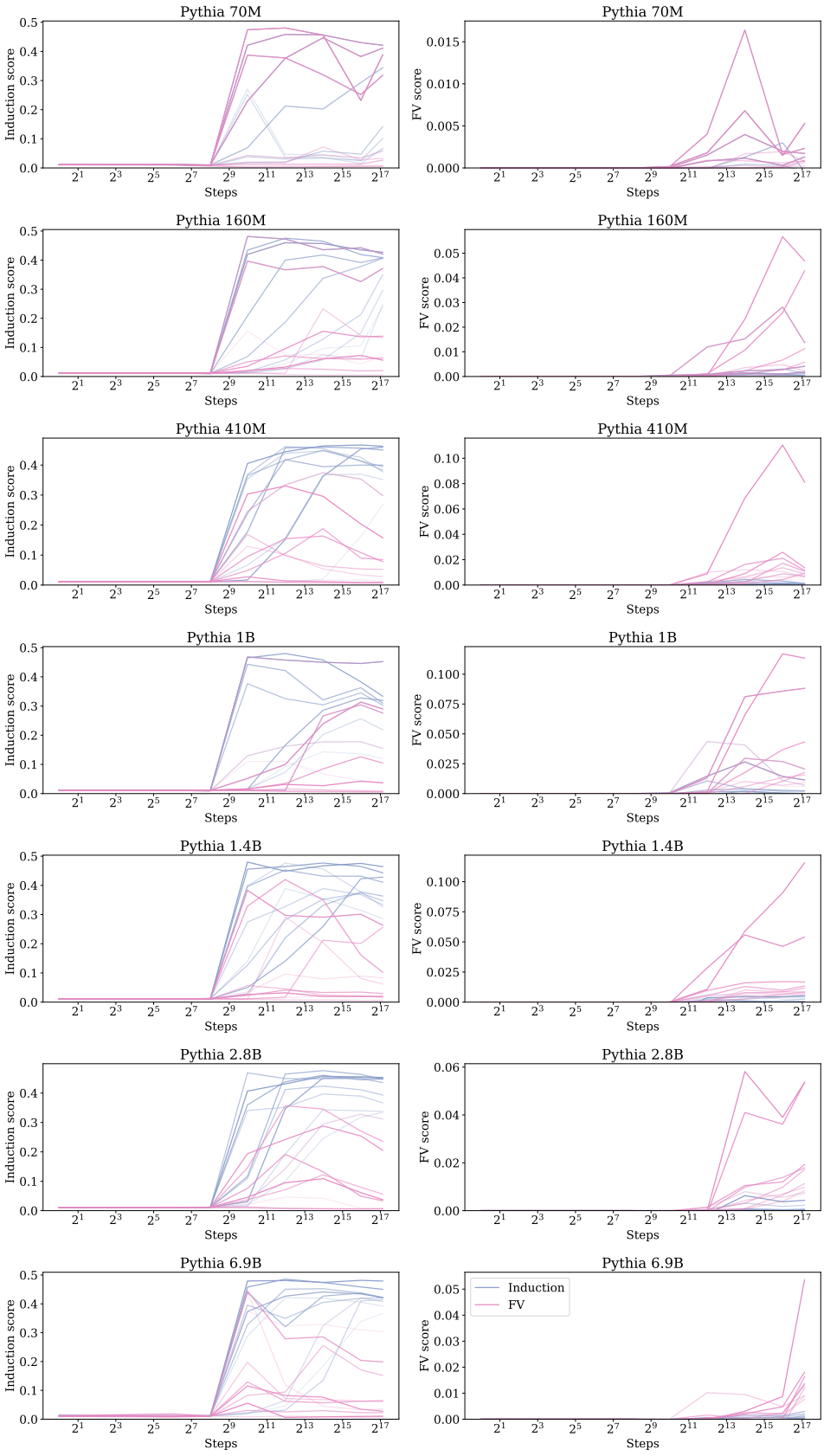


Figure 30. Evolution of induction scores (left) and FV scores (right) of individual induction and FV heads across training