Understanding In-Context Learning Beyond Transformers: An Investigation of State Space and Hybrid Architectures

Shenran Wang **Timothy Tin-Long Tse** **Jian Zhu**

The University of British Columbia
shenranw@cs.ubc.ca, ttse05@student.ubc.ca, jian.zhu@ubc.ca

Abstract

We perform in-depth evaluations of in-context learning (ICL) on state-of-the-art transformer, state-space, and hybrid large language models over two categories of knowledge-based ICL tasks. Using a combination of behavioral probing and intervention-based methods, we have discovered that, while LLMs of different architectures can behave similarly in task performance, their internals could remain different. We discover that function vectors (FVs) responsible for ICL are primarily located in the self-attention and Mamba layers, and speculate that Mamba2 uses a different mechanism from FVs to perform ICL. FVs are more important for ICL involving parametric knowledge retrieval, but not for contextual knowledge understanding. Our work contributes to a more nuanced understanding across architectures and task types. Methodologically, our approach also highlights the importance of combining both behavioural and mechanistic analyses to investigate LLM capabilities.

1 Introduction

In-context learning (ICL) (Brown et al., 2020a,b), the ability to learn from few-shot demonstrations at test time, is an emergent ability from pretrained Large Language Models (LLMs). In ICL, a few training samples are presented as demonstrations in the prompt, from which the model can learn to make predictions without any parameter updates. This emergent ability has raised interest in research on how LLMs acquire and generalize patterns solely from contexts at test time.

Mechanistic interpretability studies have successfully attributed ICL in transformers to certain types of self-attention heads, notably FVs (FVs) (Todd et al., 2024; Hendel et al., 2023; Yin and Steinhardt, 2024) and induction heads (Olsson et al., 2022; Edelman et al., 2024; Yin and Steinhardt, 2024), but such analyses are limited to transformers (Vaswani et al., 2017).

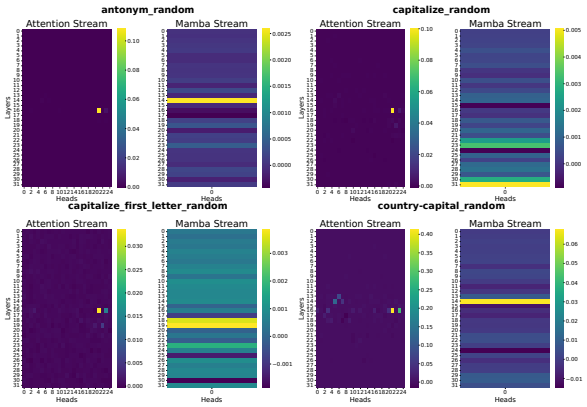

Figure 1: HYMBA-1.5B-BASE's AIE heatmap on a subset of parametric knowledge retrieval ICL. Top FV heads identified are much more concentrated in self-attention layers than in SSM layers. [X-axis: head number; Y-axis: layer number.]

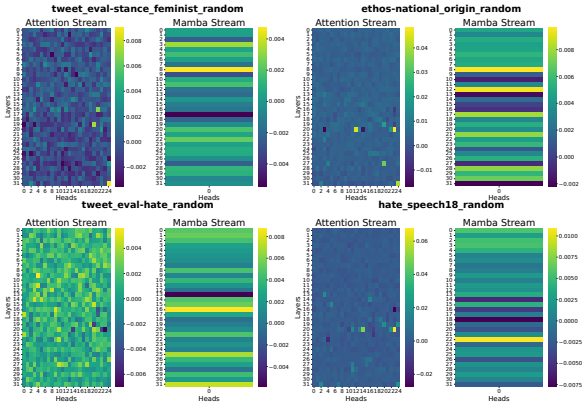

Figure 2: HYMBA-1.5B-BASE's AIE heatmap on a subset of contextual knowledge understanding ICL. The top FV heads are far less concentrated than in parametric knowledge retrieval ICL. [X-axis: head number; Y-axis: layer number.]

Some early explorations (Grazzi et al., 2024; Park et al., 2024a; Li et al., 2024; Li et al.; Bondaschi et al., 2025) show that state-space models (SSMs) like Mamba (Gu and Dao, 2023) and Mamba2 (Dao and Gu, 2024) can also perform

Findings of the Association for Computational Linguistics: ACL 2026, pages 794–820
July 2-7, 2026 ©2026 Association for Computational Linguistics

ICL, though their ICL capabilities are weaker than transformers. Transformer-SSM hybrid architectures are therefore proposed to improve the ICL in SSMs (Park et al., 2024a). However, little is known about the internal mechanisms of ICL in SSMs and hybrid models.

To address this gap, we analyze ICL in transformers, state space models, and hybrid models. Built on prior works (Min et al., 2022b; Todd et al., 2024; Yin and Steinhardt, 2024), we further distinguish two types of knowledge-based ICL tasks that require different reasoning abilities (Jin et al., 2025). A combination of behaviroal and mechanistic interpretability experiments yield several novel observations not previously reported in the literature.

- While all architectures demonstrate qualitatively similar ICL performance and robustness against noisy labels, mechanistic interpretability analyses reveal differences in internal mechanisms.
- FVs contribute more to the ICL capabilities in transformers, Mamba, and hybrid models, but less to Mamba2.
- For hybrid models, ICL capabilities are primarily contributed by the FVs located in self-attention layers, regardless of whether the self-attention layers and SSM layers are stacked in parallel or interleaved sequantially.
- FVs play a more important role in parametric knowledge retrieval ICL tasks. For ICL tasks that require contextual understanding, FVs have less impact on performance. These two types of ICL tasks do not always share the same set of FVs.

Our work has extended the prior analysis of FVs (Todd et al., 2024; Hendel et al., 2023; Yin and Steinhardt, 2024) to more diverse architectures and further contributes to a more nuanced understanding of FVs in different knowledge-based task types. Methodologically, our approach also highlights the importance of combining both behavioural and mechanistic analyses to investigate model behaviors. Our code is available at: `https://github.com/ShenranTomWang/ICL`.

2 Related Work

Mamba and Hybrid LLMs Mamba (Gu and Dao, 2023) and Mamba2 (Dao and Gu, 2024) are state-space models (SSMs) with selective state update mechanisms. Mamba-based and Mamba-transformer hybrid LLMs have also emerged as more efficient replacements for transformer-based LLMs. Zamba (Glorioso et al., 2024b), Zamba2 (Glorioso et al., 2024a), Jamba (Lenz et al., 2025), Samba (Ren et al., 2025), and Nemotron-H (NVIDIA et al., 2025) are representative hybrid models that stack self-attention and Mamba/Mamba2 sequentially. Hymba (Dong et al., 2024) is a hybrid model that integrates self-attention and Mamba in parallel. In this research, we examine the ICL capabilities of these state space and hybrid models, which have remained understudied.

In-Context Learning Since first observed by Brown et al. (2020b), ICL has been studied across different perspectives. Garg et al. (2022) and Zhang et al. (2023) provide theoretical insights into tasks that transformer models can learn in-context. Xie et al. (2022), Wies et al. (2023), Wang et al. (2023b) explores the theory behind ICL. Min et al. (2022b), (Reynolds and McDonell, 2021), and Yoo et al. (2022) studies how prompts can affect ICL performance. Other than transformer models, Park et al. (2024b) studies the performance of Mamba-based models in ICL. Gu et al. (2023) and Min et al. (2022a) propose to improve ICL performance of LLMs in training time. Todd et al. (2024) and Yin and Steinhardt (2025) examine the ICL mechanism through the lens of mechanistic interpretability. In our research, we will focus on hybrid models, and perform analysis on both the prompt level and the model internals.

Function Vectors and Function Vector Heads Recent work by Hendel et al. (2023) and Todd et al. (2024) has demonstrated that certain attention heads in LLMs are responsible for in-context learning, namely function vector (FV) heads. Yin and Steinhardt (2025) further studies the roles and importance of FV induction heads in in-context learning, identifying how effective each function head is in LLMs. However, FVs have only been identified in transformer architectures, and little is known about the internal mechanism of ICL in non-transformer architectures. To fill in this gap, we will apply similar mechanistic interpretability techniques to investigate the FVs in state-space models and hybrid architectures.

3 Models and Tasks

To understand ICL, we selected diverse but representative architectures and knowledge-based tasks.

3.1 Baseline models

For SSMs, we chose pretrained checkpoints of MAMBA-1.4B[1] (Gu and Dao, 2023), MAMBA2-1.3B[2] (Dao and Gu, 2024). For pretrained transformer LLMs, GEMMA-3-1B-PT (Team et al., 2025), LLAMA-3.2-1B (Grattafiori et al., 2024) and QWEN2.5-1.5B (Qwen et al., 2025) were selected as baselines. All models were limited to around 1B parameters for fair comparison.

3.2 Hybrid Models

We selected HYMBA-1.5B-BASE (Dong et al., 2024) and ZAMBA2-1.2B (Glorioso et al., 2024a). They were chosen becasue they had competitive performance with transformer LLMs and they represent two typical designs of hybrid models, that is, parallel and sequentially-stacked hybrid layers.

Zamba2 In ZAMBA2, a self-attention layer with LoRA (Hu et al., 2022) is stacked on top of every 6 Mamba2 layers. A linear layer is then applied to the outputs of the self-attention layer, followed by more Mamba2 layers. Our selected ZAMBA2-1.2B has 37 Mamba2 layers, each with 64 heads, and 6 self-attention layers, each with 32 attention heads.

Hymba Unlike in ZAMBA2, HYMBA incorporates Mamba and self-attention blocks in parallel. At each layer, inputs individually go through Mamba and sliding-window self-attention blocks and are taken the mean before going through the out projection. Our selected HYMBA-1.5B-BASE has 32 layers. Each layer has 25 self-attention heads and 1 Mamba head.

3.3 Tasks

Adopting the definitions in Jin et al. (2025), we classified our ICL tasks into **Contextual Knowledge Understanding** and **Parametric Knowledge Retrieval**, a distinction not explicitly made in prior works (Min et al., 2022b; Todd et al., 2024).

Parametric Knowledge Retrieval Parametric knowledge retrieval refers to questions that can be answered correctly by simply using the query and the knowledge within the model to perform a retrieval match (e.g., country-capital, country-currency). They were the focus in prior FV research (Yin and Steinhardt, 2025; Todd et al., 2024). We adapted 17 datasets from Todd et al. (2024).

Contextual Knowledge Understanding These tasks require understanding the content within a paragraph and using the information it provides to answer questions, such as hate speech classification and sentiment analysis. The associations tend to be noiser than the parametric knowledge retrieval tasks. We adapted 16 original datasets from Min et al. (2022b).

To guarantee a decent amount of testing samples, we keep 30% of all samples for testing in the parametric knowledge retrieval datasets. For the contextual knowledge understanding datasets, we use their default train-test split available on Hugging Face. Detailed statistics of our datasets are listed in Table 1 in Appendix A.

4 Behavioral Experiments

We begin with behavioral experiments (Min et al., 2022b) to test whether different architectures behave similarly in ICL.

4.1 Label Randomization Experiments

Method We followed the same setup in Min et al. (2022b): for each dataset, we sampled k question-answer pairs as demonstrations with k = 4, 8, 12, 16, and 32. To evaluate the performance, we found the logit indexes of the first token of each option, and then select the index with the maximum logits as the answer. We ran each experiment 5 times with 5 random seeds to measure the mean performance. For the parametric knowledge retrieval datasets without options, we sorted all possible answers to form the options (Todd et al., 2024). We conducted experiments under settings below:

1. **No demo**: we give the model $k = 0$ demonstration, i.e. we present only the query.
2. $\alpha\%$ **correct**: we augment our k demonstrations such that $\alpha\%$ of the demonstrations are correct. We select α to be 0, 25, 50, 75.
3. **Gold**: we give correct demonstrations to the model. This is equivalent to setting $\alpha = 100$.
4. **Random**: the demonstrations' questions and answers are shuffled.

Results As presented in Figure 3, results in parametric knowledge retrieval tasks demonstrate in-

[1] https://huggingface.co/state-spaces/mamba-1.4b-hf

[2] https://huggingface.co/AntonV/mamba2-1.3b-hf

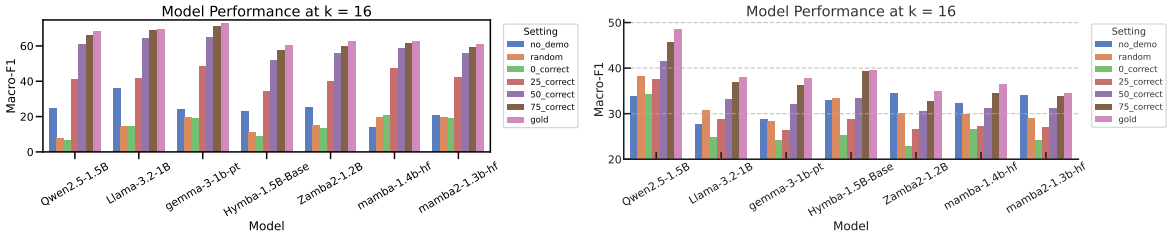

Figure 3: **Left:** Performance of different models on parametric knowledge retrieval datasets, for k = 16, on initial setting. **Right:** Performance of different models on contextual knowledge understanding datasets, for k = 16, on initial setting.

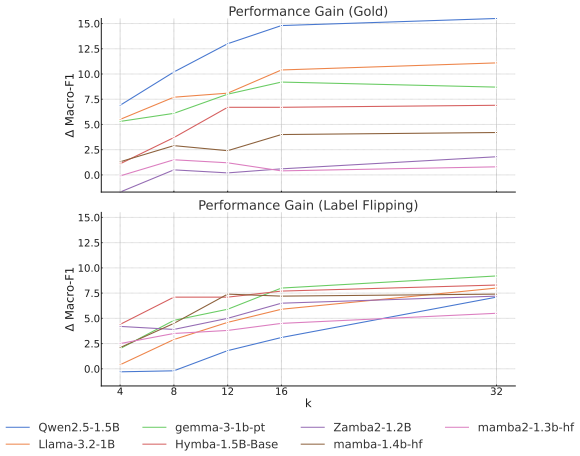

Figure 4: Performance gain of label flipping and gold conditions on contextual knowledge understanding datasets. Performance gain is the difference between the performance of gold or label flipping and the corresponding no demo conditions. Performance gain for other conditions are plotted in Figure 23 in Appendix B.

context learning being properly carried out. Surprisingly, we find that in contextual knowledge understanding datasets, while all models are affected by the percentage of correct demonstrations given, transformer-based models, HYMBA-1.5B-BASE and MAMBA-1.4B-HF are able to outperform no demo when gold demonstrations are provided, whereas MAMBA2-1.3B-HF and ZAMBA2-1.2B only demonstrate a marginal increment in performance. **Notice that the phenomenon that model performance varies significantly with different percentages of correct examples contradicts the findings of Min et al. (2022b).** From the top figure of Figure 4, we can observe that with more number of ICL examples, transformer models, HYMBA-1.5B-BASE, and MAMBA-1.4B-HF are able to perform better. However, this trend barely holds true for the models whose gold label performance merely outperforms no demo, namely ZAMBA2-1.2B and MAMBA2-1.3B-HF. **Under these regular settings, ZAMBA2-1.2B and MAMBA2-1.3B-HF have relatively weak ICL performance for tasks involving contextual knowledge understanding.**

Impact of k **In general, the impact of k shows consistent findings with Min et al. (2022b), that is, models only benefit from more demonstrations when the majority of labels are reliable.** In the 0% and 25% correct settings, more examples actually hurt the performance. This is expected as more examples mislead the model and therefore cause lower performance. The 50% correct and the random case yield similar curves as the 50% correct case, as statistically they are almost equivalent. At 75% correct, all models except ZAMBA2-1.2B and MAMBA2-1.3B-HF are demonstrating increments in performance with increased k, suggesting that they can guess the task from examples already. Performance plateaus mostly after $k = 8$ for all settings. Additional figures supporting these claims are available in Figures 14 to 22 in Appendix B.

4.2 Label-flipping Experiments

In the label randomization experiments above for contextual understanding and retrieval tasks, we have observed that ZAMBA2-1.2B and MAMBA2-1.3B-HF only demonstrate marginal gains on gold-label demonstrations. However, they experience a significant performance degradation when the given demonstrations are entirely wrong. We introduce the label-flipping experiments to look into this phenomenon.

Method To explore whether MAMBA2-1.3B-HF and ZAMBA2-1.2B can learn under a counterfactual setting and reduce the impact of memorization, we also introduced the label flipping setting: for each datapoint in the demonstrations and test set, we map the correct answer consistently to one of the incorrect options, and we expect the mod-

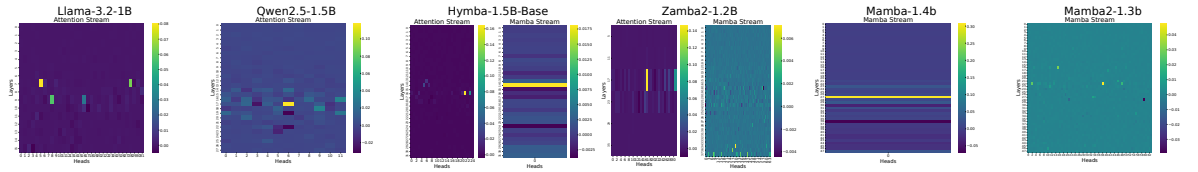

Figure 5: AIE heatmap for model heads on parametric knowledge retrieval datasets. [X-axis: head number; Y-axis: layer number.]

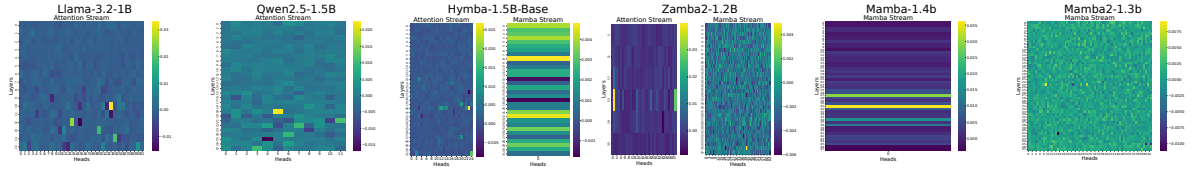

Figure 6: AIE heatmap for model heads on contextual knowledge understanding datasets. [X-axis: head number; Y-axis: layer number.]

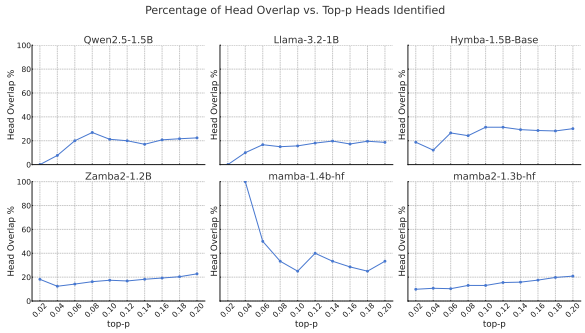

Figure 7: Percentage of intersection between top FV heads identified from contextual knowledge understanding datasets and parametric knowledge retrieval datasets. Entries are the percentage of heads in top-p that overlap. In general, different top heads are identified for different categories of tasks, as the overlaps are generally small. For MAMBA-1.4B-HF, the top-2% heads yield an empty set.

els to categorize queries to the flipped (incorrect) options. For example, for options {hate, neutral, non-hate}, we replaced "hate" with "non-hate" and used the flipped labels for evaluation. All mappings were strictly one-to-one to ensure that there are still systematic associations between questions and answers, but such associations should be absent in the training data.

Notice that this is different from label randomization: label randomization randomizes the examplars' labels but expects the model to output the correct relationship, whereas label flipping expects the model to infer the new relationship from label-flipped examplars. The goal is to examine how well LLMs can pick up counterfactual relationships in context without the influence of domain knowledge. We do this for only the contextual knowledge understanding datasets.

Results The bottom panel of Figure 4 shows that all models are able to learn in-context the label-flipped tasks for contextual knowledge understanding tasks. Models whose performance gain used to be near zero or negative (ZAMBA2-1.2B and MAMBA2-1.3B-HF) are now above zero. We also see that these two models and MAMBA-1.4B-HF achieve more gain over the increasing number of ICL examples. Nevertheless, self-attention models are still able to achieve the biggest performance increment over the increasing number of ICL examples. **Evidence shows that all models, including ZAMBA2-1.2B and MAMBA2-1.3B-HF, are able to learn unseen associations in context, and that they all exhibit qualitatively the same learning curve.**

5 Mechanistic Interpretability Analysis

The behavioral experiments have identified some differences in ICL across models. However, the internal mechanism behind this phenomenon remains unclear. In this section, we will examine the FV heads in QWEN2.5-1.5B, LLAMA-3.2-1B, HYMBA-1.5B-BASE, ZAMBA2-1.2B, MAMBA-1.4B-HF, and MAMBA2-1.3B-HF.

5.1 Identifying FV Heads

We replicated the method of Todd et al. (2024) on our models to identify FV heads. Originally, this was done only for transformers. For SSMs and hybrid models, we treated the SSM heads as analogous to attention heads.

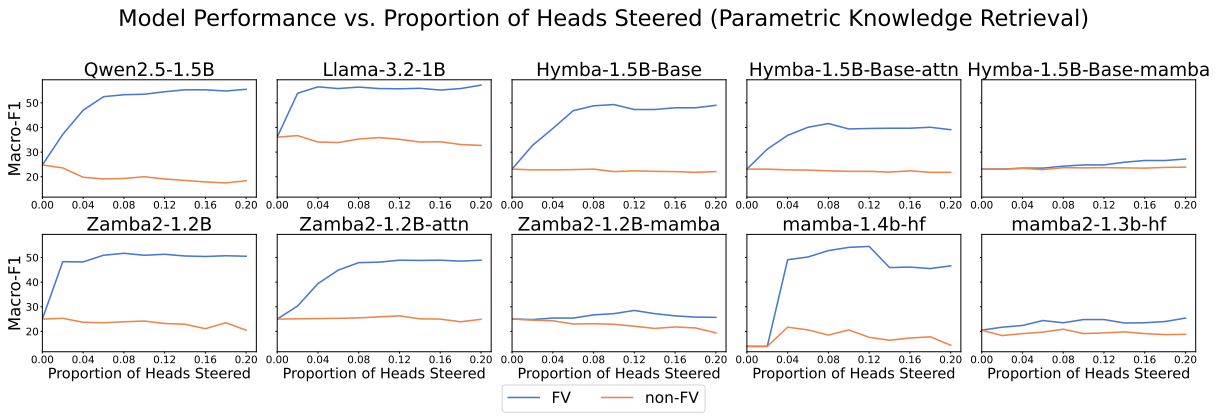

Figure 8: Steering results for models in parametric knowledge retrieval datasets. For hybrid models, we observe a significant performance increment for the self-attention stream while the mamba stream only demonstrates marginal improvement. Notably, MAMBA2-1.3B-HF does not seem to be significantly influenced by steering.

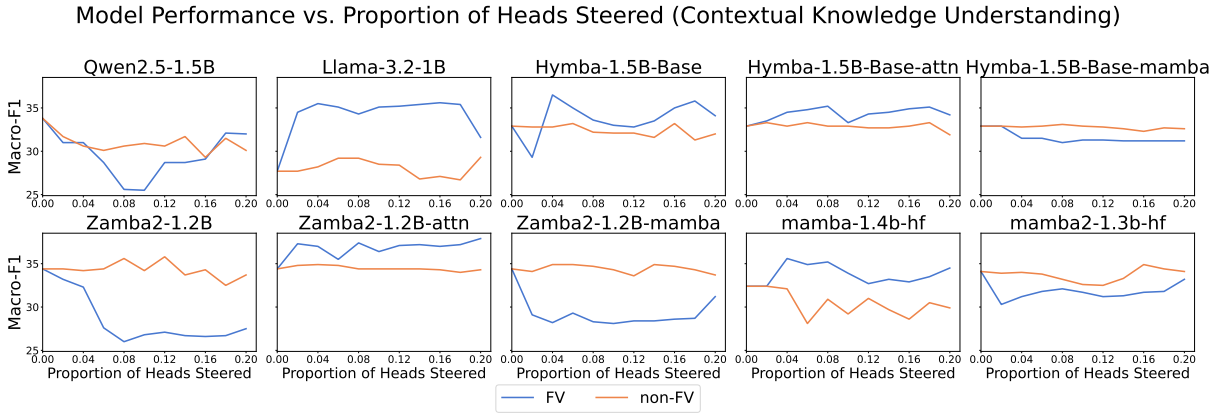

Figure 9: Steering results for models in contextual knowledge understanding datasets. Results demonstrate more fluctuations. It is worth noticing that steering the mamba stream in both hybrid models consistently decreases performance. Especially for ZAMBA2-1.2B, the decrease is so significant that it causes the overall performance to drop below non-FV steering, even though steering its self-attention stream still gives positive feedback. Surprisingly, steering QWEN2.5-1.5B hurts performance.

We first extracted the FVs for all heads at all layers from all streams at the very last token of the input prompt. Since Mamba and the Mamba stream of HYMBA-1.5B-BASE was not multi-head, we treated it as one head. Mathematically put, let $p_i^t \in P_t$ be a prompt in dataset P_t representing task t, for each attention head a_{lj} at layer l, we wish to take the task-dependent output value of this attention head, calculated as:

$$\bar{a_{lj}}^t = \frac{1}{|P_t|} \sum_{p_i^t \in P_t} a_{lj}(p_i^t)$$

We used the $k = 10$ randomly chosen demonstrations, and we extracted the FVs on the validation set. After that, for each task, we calculated the average logit-level percentage shift towards the correct logit as we replaced each head with its corresponding $\bar{a_{lj}}^t$, namely average indirect effect (AIE), on the random demonstrations setting. Formally:

$$\text{AIE}(a_{lj}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{\tilde{P}_t} \sum_{\tilde{p}_i^t \in \tilde{P}_t} \text{CIE}(a_{lj} | \tilde{p}_i^t)$$

where $\tilde{p}_i^t \in \tilde{P}_t$ is the corrupted prompt with random demonstrations from the random demonstration dataset $\tilde{P}_t$, $t \in \mathcal{T}$ is a dataset in task $\mathcal{T}$, and CIE is the causal indirect effect defined as:

$$\text{CIE}(a_{lj} | \tilde{p}_i^t) = f(\tilde{p}_i^t | a_{lj} := \bar{a_{lj}}^t)[y_{iq}] - f(\tilde{p}_i^t)[y_{iq}]$$

where f is the model's forward pass, and y_{iq} is the token of the correct answer.

We reserved 100 random samples from the training set to compute AIE, and the ICL examples are sampled at random from the remaining training samples. Since AIE is compute-heavy, we used $k = 10$ ICL examples over 25 random samples

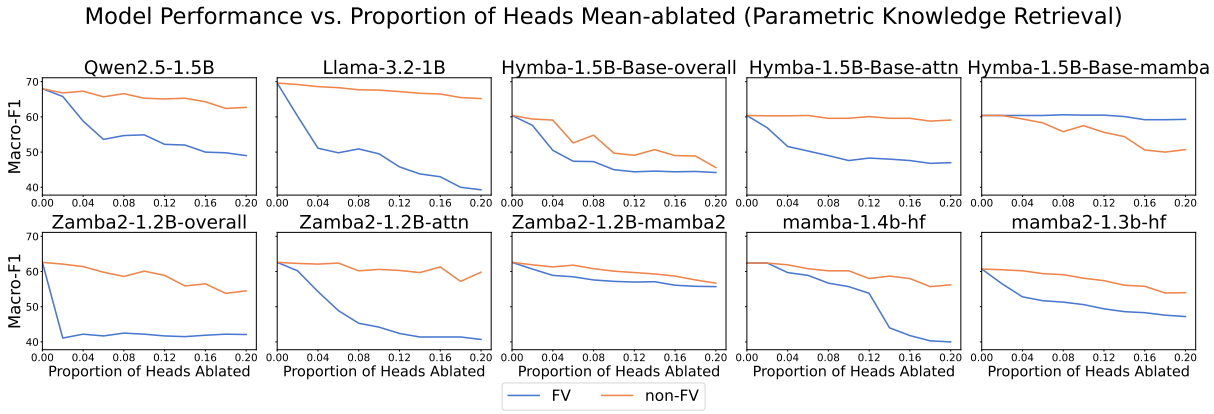

Figure 10: Mean ablation results for parametric knowledge retrieval ICL.

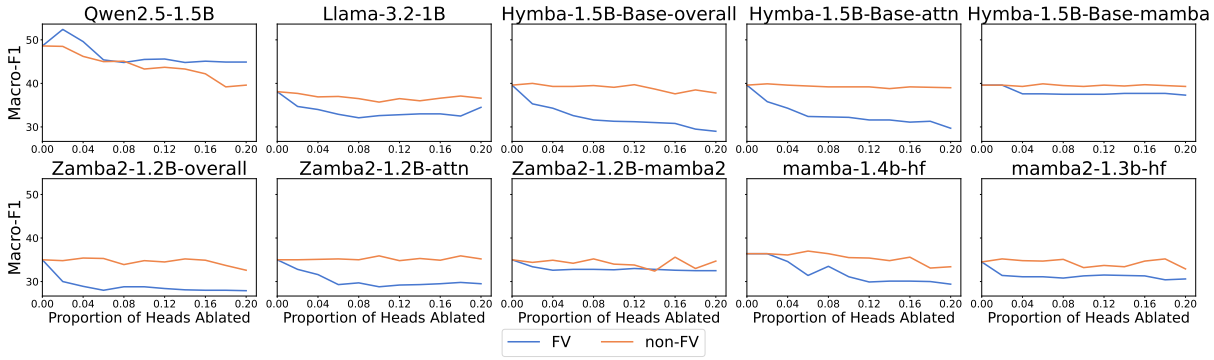

Figure 11: Mean ablation results for contextual knowledge understanding ICL.

from the development split over only one random seed. We then evaluated the steering performance on the test set.

5.2 Locations of FV heads

Top FV heads are highly consistent and concentrated for parametric knowledge retrieval ICL but not contextual understanding ICL. Figures 1 and 2 show that, in HYMBA-1.5B-BASE, certain heads are always activated in the parametric knowledge retrieval ICL, but the activations are less concentrated in contextual knowledge ICL. Such patterns also exist in all other models with full heatmaps in Appendix C. As shown in Figures 5 and 6, under different tasks, all models except MAMBA-1.4B-HF activate different heads for the two different categories of tasks. We quantified the percentage of overlapping top-p FV heads for the two categories in Figure 7. **Inspection suggests that the top FVs for the two kinds of ICL tasks do not necessarily overlap in most models except for MAMBA-1.4B-HF.**

5.3 Function vector internevtion

In the last section, we had identified the top FVs. To causally validate their roles, we conducted two intervention experiments.

Steering Function Vector Heads We took the no demo variants of each dataset, then added $\bar{a_{lj}}^t$ to the output of that head at the last token position. Formally put, let a_{lj}^t be the value of the last token of output from attention head j in layer l, we add $\bar{a_{lj}}^t$ calculated before to a_{lj}^t:

$$a_{lj}^{t\,\prime} = a_{lj}^t + \bar{a_{lj}}^t$$

We then evaluated the performance of models. We did this for the top 2 to 20 percent of heads, selected by AIE. For hybrid models, we first selected the top heads on both streams, then did the same on each stream individually, to see the effectiveness of both streams. As a comparison, we first steered 2 to 20 percent of random heads outside of the top 20% FV heads (non-FV heads) identified by AIE. We also steered all heads at each layer only.

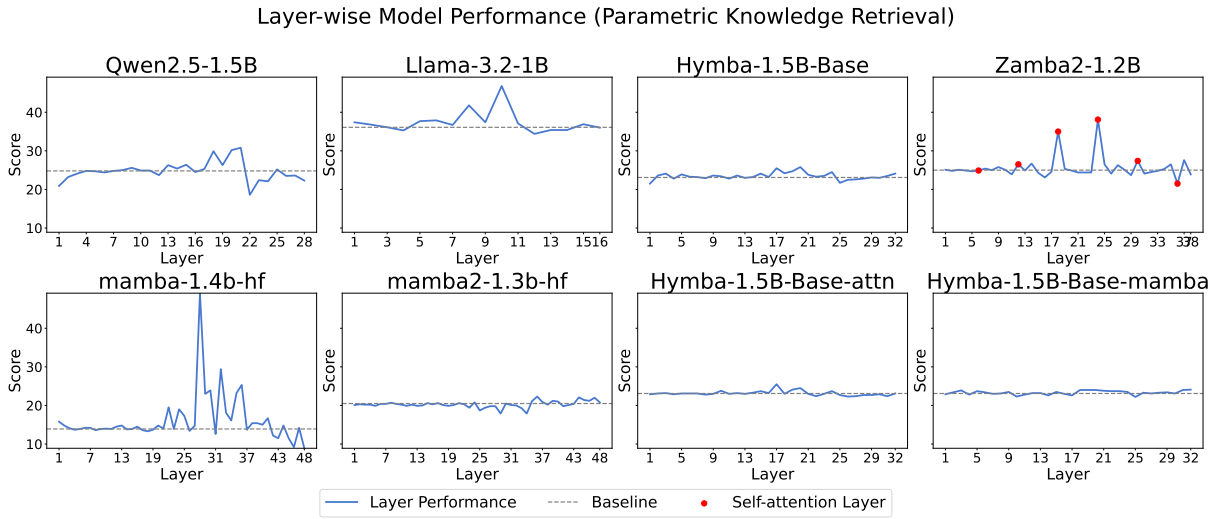

Figure 12: Layer-wise steering results for models in parametric knowledge retrieval datasets. The self-attention stream remains dominant in both hybrid models. We also see MAMBA-1.4B-HF's middle layers playing a significant role.

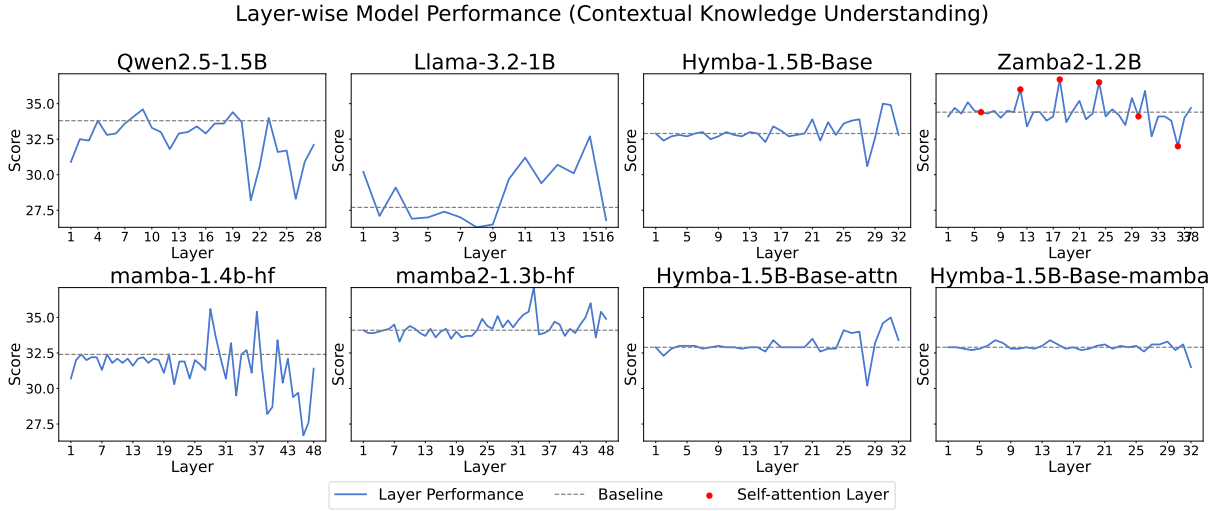

Figure 13: Layer-wise steering results for models in contextual knowledge understanding datasets. The self-attention stream remains dominant in both hybrid models.

Ablating function vector heads Adopting the experiment in Yin and Steinhardt (2025), we replaced the last token of the outputs of selected FV heads with zeros or the mean of this head's activation at the last token position over all datasets within its category. We did this for the top 2 to 20 percent of heads, selected by AIE. We did the same as the steering experiment above for hybrid models. We also ablated random 2 to 20 percent of heads outside of the top 20% identified FV heads (non-FV heads) as a comparison. We would focus on mean ablations as in Yin and Steinhardt (2024), because zero ablations may cause the out-of-distribution problem (Hase et al., 2021; Wang et al., 2023a; Zhang and Nanda, 2024).

5.4 Findings

Our two intervention-based analyses yield highly consistent results, and the main findings are summarized below.

FV heads drive ICL in self-attention models and Mamba, but not so much in Mamba2. We observe in Figures 8 and 9 that the steering performance increment of MAMBA2-1.3B-HF is insignificant compared to other models. However, unlike in the steering experiments, Figures 10 and 11 show that MAMBA2-1.3B-HF's performance change in ablations is comparable with other models in both categories. Along with our steering results, we can claim that MAMBA2-1.3B-HF does not depend on

FV heads to retrieve task knowledge, but rather uses its heads for other purposes. We hypothesize that since MAMBA2-1.4B-HF is multi-head, each head has a smaller number of hidden dimensions and therefore is less capable of capturing task-specific information at the head level.

FV heads can account for the ICL performance in parametric knowledge retrieval tasks, but not in contextual knowledge understanding tasks. In Figures 8 and 9, we can observe that steering FV heads are mostly effective in the parametric knowledge retrieval dataset, whereas there are significantly more fluctuations in the contextual knowledge understanding datasets. Performance increment for steering contextual knowledge understanding datasets is significantly less than steering parametric data retrieval datasets. Notably, we see a consistent decrease in performance when steering the Mamba streams of both hybrid models in the contextual knowledge understanding datasets. It is likely that the Mamba stream serves other functionalities than the FV heads in the attention stream. With these observations, we can conclude that FVs are not key in contextual knowledge understanding datasets.

For hybrid models, their ICL capabilities are primarily controlled by FV heads at self-attention layers. To further study FV heads in these models, we will focus on parametric knowledge retrieval datasets. Looking at the figures of HYMBA-1.5B-BASE and ZAMBA2-1.2B, FV heads are primarily located in the attention stream of hybrid models. Nevertheless, the Mamba stream of HYMBA-1.5B-BASE is also important, as simply steering the attention stream cannot match the performance of steering both streams. This can be further justified by the results in Appendix D. We also find the non-FV heads identified in HYMBA-1.5B-BASE are very important, as ablating them causes more performance drop than the FV heads in the Mamba stream (Figure 10). This is likely caused by HYMBA-1.5B-BASE taking the mean of attention and Mamba stream output at each layer, so at some layers where FV heads are located but the corresponding Mamba head is not identified, this may cause a problem.

Steering FVs in the middle or later layers tends to improve ICL performance. Layerwise steering results in Figures 12 and 13 also reinforce our findings. We can easily identify that the self-attention layers in ZAMBA2-1.2B, especially the middle few, play a significant role in both sets of datasets. We also identify that the overall performance of HYMBA-1.5B-BASE is mostly governed by its attention stream. It is worth noticing that transformer models in parametric knowledge ICL react quite significantly to layer-wise steering. However, the layers that yield the highest performance increments do not necessarily contain the top identified FV heads, suggesting that they are more sensitive to layer-wise steering.

6 Conclusions

In this research, we performed an in-depth analysis of ICL on several mainstream architectures. Through a combination of behavioral and mechanistic analysis, we report several new findings concerning how FVs contribute to ICL in different architectures and task settings. Our work extends and refines the understanding of FVs from transformers to SSMs and hybrid models.

7 Limitations

We acknowledge that our study is still limited in several ways. ICL is a complex capability in LLMs that might involve multiple mechanisms in synchrony. In this study, we only focus on one potential mechanism, namely, FV heads. Yet there might be other mechanisms, including induction heads. Future studies should investigate more mechanisms to uncover the intricacies of ICL.

Due to limited computing budget, our analyses are limited to pretrained LLMs, like all prior works (Todd et al., 2024; Yin and Steinhardt, 2024; Min et al., 2022b). While we have made our best attempts to control for the experiment settings and model parameters, we have no control over the pretraining materials. We speculate that pretraining materials and procedures could also impact ICL capabilities and the locations of FV heads, since both QWEN2.5-1.5B and LLAMA3.2-1B still exhibit slightly different internals despite using similar self-attention layers. More well-designed experiments that strictly control for training data, hyperparameters, and training procedures across architectures can better clarify these issues. We will leave these to future studies.

8 Ethics statement

Our research primarily makes empirical contributions toward the internal mechanisms of LLMs. While most findings do not have direct practical applications, a better understanding of the internal

mechanisms may be exploited by malicious users to actively jailbreak or take control over deployed LLMs, potentially leading to undesirable societal risks.

9 Acknowledgments

We thank three anonymous reviewers and the area chairs for their thoughtful comments on the original manuscript. This research was enabled in part through the computational resources provided by Advanced Research Computing at the University of British Columbia and the Digital Research Alliance of Canada.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The research activities were also supported by the NSERC Discovery Grant and the CFI-JELF Grant awarded to JZ.

References

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Kannan Ramchandran, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Vardhan Makkuva. 2025. From markov to laplace: How mamba in-context learns markov chains. *arXiv preprint arXiv:2502.10178*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. To appear in proceedings of Sinn und Bedeutung 23. Data can be found at https://github.com/mcdm/CommitmentBank/.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the International Workshop on Paraphrasing*.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen,

Yoshi Suhara, Yingyan Lin, Jan Kautz, and Pavlo Molchanov. 2024. Hymba: A hybrid-head architecture for small language models.

Ezra Edelman, Nikolaos Tsilivis, Benjamin Edelman, Eran Malach, and Surbhi Goel. 2024. The evolution of statistical induction heads: In-context learning markov chains. *Advances in neural information processing systems*, 37:64273–64311.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, Anna Golubeva, Vasudev Shyam, James Whittington, Jonathan Pilault, and Beren Millidge. 2024a. The zamba2 suite: Technical report. *Preprint*, arXiv:2411.15242.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. 2024b. Zamba: A compact 7b ssm hybrid model. *Preprint*, arXiv:2405.16712.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Riccardo Grazzi, Julien Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. 2024. Is mamba capable of in-context learning? *arXiv preprint arXiv:2402.03170*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4849–4870, Toronto, Canada. Association for Computational Linguistics.

Peter Hase, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 3650–3666. Curran Associates, Inc.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025. Massive values in self-attention modules are the key to contextual knowledge understanding. In *Forty-second International Conference on Machine Learning*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. 2025. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Hongkang Li, Songtao Lu, Xiaodong Cui, Pin-Yu Chen, and Meng Wang. Understanding mamba in in-context learning with outliers: A theoretical generalization analysis. In *High-dimensional Learning Dynamics 2025*.

Yingcong Li, Xupeng Wei, Haonan Zhao, and Taigao Ma. 2024. Can mamba in-context learn task mixtures? In *ICML 2024 Workshop on In-Context Learning*.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs. *Preprint*, arXiv:2008.13546.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *Preprint*, arXiv:2006.08328.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. 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*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

NVIDIA, :, Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, and et al. 2025. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *Preprint*, arXiv:2504.03624.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024a. Can mamba learn how to learn? a comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 39793–39812.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024b. Can mamba learn how to learn? A comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39793–39812. PMLR.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Liliang Ren, Yang Liu, Yadong Lu, yelong shen, Chen Liang, and Weizhu Chen. 2025. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *The Thirteenth International Conference on Learning Representations*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. *Preprint*, arXiv:2011.02686.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, and et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*. ArXiv:2310.15213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023b. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Kayo Yin and Jacob Steinhardt. 2024. Which attention heads matter for in-context learning? *arXiv preprint arXiv:2502.14010*.

Kayo Yin and Jacob Steinhardt. 2025. Which attention heads matter for in-context learning? *Preprint*, arXiv:2502.14010.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2023. Trained transformers learn linear models in-context. *Preprint*, arXiv:2306.09927.

A Datasets

Dataset	# Train	# Eval
Datasets category: contextual knowledge understanding		
financial_phrasebank (Malo et al., 2014)	1811	453
poem_sentiment (Sheng and Uthus, 2020)	892	105
medical_questions_pairs (McCreery et al., 2020)	2438	610
glue-mrpc (Dolan and Brockett, 2005)	3668	408
glue-wnli (Levesque et al., 2011)	635	71
climate_fever (Diggelmann et al., 2020)	1228	307
glue-rte (Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2007)	2490	277
superglue-cb (De Marneffe et al., 2019)	250	56
sick (Marelli et al., 2014)	4439	495
hate_speech18 (de Gibert et al., 2018)	8562	2141
ethos-national_origin (Mollas et al., 2020)	346	87
ethos-race (Mollas et al., 2020)	346	87
ethos-religion (Mollas et al., 2020)	346	87
tweet_eval-hate (Basile et al., 2019)	8993	999
tweet_eval-stance_athesim (Mohammad et al., 2016)	461	52
tweet_eval-stance_feminist (Mohammad et al., 2016)	597	67
Datasets category: parametric knowledge retrieval		
antonym (Nguyen et al., 2017)	1678	720
capitalize_first_letter (Todd et al., 2024)	569	244
capitalize (Todd et al., 2024)	569	244
country-capital (Todd et al., 2024)	137	60
country-currency (Todd et al., 2024)	137	60
english-french (Lample et al., 2018)	3288	1410
english-german (Lample et al., 2018)	3288	1410
english-spanish (Lample et al., 2018)	3639	1560
landmark-country (Hernandez et al., 2023)	585	251
lowercase_first_letter (Todd et al., 2024)	569	245
national_parks (Todd et al., 2024)	315	136
park-country (Todd et al., 2024)	524	225
person-sport (Hernandez et al., 2023)	222	96
present-past (Todd et al., 2024)	205	88
product-company (Hernandez et al., 2023)	365	157
singular-plural (Todd et al., 2024)	143	62
synonym (Nguyen et al., 2017)	2015	865

Table 1: All datasets being used. In experiments, we randomly sample k samples from # Train.

B Additional Behavioral Experiment Results

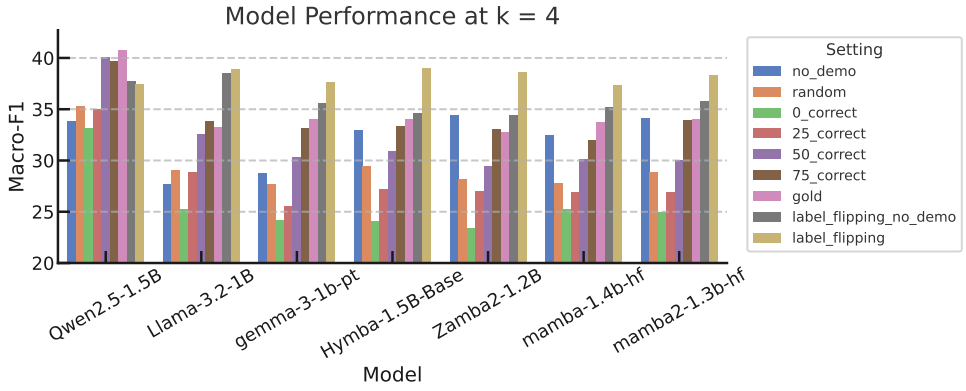

Figure 14: $k = 4$ results on contextual knowledge understanding datasets, including label flipping experiments.

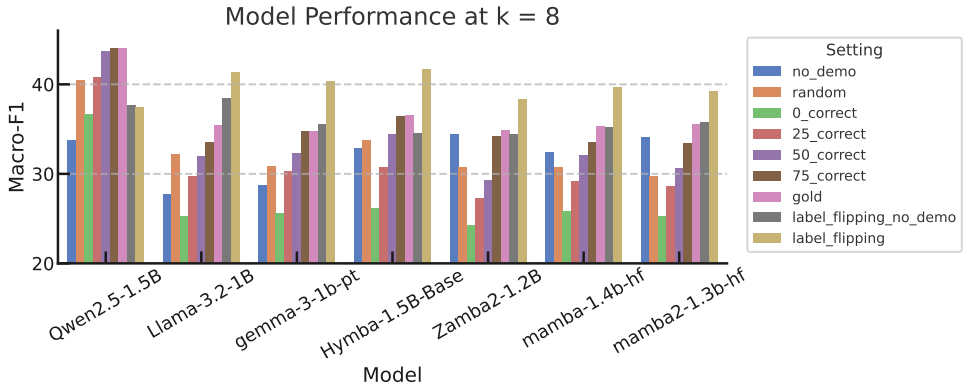

Figure 15: $k = 8$ results on contextual knowledge understanding datasets, including label flipping experiments.

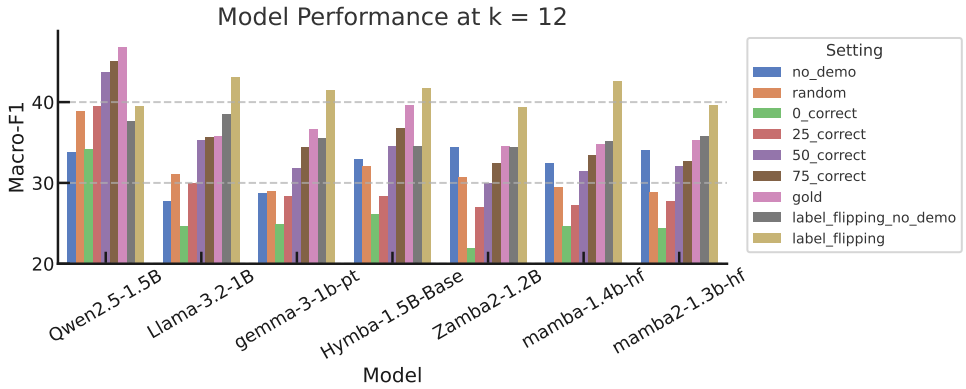

Figure 16: $k = 12$ results on contextual knowledge understanding datasets, including label flipping experiments.

setting	Qwen2.5-1.5B	Llama-3.2-1B	gemma-3-1b-pt	HYMBA-1.5B-BASE	ZAMBA2-1.2B	mamba-1.4b-hf	mamba2-1.3b-hf
no_demo	33.8	27.7	28.7	32.9	34.5	34.6	34.1
incorrect_mapping_no_demo	37.7	38.5	35.6	34.6	34.4	35.2	35.8
k = 4							
0_correct	33.1	25.2	24.2	24.1	23.4	25.2	24.9
25_correct	35.0	28.8	25.5	27.2	27.0	26.9	26.9
50_correct	40.1	32.5	30.3	30.9	29.4	30.1	30.0
75_correct	39.7	33.8	33.1	33.3	33.0	32.0	33.9
gold	40.7	33.2	34.0	34.0	32.7	33.7	34.0
random	35.3	29.0	27.7	29.4	28.2	27.8	28.8
incorrect_mapping	37.4	38.9	37.6	39.0	38.6	37.3	38.3
k = 8							
0_correct	36.7	25.3	25.6	26.2	24.3	25.8	25.3
25_correct	40.8	29.8	30.3	30.8	27.3	29.2	28.6
50_correct	43.7	32.0	32.3	34.4	29.3	32.1	30.6
75_correct	44.0	33.6	34.8	36.4	34.2	33.5	33.4
gold	44.0	35.4	34.8	36.6	34.9	35.3	35.6
random	40.5	32.2	30.9	33.8	30.8	30.8	29.8
incorrect_mapping	37.5	41.4	40.4	41.7	38.3	39.7	39.3
k = 12							
0_correct	34.2	24.7	24.9	26.1	21.9	24.6	24.4
25_correct	39.5	30.0	28.4	28.4	27.0	27.3	27.7
50_correct	43.7	35.3	31.8	34.5	30.0	31.4	32.1
75_correct	45.1	35.7	34.4	36.8	32.5	33.4	32.7
gold	46.8	35.8	36.7	39.6	34.6	34.8	35.3
random	38.9	31.1	29.0	32.1	30.7	29.5	28.9
incorrect_mapping	39.5	43.1	41.5	41.7	39.4	42.6	39.6
k = 16							
0_correct	34.2	24.9	24.2	25.2	22.9	26.7	24.1
25_correct	37.5	28.7	26.4	28.9	26.6	27.3	27.1
50_correct	41.6	33.1	32.0	33.4	30.6	31.2	31.3
75_correct	45.8	37.0	36.3	39.4	32.7	34.6	33.8
gold	48.6	38.1	37.9	39.6	35.0	36.4	34.5
random	38.3	30.8	28.4	33.3	30.1	30.0	29.0
incorrect_mapping	40.8	44.4	43.6	42.3	40.9	42.4	40.3
k = 32							
0_correct	29.0	23.9	21.8	23.5	22.9	23.8	23.8
25_correct	37.0	27.4	24.8	29.4	27.5	27.8	29.4
50_correct	45.1	34.3	33.2	34.9	32.1	33.5	31.6
75_correct	47.8	37.2	35.9	40.2	32.9	35.0	32.9
gold	49.3	38.8	37.4	39.8	36.2	36.6	34.9
random	40.8	33.7	31.3	35.8	31.7	32.1	31.4
incorrect_mapping	44.8	46.5	44.8	42.9	41.6	42.6	41.3

Table 2: experimental results of contextual knowledge understanding datasets, reported in Macro-F1.

setting	Qwen2.5-1.5B	Llama-3.2-1B	gemma-3-1b-pt	HYMBA-1.5B-BASE	ZAMBA2-1.2B	mamba-1.4b-hf	mamba2-1.3b-hf
no_demo	24.8	36.1	24.2	23.1	25.0	13.9	20.5
k = 4							
0_correct	12.3	19.5	27.0	13.2	19.4	22.1	22.0
25_correct	45.5	44.1	51.5	32.9	39.8	44.8	42.8
50_correct	60.8	59.4	64.8	48.6	53.0	54.5	53.4
75_correct	63.4	61.9	69.1	54.1	57.0	57.2	56.2
gold	64.1	66.9	70.2	58.3	59.2	58.8	58.7
random	16.9	20.1	27.1	13.7	22.2	25.7	24.2
k = 8							
0_correct	7.7	17.2	22.2	11.3	15.9	21.5	21.2
25_correct	33.5	37.7	47.6	29.9	36.3	40.7	39.9
50_correct	54.7	57.1	64.6	48.3	51.4	54.8	53.2
75_correct	64.8	67.2	70.8	57.6	59.4	61.0	59.2
gold	67.0	68.7	72.1	59.9	61.3	61.9	60.5
random	10.3	17.1	21.9	11.9	15.7	20.5	18.8
k = 12							
0_correct	6.0	12.9	18.7	10.8	15.7	19.7	17.9
25_correct	41.7	42.7	49.2	32.9	39.3	48.5	42.4
50_correct	61.3	63.6	65.9	50.8	54.4	58.6	55.2
75_correct	65.1	67.8	70.9	56.1	59.8	60.7	58.9
gold	67.6	70.0	72.8	60.1	62.0	62.4	60.9
random	8.4	15.5	19.1	9.5	14.9	19.7	19.9
k = 16							
0_correct	6.6	14.7	18.7	8.6	13.1	20.9	18.7
25_correct	40.9	41.4	48.5	34.3	40.0	47.4	42.1
50_correct	60.7	64.0	65.1	51.8	55.9	58.8	56.0
75_correct	66.1	68.7	71.3	57.3	59.8	61.3	59.3
gold	68.0	69.6	72.7	60.4	62.6	62.4	60.7
random	7.5	14.7	19.4	10.9	15.1	19.7	19.3
k = 32							
0_correct	5.2	13.2	15.6	7.1	11.6	19.3	15.8
25_correct	40.6	46.0	43.2	32.2	42.2	49.4	41.5
50_correct	61.9	64.9	64.8	52.7	55.7	59.2	54.4
75_correct	66.9	69.0	71.2	58.1	60.9	61.7	59.5
gold	68.4	69.8	72.9	61.1	63.0	62.8	61.5
random	7.7	13.5	16.7	9.1	12.7	20.4	17.6

Table 3: Experimental results for parametric knowledge retrieval datasets, reported in Macro-F1.

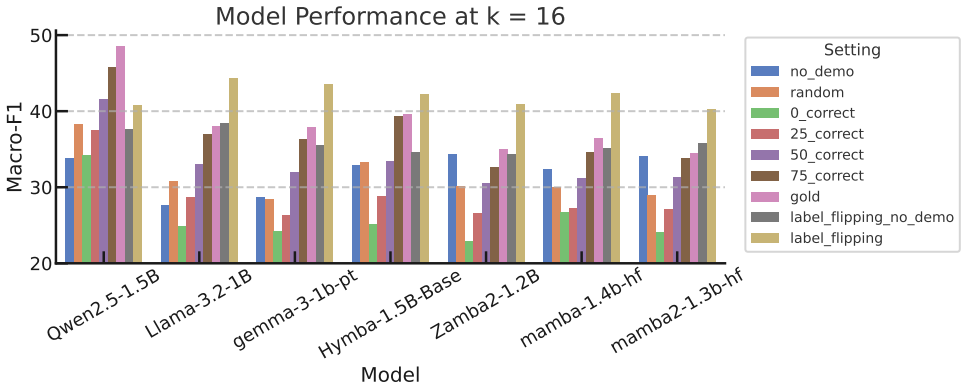

Figure 17: $k = 16$ results on contextual knowledge understanding datasets, including label flipping experiments.

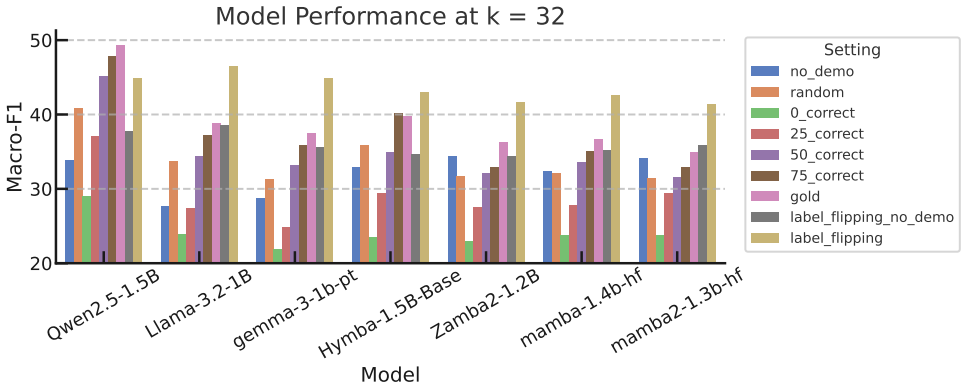

Figure 18: $k = 32$ results on contextual knowledge understanding datasets, including label flipping experiments.

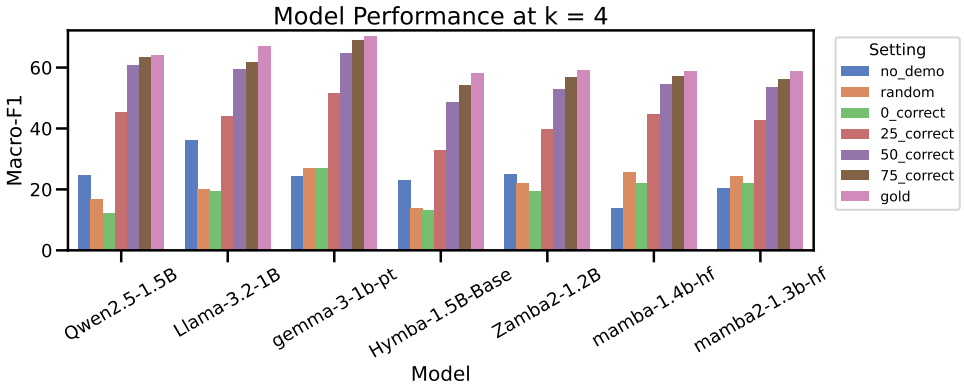

Figure 19: $k = 4$ results on parametric knowledge retrieval datasets, on initial setting.

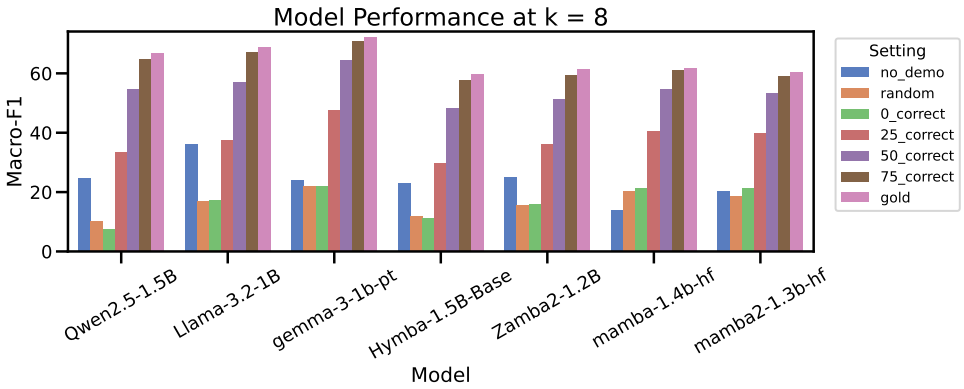

Figure 20: $k = 8$ results on parametric knowledge retrieval datasets, on initial setting.

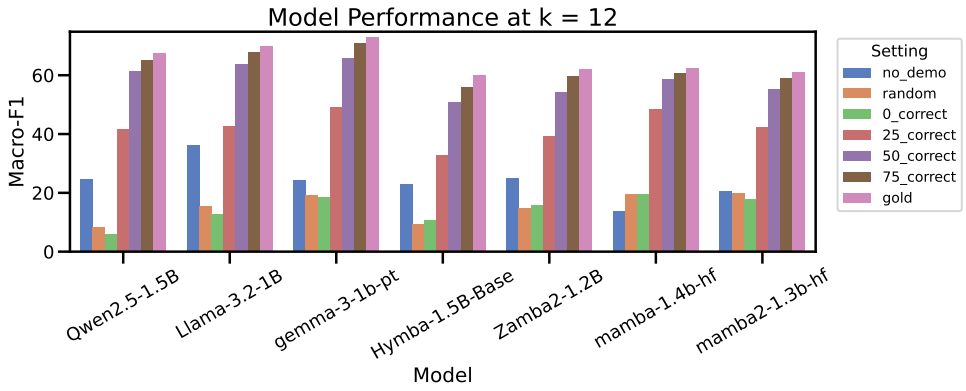

Figure 21: $k = 12$ results on parametric knowledge retrieval datasets, on initial setting.

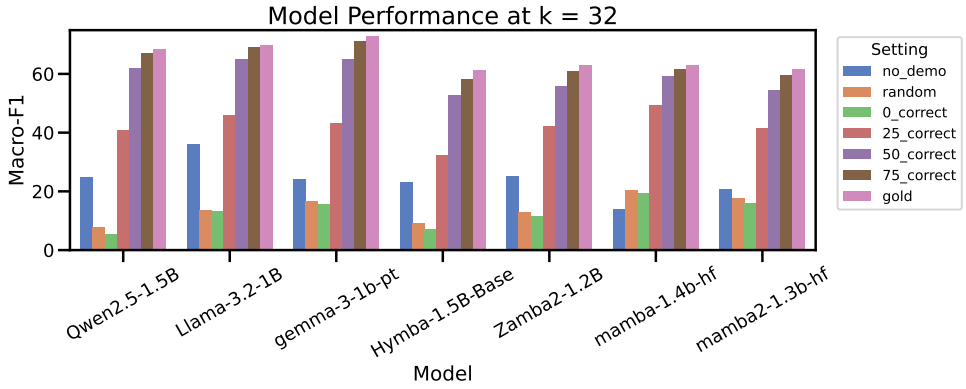

Figure 22: $k = 32$ results on parametric knowledge retrieval datasets, on initial setting.

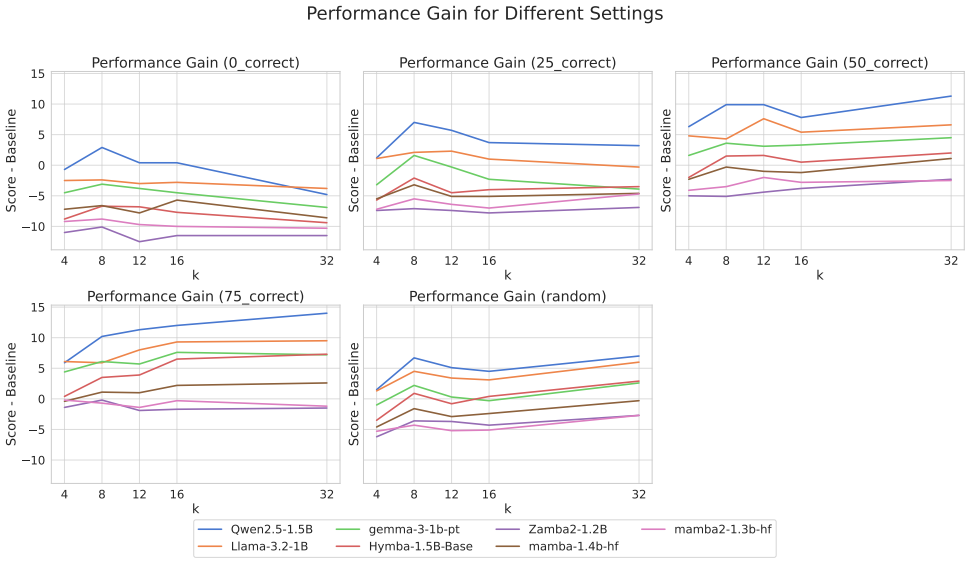

Figure 23: Performance gain of other conditions on contextual knowledge understanding datasets.

C Additional Mechanistic Interpretability Experiment Results

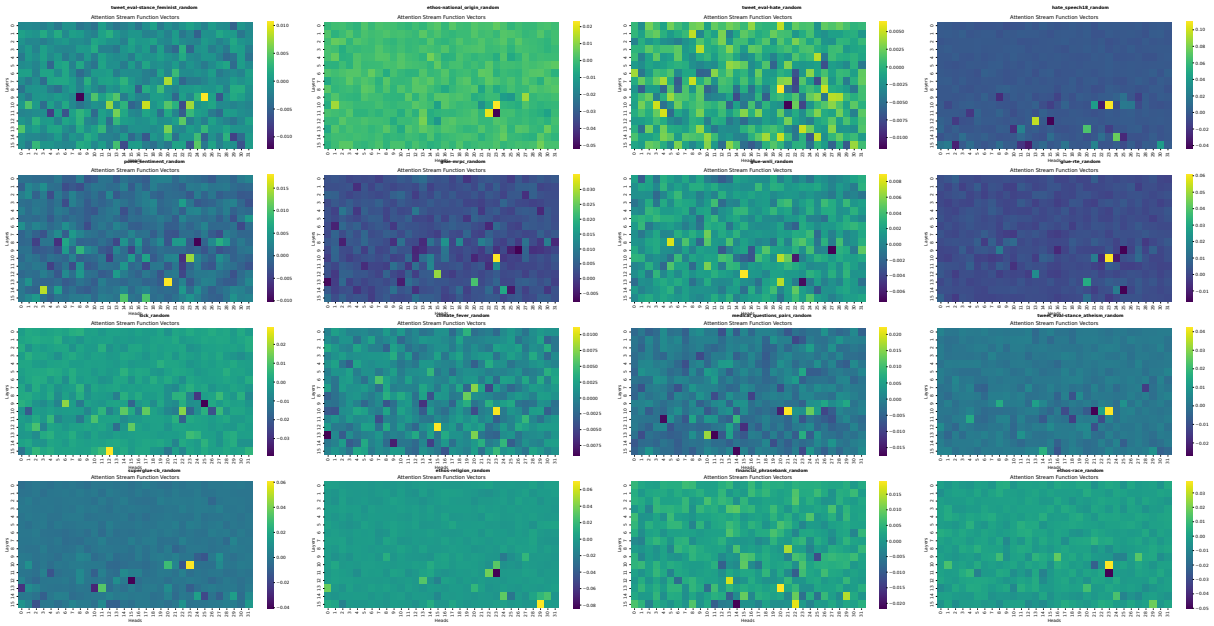

Figure 24: Llama-3.2-1B's AIE heatmaps on contextual knowledge understanding datasets.

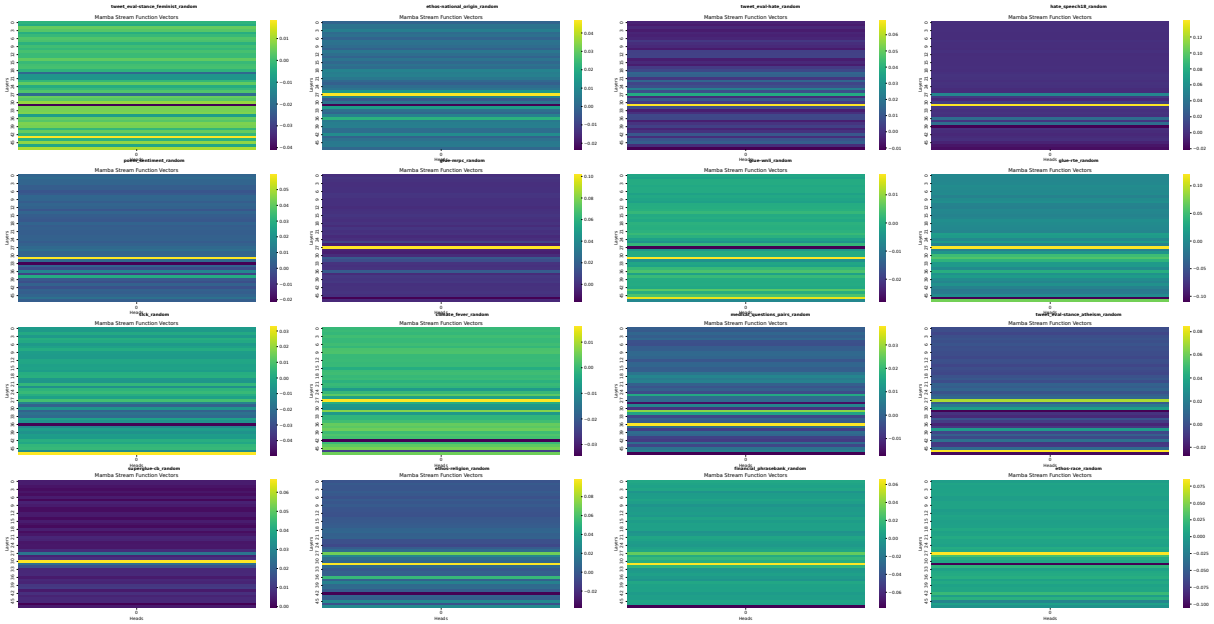

Figure 25: mamba-1.4b-hf's AIE heatmaps on contextual knowledge understanding datasets.

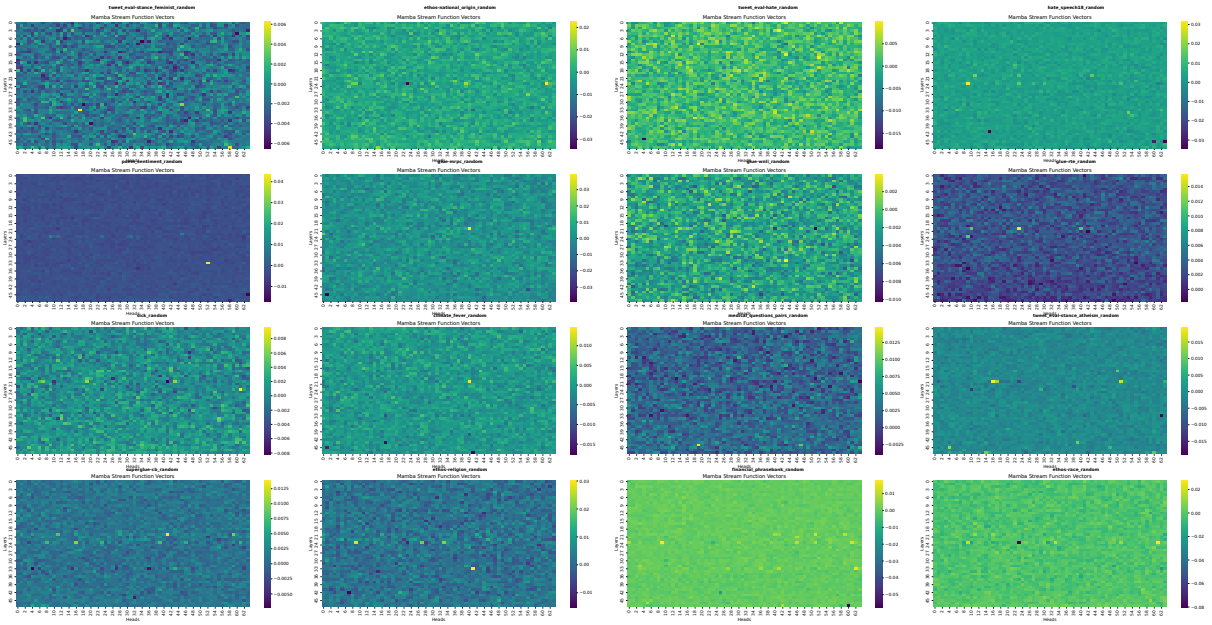

Figure 26: mamba2-1.3b-hf's AIE heatmaps on contextual knowledge understanding datasets.

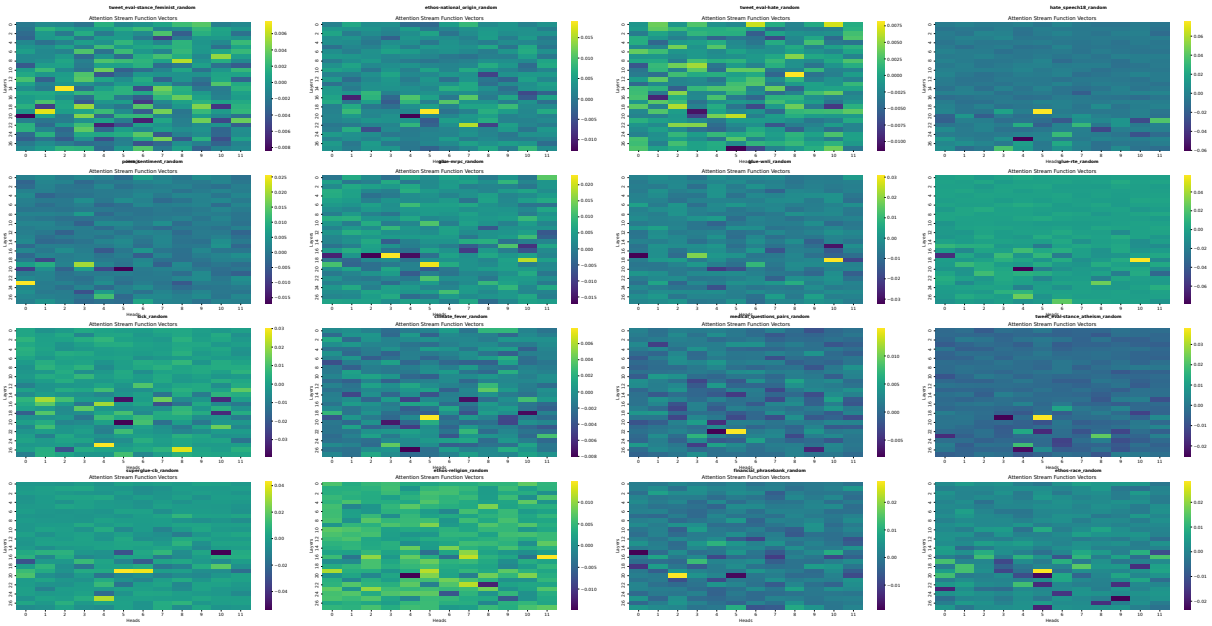

Figure 27: Qwen2.5-1.5B's AIE heatmaps on contextual knowledge understanding datasets.

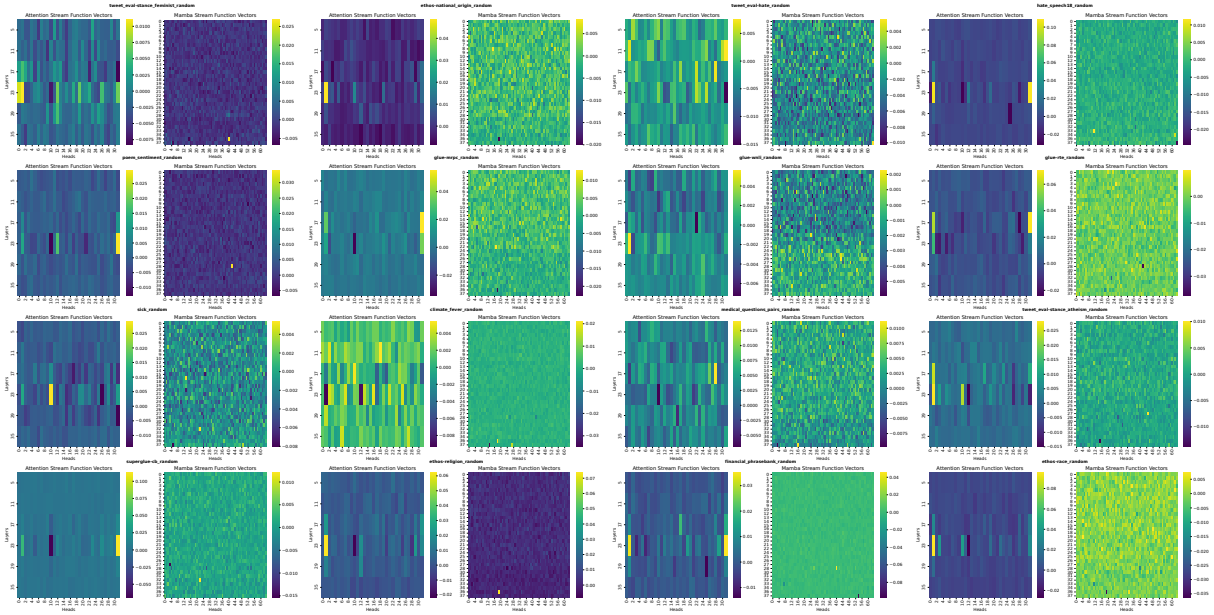

Figure 28: ZAMBA2-1.2B's AIE heatmaps on contextual knowledge understanding datasets.

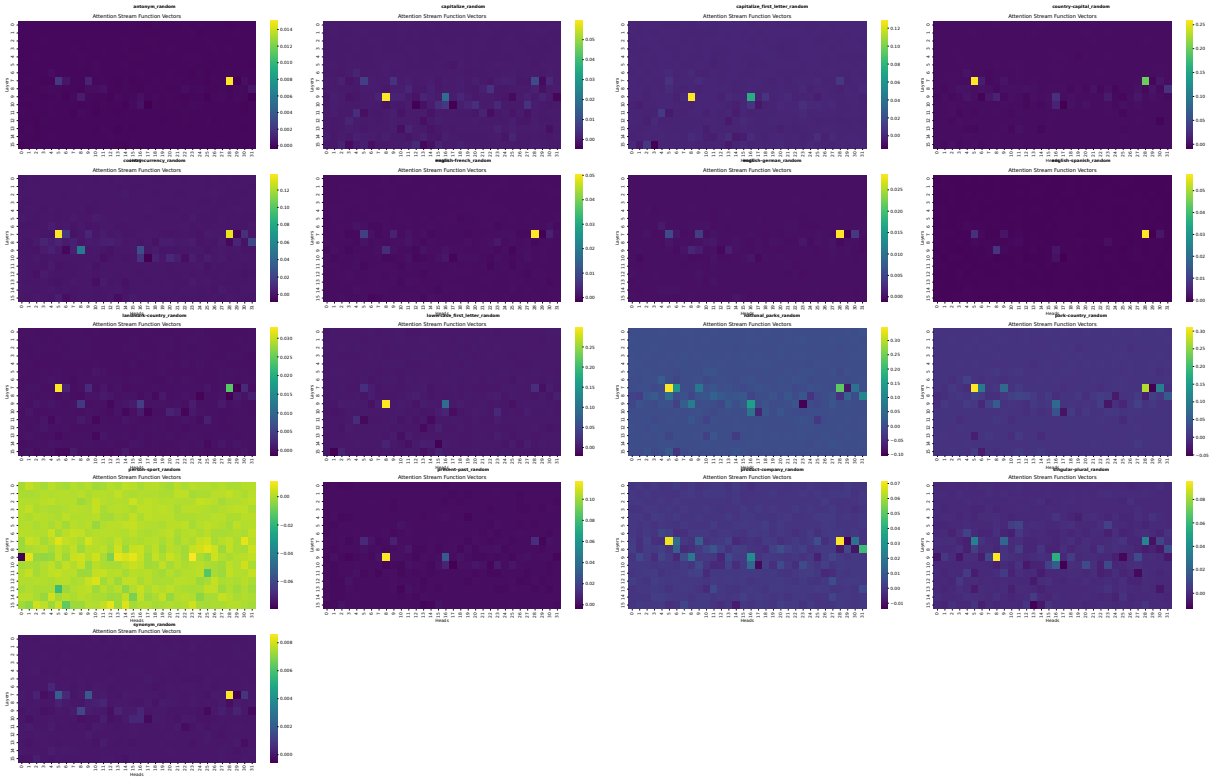

Figure 29: Llama-3.2-1B's AIE heatmaps on parametric knowledge retrieval datasets.

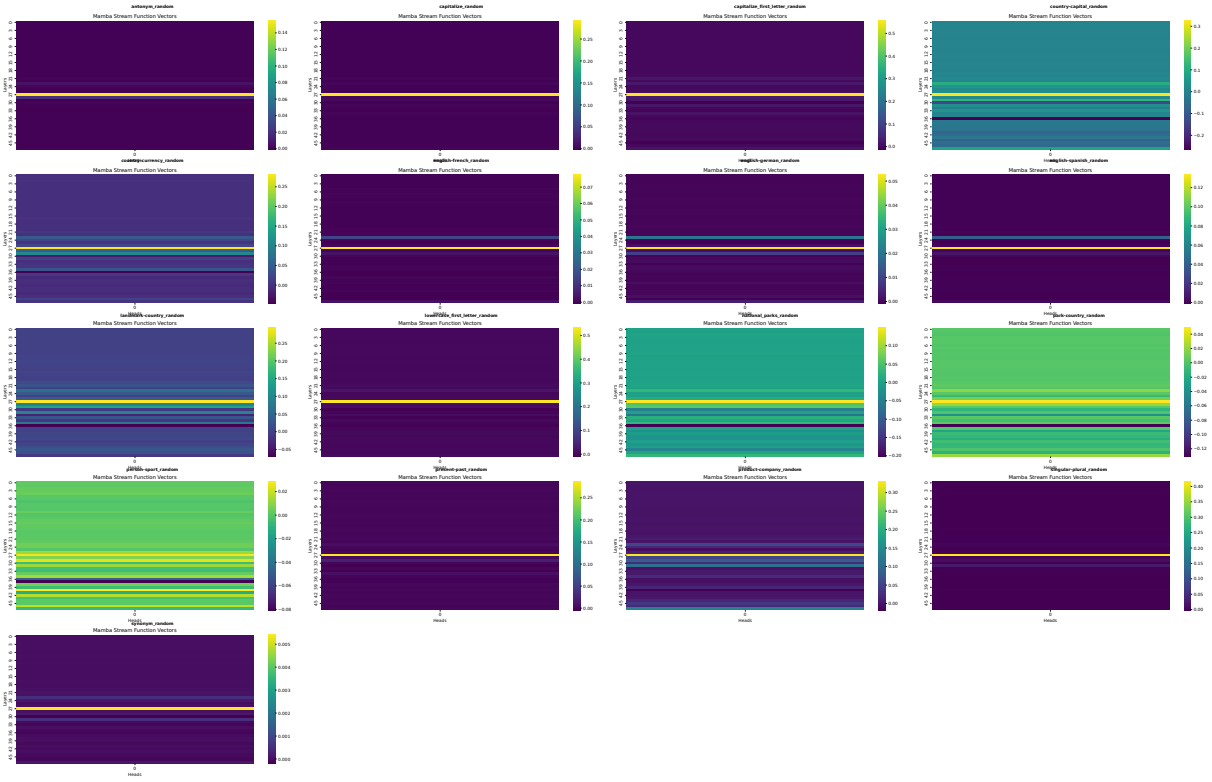

Figure 30: mamba-1.4b-hf's AIE heatmaps on parametric knowledge retrieval datasets.

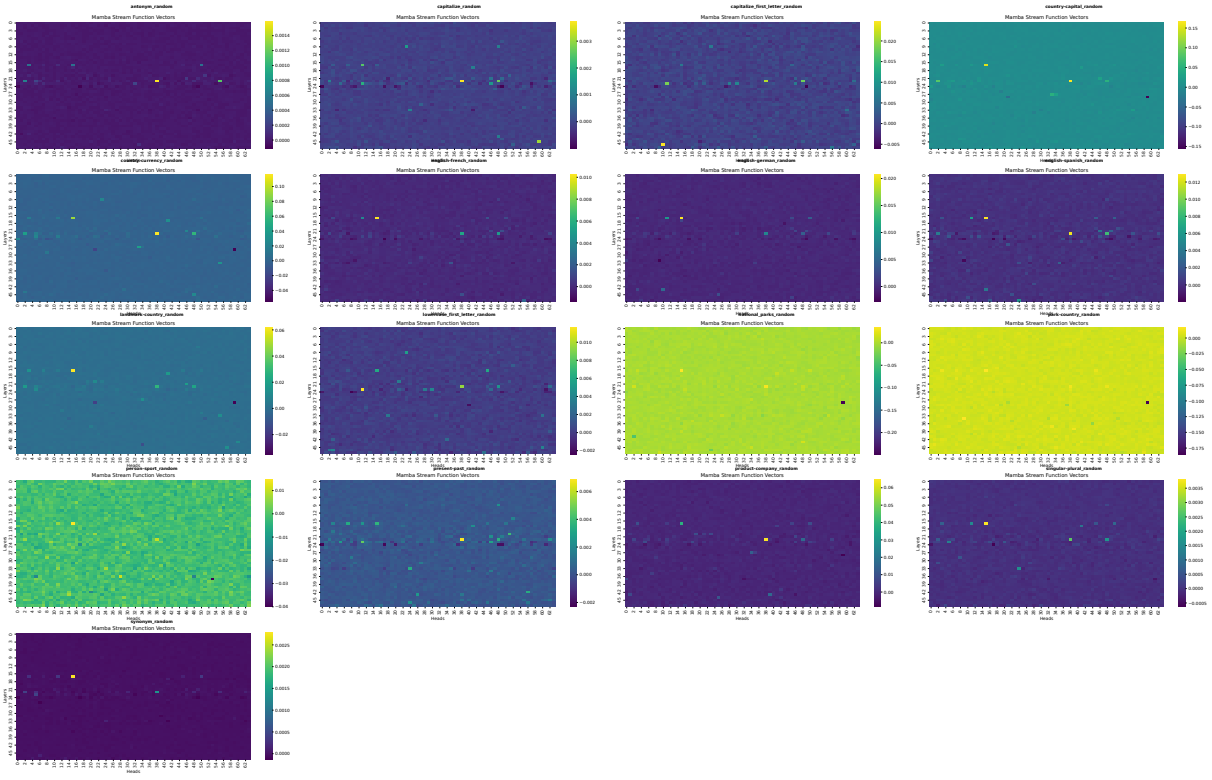

Figure 31: mamba2-1.3b-hf's AIE heatmaps on parametric knowledge retrieval datasets.

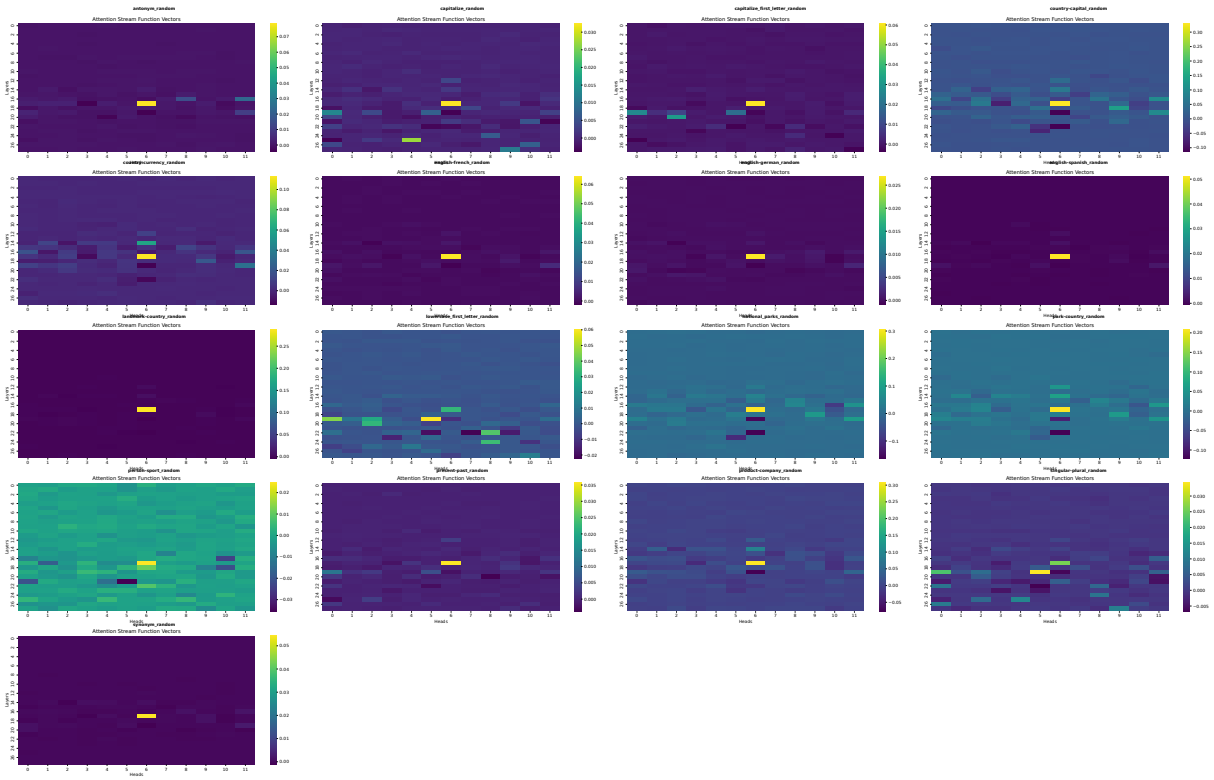

Figure 32: Qwen2.5-1.5B's AIE heatmaps on parametric knowledge retrieval datasets.

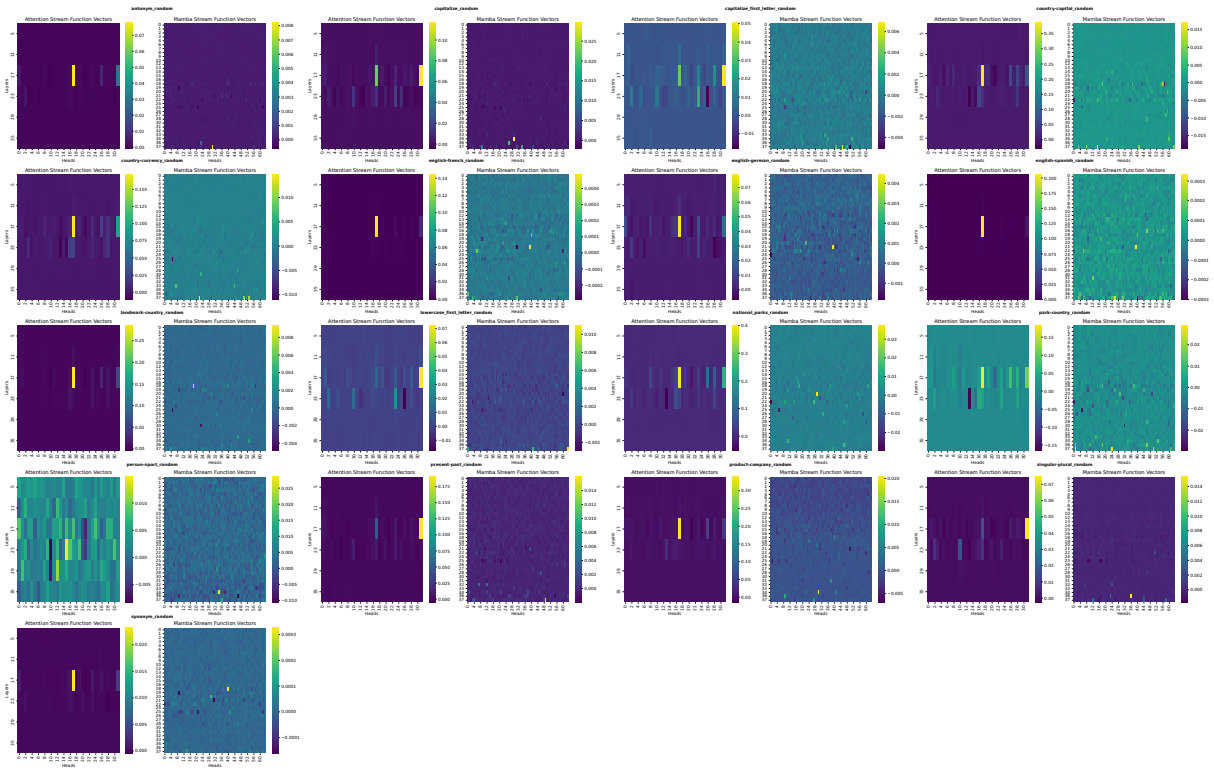

Figure 33: ZAMBA2-1.2B's AIE heatmaps on parametric knowledge retrieval datasets.

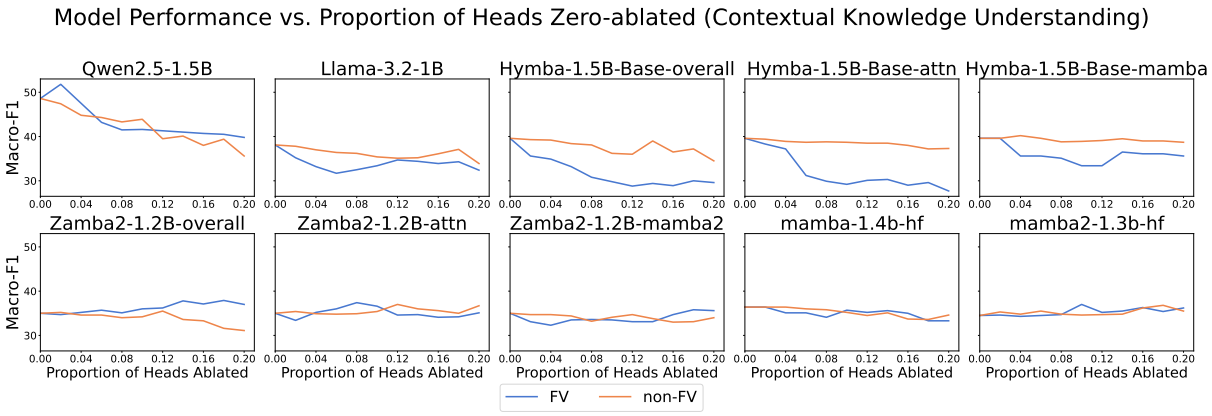

Figure 34: Zero ablation results on contextual knowledge understanding datasets.

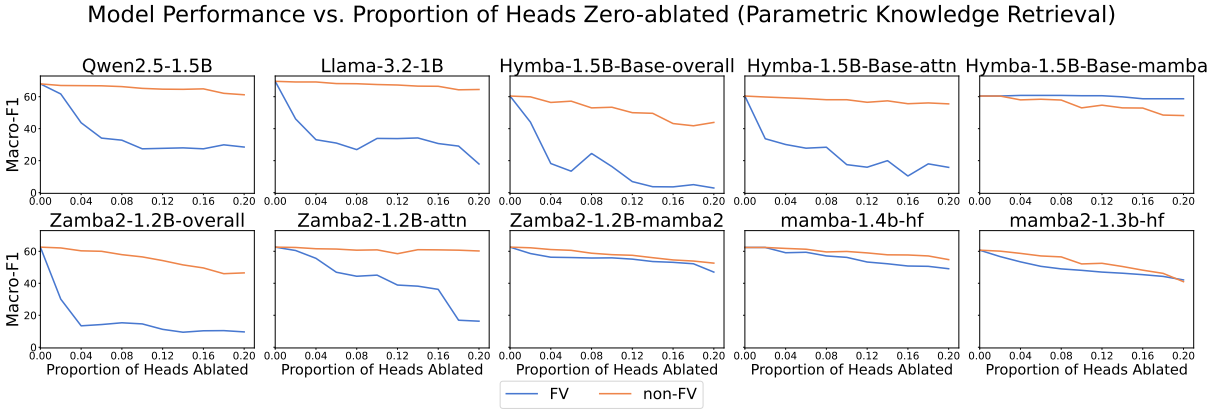

Figure 35: Zero ablation results on parametric knowledge retrieval datasets.

top-p	Qwen2.5-1.5B	Llama-3.2-1B	HYMBA-1.5B-BASE	ZAMBA2-1.2B	mamba-1.4b-hf	mamba2-1.3b-hf
0.02	0.00%	0.00%	18.75%	18.18%	-	9.84%
0.04	7.69%	10.00%	12.12%	12.36%	100.00%	10.66%
0.06	20.00%	16.67%	26.53%	14.18%	50.00%	10.33%
0.08	26.92%	15.00%	24.24%	16.20%	33.33%	13.06%
0.10	21.21%	15.69%	31.33%	17.41%	25.00%	13.03%
0.12	20.00%	18.03%	31.31%	16.79%	40.00%	15.49%
0.14	17.02%	19.72%	29.31%	18.21%	33.33%	15.81%
0.16	20.75%	17.28%	28.57%	19.27%	28.57%	17.52%
0.18	21.67%	19.57%	28.19%	20.35%	25.00%	19.75%
0.20	22.39%	18.63%	30.12%	22.77%	33.33%	20.85%

Table 4: Percentage of intersection between top FV heads identified from contextual knowledge understanding datasets and parametric knowledge retrieval datasets. Entries are the percentage of heads in top-p that overlap.

D Additional Hymba Results

To analyze the less consistent behavior in HYMBA-1.5B-BASE's in steering and ablation experiments, we introduce the negative AIE experiment: on the gold-labelled dataset, we mean-ablated each head to observe how much drop in the logit confidence this ablation introduced. We present our results in Figures 36 and 37. Surprisingly, we find the results to be extremely consistent on the parametric knowledge retrieval tasks: the Mamba head on the very first layer always cause the most significant drop in performance. Nevertheless, results on the contextual understanding tasks, as opposed to the other, appear to be less consistent. The attention stream seemed to have played a more important role in contextual knowledge understanding, as ablating attention heads caused a more significant drop in AIE magnitude compared to ablating the mamba stream.

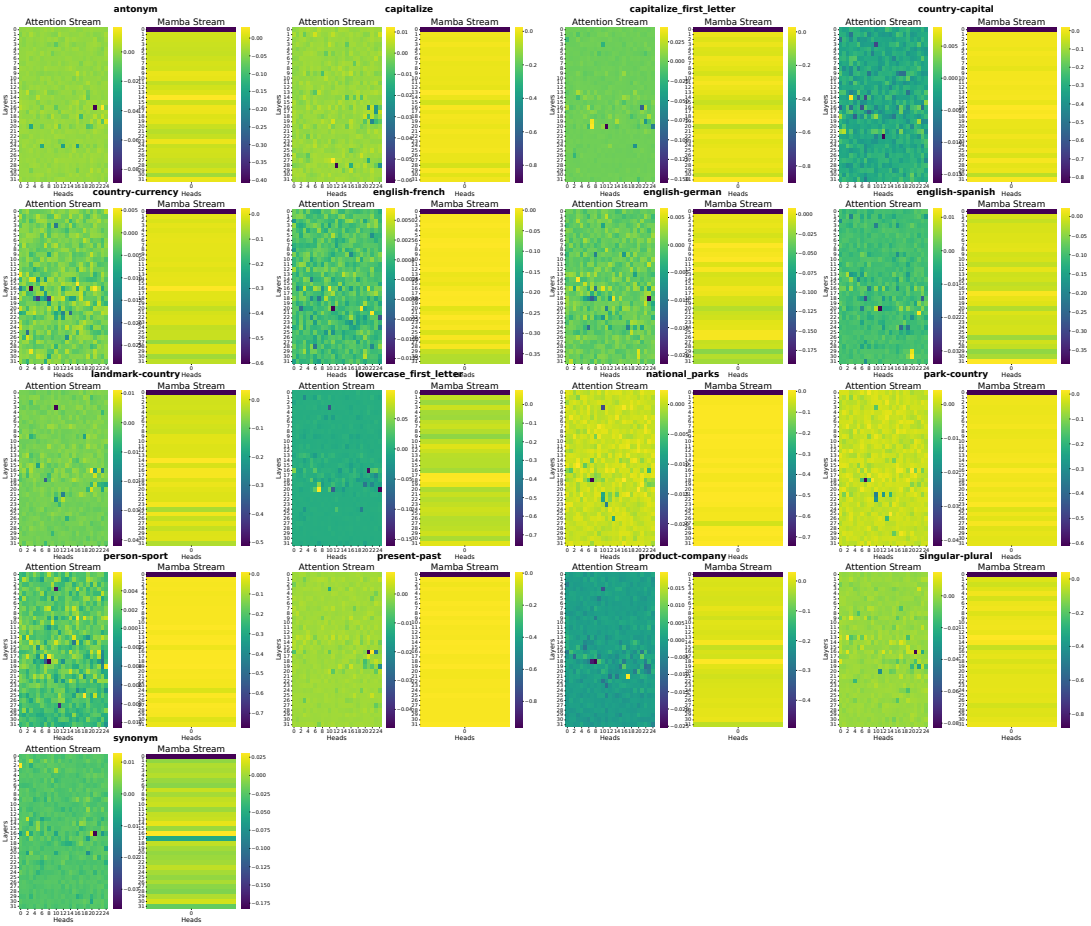

Figure 36: Negative AIE experiment results for HYMBA-1.5B-BASE on parametric knowledge retrieval datasets. Darker heads are those that caused a greater drop in performance upon removal.

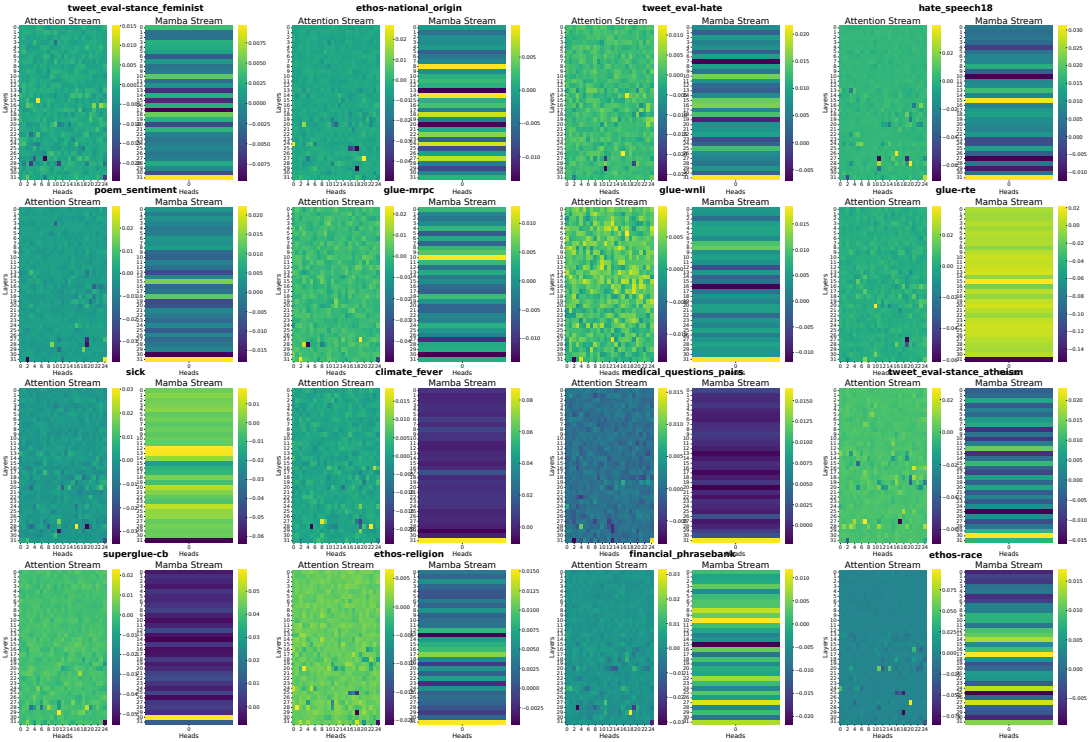

Figure 37: Negative AIE experiment results for HYMBA-1.5B-BASE on contextual knowledge understanding datasets. Darker heads are those that caused a greater drop in performance upon removal.