Quantifying Logical Consistency in Transformers via Query-Key Alignment

Eduard Tulchinskii[1,2], Anastasia Voznyuk[3], Laida Kushnareva[2], Andrei Andriiainen[1,3], Irina Piontkovskaya[2], Evgeny Burnaev[1,5], Serguei Barannikov[1,4],

[1]Skolkovo Institute of Science and Technology, [2]AI Foundation and Algorithm Lab
[3]Moscow Institute of Physics and Technology, [4]CNRS, Université Paris Cité, France
[5]Artificial Intelligence Research Institute (AIRI)

Abstract

Large language models (LLMs) excel at many NLP tasks, yet their multi-step logical reasoning remains unreliable. Existing solutions such as Chain-of-Thought prompting generate intermediate steps but provide no internal check of their logical coherence. In this paper, we use the "QK-score", a lightweight metric based on query–key alignments within transformer attention heads, to evaluate the logical reasoning capabilities of LLMs. Our method automatically identifies attention heads that play a key role in distinguishing valid from invalid logical inferences, enabling efficient inference-time evaluation via a single forward pass. It reveals latent reasoning structure in LLMs and provides a scalable mechanistic alternative to ablation-based analysis. Across three benchmarks: ProntoQA-OOD, PARARULE-Plus, and MultiLogicEval, with models ranging from 1.5B to 70B parameters, the selected heads predict logical validity up to 14% better than the model probabilities, and remain robust under distractors and increasing reasoning depth of $d \leq 6$.

1 Introduction

Large language models (LLMs) have achieved remarkable success in a range of NLP tasks, yet they still struggle with reliable multi-step logical reasoning (Wei et al., 2022; Kojima et al., 2022; Yang et al., 2024; Seals and Shalin, 2023; Wan et al., 2024). While chain-of-thought prompting has improved performance by allowing models to generate intermediate reasoning steps, it lacks a mechanism to assess the coherence of these transitions.

In this work, we propose a novel evaluation method that uses internal Query-Key (QK) vectors interactions within transformer heads as a proxy for logical consistency. For a triplet (context c, statement s, and candidate answer a_i), where $a_0 =$ true or $a_1 =$ false, we define the QK-score as:

$$S_{QK}^{(l,h)}(c, s, a_i) = \boldsymbol{q}_{a_i}^{(l,h)\top} \boldsymbol{k}_s^{(l,h)},$$

where $\boldsymbol{q}_{a_i}^{(l,h)}$ is the query vector for the answer token and $\boldsymbol{k}_s^{(l,h)}$ is the key vector corresponding to the statement. Our method efficiently identifies transformer heads capable of accurately evaluating the validity of logical transitions. It processes all heads in a single run, making it fast and scalable.

The contributions of this paper are as follows:

- We propose a novel QK-score mechanism that uses the interactions between certain query- and key-vectors to assess the correctness of logical transitions in the corresponding tasks.
- We identify particular attention heads responsible for solving various types of logical problems across several LLMs, enhancing our understanding of how transformers perform logical reasoning. We observe that individual "logic heads" (such as (22, 26) in DeepSeek-R1-Distill-Qwen-7B) generalise across datasets, acting as model-internal verification anchors. Several heads that are selected by their QK-score on one logical task, perform quite well on another ones, showing some degree of universality.
- We conduct extensive experiments with models ranging from 1.5B to 70B parameters on three logical benchmarks and demonstrate that our single forward-pass probe predicts logical validity by up to 14% better compared to model itself, including multi-step reasoning scenarios (with ≤ 6 steps) or scenarios when context contains reasoning distractors.

2 Related Work

A significant line of research has focused on improving multi-step logical reasoning in LLMs via chain-of-thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022; Yang et al., 2024; Seals and Shalin, 2023; Wan et al., 2024). Although CoT

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35185–35200
November 4-9, 2025

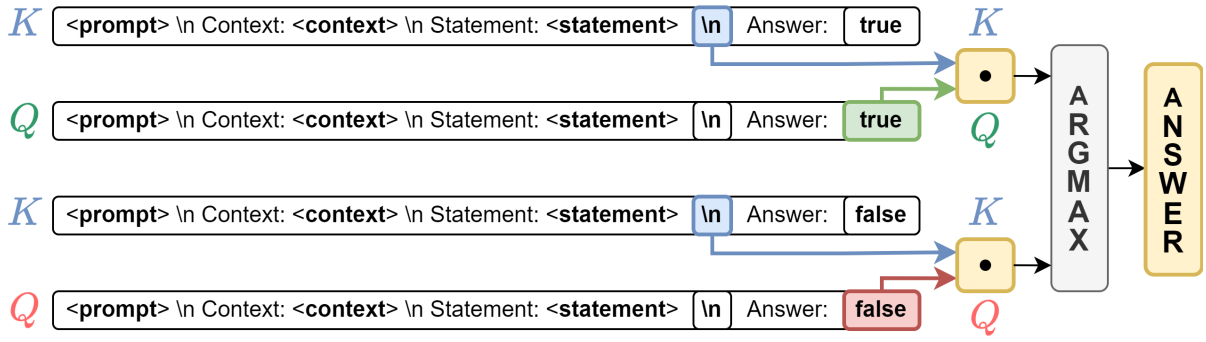


Figure 1: Our method calculates the Query-Key score between the end-of-line token immediately after the statement and the "true"/"false" tokens, for the designated head, from which we derive the answer.

methods allow models to generate intermediate reasoning steps, they lack a mechanism to assess the coherence of these logical transitions.

Mechanistic interpretability studies have examined the roles of transformer attention heads. Recent work revealed that attention layers operate in phases—knowledge recalling, latent reasoning, and expression preparation (Elhage et al., 2021; Olah et al., 2020; Zheng et al., 2024). Subsequent studies have shown that some attention heads introduce biases by evenly splitting probabilities between answer options (Lieberum et al., 2023; Yu et al., 2024), while others suppress such behaviors during the final expression phase (Kim et al., 2024). In addition, recent investigations have attempted to analyze model behavior by disabling specific components (Zhang and Nanda, 2024; Todd et al., 2024; Yao et al., 2024), though these approaches are often computationally expensive or limited to simpler tasks (Wang et al., 2023; Yao et al., 2024).

For an expanded discussion on related works, see Appendix A.

3 Approach

Our method evaluates logical consistency by examining certain internal Query-Key (QK) interactions within transformer heads. In our setup, each input consists of a context c (which provides the premises), a statement s (a candidate conclusion), and a candidate answer a_i (with a_0 = true and a_1 = false), see Figures 2, 4, 5 for examples.

For every attention head, indexed by h in layer l, we compute a QK-score that quantifies the alignment between the representation of the statement and that of the candidate answer. Given the triplet (c, s, a_i), the **QK-score** is the dot-product $S_{QK}^{(l,h)}(c, s, a_i) = \boldsymbol{q}_{a_i}^{(l,h)\top} \boldsymbol{k}_s^{(l,h)}$, where $\boldsymbol{q}_{a_i}^{(l,h)}$ is the query vector associated with the token representing a_i (either "true" or "false") and $\boldsymbol{k}_s^{(l,h)}$ is the key vector corresponding to the token marking the end of the statement (see Figure 1). This score reflects how well a head can distinguish valid logical transitions.

By evaluating all heads in a single forward pass, our approach identifies those that reliably assess logical consistency. This efficient procedure avoids the need for extensive model modifications or head-by-head ablation studies that are common in prior work.

Use provided context and answer whether the statement is true or false.
CONTEXT: Numpuses are zumpuses. Each zumpus is not opaque. Shumpuses are numpuses. Every jompus is opaque. Impuses are shumpuses. Polly is an impus. Polly is a yumpus.
STATEMENT: Polly is opaque.
ANSWER: true.

Figure 2: PrOntoQA-OOD Example

4 Experiments

4.1 Datasets

We evaluate our method on three diverse benchmarks for logical reasoning:

ProntoQA-OOD (Saparov et al., 2023) is a synthetic dataset for chain-of-thought first-order logic question-answering over abstract categories (see Figure 2). We consider two deduction settings: (1) tasks requiring only repeated applications of Modus Ponens; and (2) tasks involving all six propositional logic rules supported in the existing ProntoQA-OOD benchmark. For each setup, we partition the data by reasoning depth (how many times a logic rule must be applied to the premises and results of previous steps to derive the

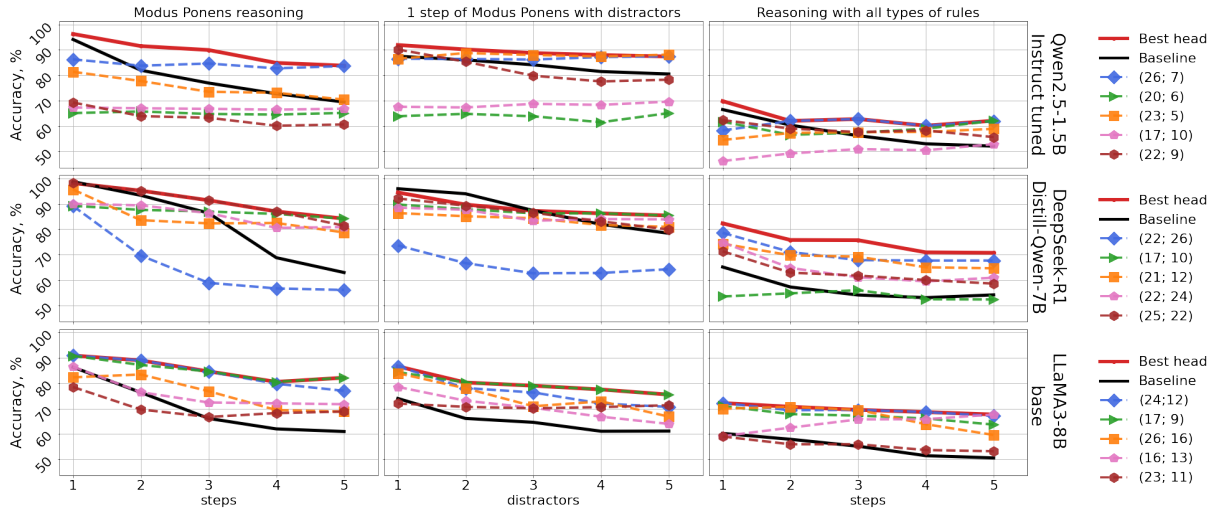


Figure 3: In-domain performance on ProntoQA-OOD. Best heads are selected per case using calibration data. The notation (26, 7) means the seventh head from 26th layer and so on.

answer), selecting 600 examples (300 "true" and 300 "false") for calibration and using the remaining 1,000+ examples for evaluation. To further test robustness, we adapt the scripts to vary the number of distractors (irrelevant context statements).

PARARULE Plus (Bao et al., 2022) is a dataset of true/false questions with fixed-depth reasoning. We use its test subset without modifications.

Extended-Multi-LogiEval builds on Multi-LogiEval (Patel et al., 2024). The original dataset is constrained by only 10 samples per logical scheme and an imbalanced answer distribution. We address these limitations by generating additional samples (100 per scheme) and enforcing a balanced 50/50 split for "yes" and "no" responses (treated as a_0 and a_1, respectively, for QK-score computation).

All experiments are performed in a zero-shot setup. Further details are provided in Appendices B and D.

4.2 Experimental Setup

All experiments were performed with frozen pre-trained LLMs of size 1.5B to 70B (for models larger than 7B, see Appendix E).

Our questions assume single-word answers; the standard approach in such setup is to select from a_0 and a_1 the option that was assigned the highest output probability by LLM, when the models are given the samples and asked to provide an answer. We refer to this method as the **BASELINE**.

Our experiments were performed in two steps. First, for each explored setup from ProntoQA-OOD, we chose the best head in terms of QK-score on calibration set. Then we report the accuracy of QK-Scoring via the chosen head and the accuracy of BASELINE on the evaluation subset.

Second, we selected five heads from those that achieved the top-10 performance in various ProntoQA-OOD setups (we aimed to choose heads that cover higher number of setups). Then, we evaluated their performance on the PARARULE Plus and Extended-Multi-LogiEval datasets to assess the generalization capabilities of the QK-score informed head selection.

5 Results

5.1 In-Domain Evaluation

For each deduction setup in ProntoQA-OOD, we select the best-performing head on the calibration set (referred to as BEST HEAD) and evaluate its performance on the held-out evaluation set. Figure 3 shows that the BEST HEAD consistently outperforms the baseline across varying reasoning depths and under different numbers of distractors. In addition, we report the performance of five individual heads drawn from the top-five performers in each setup. These results indicate that the selected heads maintain stable performance regardless of the number of reasoning steps, thereby confirming that our QK-score reliably identifies transformer components that contribute to logical consistency. Results for three LLMs are presented here; further details are provided in Appendix E and F.)

		PARARULE Plus Reasoning depth				Extended-Multi-LogiEval Reasoning depth			
	Head	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**
Qwen2.5	(26, 7)	0.6043	0.6772	0.6483	0.7095	0.4730	0.4663	0.5163	0.4413
1.5B, instruct	(20, 6)	0.5310	0.5844	0.5436	0.6097	**0.6069**	**0.6063**	**0.5804**	**0.5516**
	(23, 5)	0.5999	0.6488	0.6468	0.6694	0.5026	0.5587	0.4859	0.3840
	(17, 10)	0.4552	0.5569	0.4800	0.4549	0.5003	0.5165	0.5413	0.4943
	(22, 9)	**0.7095**	**0.7529**	**0.7719**	**0.8003**	0.2445	0.4531	0.4141	0.4527
Baseline	-	0.6240	0.6263	0.6521	0.6423	0.4996	0.4834	0.5011	0.5057
DeepSeek-R1	(22, 26)	**0.6736**	**0.7418**	0.5495	0.6145	0.5236	0.5112	0.5402	0.5115
Distill-Qwen-7B	(22, 16)	0.3206	0.3422	0.2763	0.2388	0.4689	0.5271	0.5185	0.5143
	(21, 12)	0.4982	0.5203	0.4452	0.4740	**0.7682**	**0.5733**	**0.5989**	**0.6218**
	(22, 24)	0.6523	0.6454	**0.6149**	**0.6557**	0.6572	0.5495	0.5771	0.5888
	(25, 22)	0.6227	0.6233	0.6145	0.5357	0.5093	0.4768	0.4902	0.5186
Baseline	-	0.4969	0.5212	0.4523	0.4717	0.5153	0.4835	0.5010	0.5057

Table 1: Performance of QK-score on heads selected on ProntoQA-OOD in cross-domain evaluation. PARARULE Plus and Extended-Multi-LogiEval datasets are used. Best results are highlighted in **bold**. Those results that are better than baseline are underlined

5.2 Transfer Learning Evaluation

We further assessed the cross-dataset generalization of our approach by evaluating the QK-scoring accuracy of the heads selected in the above experiments on two additional datasets: PARARULE Plus and Extended-Multi-LogiEval. As reported in Table 1, three out of five selected heads consistently achieve accuracy exceeding the baseline on PARARULE Plus for DeepSeekR1 Distill Qwen-7B, namely (22, 26), (22, 24), and (25, 22). Note that for smaller reasoning depths, the accuracy gap exceeds 10%. The heads (22, 24) and (22, 26) also consistently outperform the baseline on Extended-Multi-LogiEval, together with head (21, 12). For Qwen 2.5-1.5B-Instruct, only one head consistently outperforms the baseline across all reasoning depths of PARARULE Plus—head (22, 9)—and on Extended-Multi-LogiEval—head (20, 6)—which may be explained by smaller size of this model. Also, it is possible to use a mixture of datasets to perform a selection of heads; we explore this possibility in Appendix G .

Overall, these results demonstrate that our QK-score method identifies transformer heads capable of reliably assessing logical transitions, with robust performance observed both within the ProntoQA-OOD domain and in cross-dataset setups.

5.3 Ablation study

To analyze the effect on model reasoning capabilities that heads with high QK-score have, we performed the following intervention experiment: on ProntoQA-OOD dataset we evaluated performance of a model when its K best attention heads are pruned (i.e., their outputs are zeroed out). We compared it with the case when K random heads are pruned. Table 2 provides the results of this comparison for model LLaMA-3.1-8B and $K = 10, 20$; averaging was done over 7 restarts. The full results for all setups from ProntoQA-OOD are given in Appendix H.

It can be seen that deletion of the 'logic-related' heads detected by our method often causes greater disruption to the model's performance than deletion of the same number of randomly selected heads. This effect is clearly observed for reasoning with pure Modus Ponens; however, for reasoning with composition of rules, or for Modus Ponens with distractors (multiple irrelevant premises) we can sometimes see that models performance slightly improves after removal of a few best heads. We hypothesize that this might happen because of a contradiction between different tasks: various logic rules or logic reasoning vs. filtering irrelevant information. The relatively small overall magnitude of the performance drop suggests that other parts of the LLM, beyond the selected heads, may contribute to reasoning. These components remain functional even when the 'logic-related' heads are ablated.

Ruleset	Depth	Distractors	No pruning	# of heads pruned			
				10 best	10 random	20 best	20 random
Modus Ponens	1	0	86.32	81.44	84.66 ± 2.14	79.08	81.28 ± 4.26
	2	0	76.36	73.40	78.04 ± 3.56	73.89	75.99 ± 3.47
	1	1	86.32	73.37	83.80 ± 2.00	79.55	81.34 ± 2.74
	1	5	61.14	64.54	60.37 ± 2.39	64.74	66.49 ± 5.21
All rules	1	0	60.23	63.77	59.94 ± 1.25	62.31	57.95 ± 5.20
	2	0	57.91	57.55	58.42 ± 0.18	60.03	57.22 ± 7.40

Table 2: Performance of LLaMA-3-8B model on various setups from PRONTOQA-OOD after pruning a number of attention heads.

6 Analysis

Our findings suggest that QK-score offer a distinctive lens on how LLMs process logical structure. Unlike raw attention weights, QK-scores are independent of positional encoding, thereby focusing purely on semantic alignment between the statement and candidate answers. Moreover, heads selected via QK-scores often outperform the model's final probabilities, confirming that essential reasoning signals can be sometimes obscured by later-stage processing (Kim et al., 2024; Lieberum et al., 2023).

We also observe that high QK-scoring heads consistently identify valid inferences even with distractors in the input. This stability indicates that such heads act as "verification anchors," largely unaffected by irrelevant context. Consequently, QK-scores may bolster both interpretability and robustness: by highlighting the heads that preserve logical consistency, they help clarify how multi-step reasoning unfolds within LLMs.

The results of the ablation study show that removing top-10 highest QK-scoring heads from the model causes a greater disruption in performance compared to the removal of an equivalent number of randomly selected heads, on average for Modus Ponens rule. This suggests a causal relationship between the high QK-scoring heads and LLM logic processing capabilities.

For additional results on where 'logic' heads typically arise across specific layers of LLMs, see Appendix I.

7 Conclusion

We introduced a QK-score framework for finding attention heads that consistently capture logical validity. Our experiments show that in multi-step inferences with distractors, certain heads outperform the model's own predictions from the final layer. Crucially, this single-pass procedure sidesteps the computational overhead of ablation-based methods and generalizes relatively well across datasets. By identifying attention heads that act as "reasoning checkpoints," our approach offers an opportunity for better understanding into how these models process complex logical relationships. Future work may refine QK-scoring for specialized tasks, explore synergy with chain-of-thought prompting, and extend this analysis to other interpretability settings. Ultimately, bridging internal alignment signals with logical consistency represents a key step toward transparent and reliable LLMs.

8 Limitations

Our method requires a sufficiently large calibration dataset (at least ≈ 400 reasoning questions), with balanced coverage of logical rules and careful debiasing to avoid selecting heads that exploit how the question is phrased rather than the logic behind the question.

In our paper, we locate some "universal logic heads", however, there would be no single head that outperforms all the other heads on every possible logical task, as logical reasoning is an inherently complex task. Moreover, dataset-specific biases also affect the set of best performing heads, as well as the nuances of each LLM's architecture. Finally, we do not state that the heads identified by our method are the *only* heads responsible for the logical inference within the model.

Then, we conducted head selection on only two sets of deduction rules: "Modus Ponens" alone and a broader set including conjunction/disjunction, introduction/elimination, and proof by contradiction. Some heads may specialize in different logical inference rules, requiring more extensive experi-

ments with other logical rules.

Our QK-score method relies on multiplying the query and key vectors to compare the elements of the resulting matrix corresponding to specific tokens, as this directly aligns with the core operations of transformer inference. However, it does not directly incorporate information from value vectors or output aggregation matrices, which are also important components of the attention mechanism. Studying the role of these components in logical reasoning requires developing different approaches, which falls outside the scope of this paper.

Finally, in this work, we focused only on locating individual heads rather than on the head combinations, which is left for future work.

Acknowledgments

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4F0002 and the agreement with Skoltech №139-10-2025-033.

References

Qiming Bao, Alex Yuxuan Peng, Tim Hartill, Neset Tan, Zhenyun Deng, Michael Witbrock, and Jiamou Liu. 2022. Multi-step deductive reasoning over natural language: An empirical study on out-of-distribution generalisation. In *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022)*, pages 202–217, Cumberland Lodge, Windsor Great Park, United Kingdom.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Geonhee Kim, Marco Valentino, and André Freitas. 2024. A mechanistic interpretation of syllogistic reasoning in auto-regressive language models. *arXiv preprint arXiv:2408.08590*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. In *Annual Conference on Neural Information Processing Systems*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large

language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20856–20879, Miami, Florida, USA. Association for Computational Linguistics.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using OOD examples. *CoRR*, abs/2305.15269.

S. M. Seals and Valerie L. Shalin. 2023. Evaluating the deductive competence of large language models. In *North American Chapter of the Association for Computational Linguistics*.

Eric Todd, Millicent Li, Arnab Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations*. ICLR.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2155, Miami, Florida, USA. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Annual Meeting of the Association for Computational Linguistics*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118571–118602. Curran Associates, Inc.

Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seongho Joe, Taehee Lee, Youngjune L Gwon, and Sungroh Yoon. 2024. Correcting negative bias in large language models through negative attention score alignment. *arXiv preprint arXiv:2408.00137*.

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*.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

A Extended Related Work

Here we discuss the work related to logical reasoning and internal model interpretability in more detail.

Chain-of-Thought (CoT) and Logical Reasoning. CoT prompting has been widely adopted to improve reasoning in LLMs (Wei et al., 2022; Kojima et al., 2022; Yang et al., 2024; Seals and Shalin, 2023; Wan et al., 2024). These methods prompt models to produce intermediate steps in the reasoning process. However, despite the success in various tasks, they do not offer a measure to quantify the coherence of the reasoning transitions, leaving a gap that our QK-score method aims to fill.

Mechanistic Interpretability. A complementary line of research has focused on understanding transformer models through mechanistic interpretability. Work by Elhage et al. (2021) and Olah et al. (2020) has shown that transformer attention heads can be categorized into different functional phases, including knowledge recalling, latent reasoning, and expression preparation (Zheng et al., 2024). More recent studies have explored specific reasoning tasks: for example, Kim et al. (2024) examined syllogistic reasoning, showing that models learn content-independent reasoning mechanisms transferable across different logical schemes. Other investigations (Lieberum et al., 2023; Yu et al., 2024) have highlighted that certain heads can adversely affect the final decision by introducing latent biases. While ablation-based techniques (Zhang and Nanda, 2024; Todd et al., 2024; Yao et al., 2024) have been used to study these phenomena, they are often resource-intensive or limited to smaller models (Wang et al., 2023).

Benchmark Datasets. Several benchmarks have been designed to evaluate logical reasoning in LLMs. For example, LogicBench (Parmar et al., 2024) and Multi-LogiEval (Patel et al., 2024) test models on tasks such as truth table reasoning, logical entailment, and satisfiability. Datasets specifically tailored for chain-of-thought evaluation, such as PrOntoQA (Saparov and He, 2022),

FOLIO (Han et al., 2022), and FLD (Morishita et al., 2024), demonstrate that unstructured intermediate reasoning steps can enhance performance.

Alternative Evaluation Strategies. Other lines of work have explored neuro-symbolic and data-driven methods to assess reasoning quality. Some approaches reformulate reasoning tasks into structured formats, while others propose direct evaluation metrics based on model outputs.

Summary. While previous work has substantially advanced our understanding of how LLMs reason and how their internal representations operate, there remains a need for efficient, scalable methods to assess logical consistency. Our method, which relies on the natural alignment between specific query and the last statement token key vectors, complements existing techniques and offers an efficient alternative to ablation-based analysis.

B ProntoQA-OOD: additional details

Questions in PrOntoQA-OOD are organized in a following way: given a set of axioms (context) it is required to prove a theorem (statement). For our study we reformulate them into answering if the statement is true or false given the context.

Note that there are six propositional logic rules supported in the existing ProntoQA-OOD benchmark: modus ponens, conjunction introduction/elimination, disjunction introduction/elimination, and proof by contradiction.

We used scripts published by the authors of the dataset to generate the data. We used following command line flags:

- For Modus Ponens only:
  `--ordering random --distractors none --deduction-rule ModusPonens --proofs-only --ontology fictional --min-hops 1 --max-hops 5`
- For composition of deduction rules:
  `--ordering random --deduction-rule Composed --proofs-only --distractors none --ontology fictional --min-hops 1 –max-hops 5`

We also modified the original scripts to make possible variating the number of distractors in prompts, and generated data for $1-$hop questions on Modus Ponens with $1, 5$ distractors respectively.

In our experiments we used training data from *in context examples*. For each setup (deduction rule+number of hops + number of distractors) scripts yielded $4,000$ samples. To exclude possible biases, in each case we select equal number of questions where the statement is given in positive (`X is Y.`) and negative (`X is not Y.`) forms and for each sample we also generate its counterpart where its statement and ground-truth answer are negated (thus balancing the classes). Depending on the setup, this resulted in $2,600 - 7,000$ questions, out of which 600 ($300/300$ with answers "true" and "false" respectively) were taken for calibration and the rest were used as evaluation set (see Table 3). We ensured that no pair axioms+theorem is included in both subsets.

Deduction Rule	Number of hops 1	2	3	4	5
M. P.	6,144	6,204	6,364	6,288	6,176
M. P. + distr.	6,268	-	-	-	-
Comp.	2,752	2,764	2,884	2,956	2,892

Table 3: Size of ProntoQA-OOD evaluation set that was used in different setups

C Pararule-plus: additional details

Here we demonstrate the example from this dataset:

CONTEXT: Harry is strong. Harry is big. Harry is high. Anne is thin. Anne is little. Gary is smart. Gary is quiet. Gary is kind. Fiona is poor. Fiona is rough. Fiona is sad. Strong people are smart. If someone is thin and little then they are short. If someone is poor and rough then they are bad. If someone is smart and quiet then they are nice. All short people are small. All smart people are quiet. All nice people are wealthy. All bad people are dull.
STATEMENT: Harry is quiet.
ANSWER: true

Figure 4: PARARULE-PLUS prompt example of reasoning, depth 2

On PARARULE Plus, three out of five heads, selected on the ProntoQA-OOD dataset, reach accuracy higher than the baseline, in most cases by more than 10%. Interestingly, head `(22, 16)` from DeepSeek-R1 consistently yields an accuracy below 0.35 on PARARULE Plus, suggesting that its QK-score distinguishes correct from incorrect logi-

cal implications but in a reversed manner. Similar effect also occurs in some setups on other heads.

D Extended Multi-Logi-Eval: additional details

Here we demonstrate the example from this dataset and how it was prompted:

Use provided Context to answer the Question.
Print 'yes' or 'no' only.
CONTEXT: If a person uses a fishing rod, they catch fish. Michael uses a fishing rod.
QUESTION: Does Michael catch fish?
ANSWER: yes.

Figure 5: Multi-LogiEval Modus Ponens Example of reasoning depth 1

While the original dataset Multi-Logi-Eval consisted of three types of logic, namely First-Order Logic, Nonmonotonic Logic and Propositional Logic, we extended only a part with first-order logic. We used GPT-4o to generate more samples for each scheme. Table 4 compares the statistics of generated dataset with the statistics of the original (Multi-LogiEval dataset):

	Depth			
Dataset	1	2	3	4
Multi-LogiEval (FOL)	130	105	135	120
Extended Multi-LogiEval	1300	700	900	700

Table 4: Statistics of generated Extended-MultiLogiEval dataset.

We refer to original paper (Patel et al., 2024) for the detailed description of logical schemes, and we did not add any new schemes or changed them in any way on all levels of reasoning. Every scheme of reasoning depth k consists of k atomic rules from reasoning depth 1, see them in Table 5.

E Extended results of in-domain evaluation on ProntoQA-OOD

Tables 6 and 7 provide the in-domain numerical evaluation results on ProntoQA-OOD for BEST HEAD and BASELINE methods calculated for various LLMs. Figure 6 presents the plotted results for larger models (LLaMA3.1-70B-Instruct and Qwen2.5-32B-Instruct).

F Comparison of Head performances on Base and Instruct-tuned models

Fine-tuning a model potentially can alter the roles (and performance) of individual heads. Here we want to investigate this question. Figure 7 shows correlation between accuracies of corresponding heads of base and Instruct-tuned LLaMA3-8B models for three of the setups covered in ProntoQA-OOD dataset (these pictures are typical for other setups as well). From the picture we can see two main patterns: first, a cross-shaped cloud of heads that work only on one 'version' of the model (and yield constant or random predictions on the other), and second, smaller diagonal band of heads that mostly preserve their performance between untuned and tuned versions of the model. Also, we can see that many of the best heads on -base model are also among the best on the -Instruct tuned model and vice versa (top right corner of the scatterplot).

G Head selection on mixture of datasets

In the main text, we performed selection of heads via QK-scores on each dataset individually. Here we would like to briefly study what will happen if in calibration set we use a mixture of two datasets. In this experiment we studied transferability from mixture of ProntoQA-OOD and FOLIO to PARARULE Plus in the setup similar to Section 5.2. FOLIO (Han et al., 2024) is an expert-written dataset for natural language reasoning with first-order logic. We use only its train subset from which we selected 300 positive and 300 negative entries.

Essentially, questions in FOLIO are close to "Composition of logical rules without distractors" in ProntoQA-OOD with varying depth of reasoning For each head we separately calculated the accuracy of QK-score on all 5 reasoning depths in ProntoQA-OOD and on FOLIO, then selected heads that achieved the highest average accuracy; Table 8 presents the results.

From results in the table we can see that usage of the mixture of ProntoQA-OOD and FOLIO yields heads with, on average, better accuracy on the target dataset than any of the individual datasets. Still, there are heads present among top 5 selected heads in any case (heads $(22, 26)$ and $(18, 1)$). This additionally supports the claim that our method selects

Rule	First-order Logic
MP	$(\forall x(p(x) \rightarrow q(x)) \land p(a)) \vdash q(a)$
MT	$(\forall x(p(x) \rightarrow q(x)) \land \neg q(a)) \vdash \neg p(a)$
HS	$(\forall x((p(x) \rightarrow q(x)) \land (q(x) \rightarrow r(x))) \vdash (p(a) \rightarrow r(a))$
DS	$(\forall x(p(x) \lor q(x)) \land \neg p(a)) \vdash q(a)$
CD	$(\forall x((p(x) \rightarrow q(x)) \land (r(x) \rightarrow s(x))) \land (p(a) \lor r(a))) \vdash (q(a) \lor s(a))$
DD	$(\forall x((p(x) \rightarrow q(x)) \land (r(x) \rightarrow s(x))) \land (\neg q(a) \lor \neg s(a))) \vdash (\neg p(a) \lor \neg r(a))$
BD	$(\forall x((p(x) \rightarrow q(x)) \land (r(x) \rightarrow s(x))) \land (p(a) \lor \neg s(a))) \vdash (q(a) \lor \neg r(a))$
CT	$\forall x(p(x) \lor q(x)) \dashv\vdash \forall x(q(x) \lor p(x))$
DMT	$\neg\forall x(p(x) \land q(x)) \dashv\vdash \exists x(\neg p(x) \lor \neg q(x))$
CO	$\forall x((p(x) \rightarrow q(x)) \land (p(x) \rightarrow r(x))) \vdash \forall x(p(x) \rightarrow (q(x) \land r(x)))$
IM	$\forall x(p(x) \rightarrow (q(x) \rightarrow r(x))) \dashv\vdash \forall x((p(x) \land q(x)) \rightarrow r(x))$
EG	$p(a) \vdash \exists x(p(x))$
UI	$\forall x(p(x)) \vdash p(a)$

Table 5: Inference rules that establish the relationship between premises and their corresponding conclusions in Extended Multi-LogiEval. The schemes are MP: Modus Ponens, MT: Modus Tollens, HS: Hypothetical Syllogism, DS: Disjunctive Syllogism, CD: Constructive Dilemma, DD: Destructive Dilemma, BD: Bidirectional Dilemma, CT: Commutation, DMT: De Morgan's Theorem, CO: Composition, IM: Importation, EG: Existential Generalization, UI: Universal Instantiation

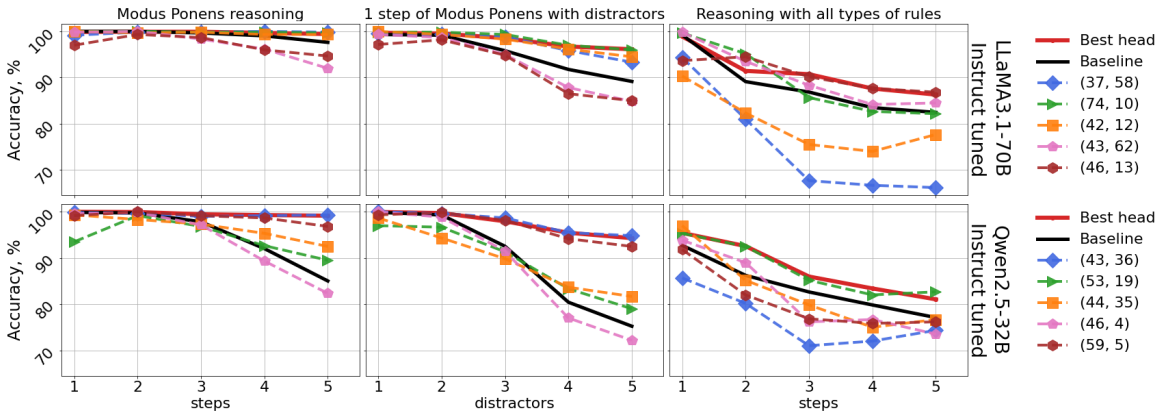


Figure 6: In-domain performance of large models on ProntoQA-OOD. Best heads are selected per case using calibration data. The notation (37, 58) means the 58th head from 37th layer and so on.

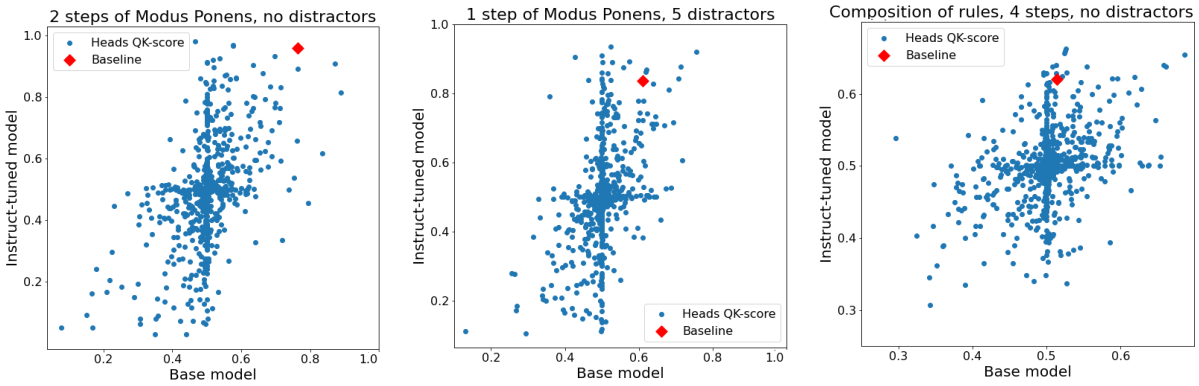


Figure 7: Correlation of baseline and head-wise QK-score accuracies on calibration subset of ProntoQA-OOD dataset between LLaMA3-8B-base and -Instruct tuned models.

heads responsible for logic processing.

H Full results of ablation study

In Section 5.3 we described the experiment with removal of high QK-scroring heads on ProntoQA-OOD dataset. We evaluated performance of a model when its K best attention heads are pruned (i.e., their outputs are zeroed out). We compared it with the case when K random heads are pruned.

Here, Table 9 shows the full results of this comparison on all setups from ProntoQA-OOD for model LLaMA-3.1-8B and $K = 10, 20$; averaging was done over 7 restarts.

I Distribution of "logic" heads in LLMs

As shown in Figures 8 and 9, strongly pronounced "logic" heads typically emerge after the first third of layers in the LLMs we analyzed. We hypothesize that they don't appear earlier because logical reasoning requires high-level abstractions, which the initial layers cannot provide. Interestingly, logical heads in Qwen models concentrate primarily in the final half or final third of the layers, while in LLaMA models, they populate the latter two-thirds of the layers. It is also interesting that such heads alternate with "inverse-logic" heads that consistently predict the wrong answer, and these "inverse-logic" heads generally populate the same layers.

J Licenses for the artifacts used

Datasets:

- ProntoQA-OOD – Apache-2.0 license
- PARARULE Plus – MIT license
- MultiLogiEval – MIT license

		Modus ponens only					Composition of rules				
Reasoning depth:		**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
LLaMA2	QK	**0.8889**	**0.8427**	**0.8474**	**0.8524**	**0.8444**	**0.7474**	**0.7847**	**0.7237**	**0.7013**	**0.7004**
7B Chat	Ba.	0.6772	0.6349	0.6506	0.6197	0.6072	0.5678	0.5414	0.5427	0.5232	0.5346
LLaMA2	QK	**0.9778**	**0.9637**	**0.9283**	**0.9187**	**0.9107**	**0.7584**	**0.7134**	**0.7044**	**0.6777**	**0.6964**
13B Chat	Ba.	0.8944	0.7784	0.7527	0.7385	0.7414	0.6556	0.4944	0.4932	0.4751	0.4786
LLaMA3	QK	**0.9090**	**0.8899**	**0.8465**	**0.8054**	**0.8217**	**0.7219**	**0.7073**	**0.6951**	**0.6863**	**0.6767**
8B Base	Ba.	0.8632	0.7636	0.6613	0.6199	0.6097	0.6023	0.5791	0.5517	0.5145	0.5057
LLaMA3	QK	0.9831	**0.9755**	**0.9483**	**0.9358**	**0.9324**	0.8197	**0.6756**	**0.6873**	**0.6620**	**0.6866**
8B Instruct	Ba.	**0.9840**	0.9591	0.9133	0.8638	0.8320	**0.8379**	0.6727	0.6603	0.6206	0.5945
LLaMA3.1	QK	**0.8909**	0.7947	0.7170	0.6652	0.6851	**0.6553**	**0.6449**	**0.6288**	**0.6419**	**0.6988**
8B Base	Ba.	0.5944	**0.6490**	**0.6791**	**0.6906**	**0.7208**	0.5982	0.5930	0.5954	0.5723	0.5770
LLaMA3.1	QK	**0.9881**	0.9577	0.9192	0.8799	0.8676	**0.8084**	**0.7184**	**0.7188**	**0.7239**	**0.7052**
8B Instruct	Ba.	0.9843	**0.9785**	**0.9473**	**0.9035**	**0.8775**	0.6848	0.6206	0.5847	0.5888	0.5731
LLaMA3	QK	**1.0000**	**0.9999**	0.9963	**0.9981**	**0.9960**	0.**9890**	**0.9214**	**0.8799**	**0.8617**	**0.8353**
70B Instruct	Ba.	0.9993	**0.9999**	**0.9981**	0.9868	0.9739	0.9805	0.8767	0.8324	0.8171	0.7948
LLaMA3.1	QK	0.9987	0.9993	**0.9993**	**0.9968**	**0.9942**	0.9890	**0.9144**	**0.9073**	**0.8762**	**0.8631**
70B Instruct	Ba.	**1.0000**	**0.9999**	0.9967	0.9904	0.9762	**0.9922**	0.8912	0.8685	0.8348	0.8246
Qwen-2.5	QK	**0.9627**	**0.9149**	**0.8995**	**0.8487**	**0.8379**	**0.6981**	**0.6201**	**0.6275**	**0.6135**	**0.6209**
1.5B Instruct	Ba.	0.9410	0.8197	0.7694	0.7273	0.6946	0.6648	0.6036	0.5617	0.5299	0.5207
Qwen-2.5	QK	**0.9988**	**0.9966**	**0.9869**	**0.9791**	**0.9683**	**0.9724**	**0.8813**	**0.8313**	**0.8332**	**0.8397**
14B Instruct	Ba.	0.9660	0.9656	0.9388	0.9146	0.9060	0.9050	0.8374	0.7720	0.7323	0.7451
Qwen-2.5	QK	**1.0000**	**0.9996**	**0.9947**	**0.9929**	**0.9916**	**0.9543**	**0.9257**	**0.8598**	**0.8340**	**0.8104**
32B Instruct	Ba.	0.9988	0.9975	0.9783	0.9203	0.8502	0.9274	0.8623	0.8264	0.7982	0.7710
DeepSeek-R1	QK	**0.9149**	**0.8039**	**0.8995**	**0.8487**	**0.8379**	**0.6718**	**0.6393**	**0.5976**	**0.5825**	**0.5894**
-Distill-Qwen-1.5B	Ba.	0.5361	0.5103	0.5053	0.5011	0.5020	0.4873	0.4980	0.4959	0.4971	0.4985
DeepSeek-R1	QK	0.9816	**0.9519**	**0.9141**	**0.8701**	**0.8418**	**0.8233**	**0.7577**	**0.7568**	**0.7090**	**0.7071**
-Distill-Qwen-7B	Ba.	**0.9873**	0.9336	0.8638	0.6879	0.6295	0.6512	0.5723	0.5408	0.5304	0.5413

Table 6: Comparison of models with different reasoning depths on ProntoQA-OOD. "Ba." stands for the Baseline. Best results are highlighted in bold.

		Distractors added				
		1	**2**	**3**	**4**	**5**
LLaMA2 7B Chat	QK	**0.8725**	**0.8654**	**0.8595**	**0.8487**	**0.8498**
	Baseline	0.6577	0.6583	0.6422	0.6280	0.6248
LLaMA2 13B Chat	QK	**0.9627**	**0.8972**	**0.8981**	**0.8920**	**0.8552**
	Baseline	0.7959	0.7348	0.7160	0.7005	0.6677
LLaMA3 8B Base	QK	**0.8665**	**0.8028**	**0.7902**	**0.7753**	**0.7551**
	Baseline	0.7398	0.6612	0.6453	0.6107	0.6114
LLaMA3 8B Instruct	QK	**0.9611**	**0.9515**	**0.9371**	**0.9266**	**0.9220**
	Baseline	0.9474	0.8993	0.8736	0.8596	0.8361
LLaMA3.1 8B Base	QK	**0.7727**	**0.6958**	0.6763	**0.6503**	0.6851
	Baseline	0.6708	0.6906	**0.6792**	**0.6909**	**0.7002**
LLaMA3.1 8B Instruct	QK	0.9527	0.8577	**0.8170**	**0.7895**	**0.7699**
	Baseline	**0.9629**	**0.8691**	0.7830	0.7212	0.6672
LLaMA3 70B Instruct	QK	**0.9967**	**0.9947**	**0.9897**	**0.9849**	**0.9845**
	Baseline	0.9961	0.9882	0.9463	0.9106	0.8926
LLaMA3.1 70B Instruct	QK	**0.9971**	**0.9954**	**0.9853**	**0.9672**	**0.9613**
	Baseline	0.9933	0.9918	0.9582	0.9176	0.8916
Qwen-2.5 1.5B Instruct	QK	**0.9627**	**0.9149**	**0.8995**	**0.8487**	**0.8379**
	Baseline	0.9410	0.8197	0.7694	0.7273	0.6946
Qwen-2.5 14B Instruct	QK	**0.9955**	**0.9823**	**0.9556**	**0.9362**	**0.9320**
	Baseline	0.9687	0.9528	0.8896	0.8336	0.8130
Qwen-2.5 32B Instruct	QK	**0.9997**	**0.9973**	**0.9794**	**0.9546**	**0.9429**
	Baseline	0.9990	0.9931	0.9253	0.8042	0.7522
Deepseek R1-Distill-Qwen-1.5B	QK	**0.8290**	**0.7650**	**0.7485**	**0.7574**	**0.7501**
	Baseline	0.5361	0.5096	0.5098	0.5016	0.5012
Deepseek R1-Distill-Qwen-7B	QK	0.9448	0.8972	0.8725	**0.8625**	**0.8540**
	Baseline	**0.9597**	**0.9397**	**0.8736**	0.8191	0.7840

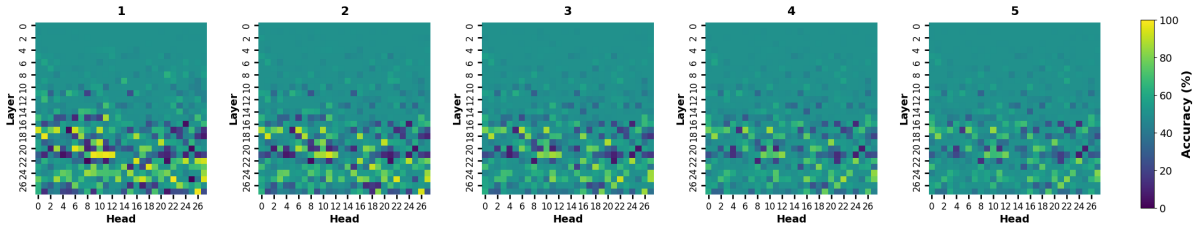
Table 7: Effect of distractors added to the prompt on ProntoQA-OOD. Only Modus Ponens inference.

		PARARULE Plus Reasoning depth			
	Head	**2**	**3**	**4**	**5**
ProntoQA	(22, 16)	0.3206	0.3422	0.2763	0.2388
-OOD only	(22, 26)	**0.6736**	**0.7418**	0.5495	0.6145
	(18, 16)	0.5155	0.5340	0.5526	0.4452
	(21, 12)	0.4982	0.5203	0.4452	0.4740
	(18, 1)	0.5292	0.5340	0.5155	0.4933
FOLIO	(18, 1)	0.5292	0.5340	0.5155	0.4933
only	(17, 0)	0.5771	0.5703	0.6552	0.6732
	(20, 6)	0.5691	0.5632	0.5469	**0.6405**
	(21, 9)	0.6058	0.6158	0.5990	0.6225
	(22, 26)	**0.6736**	**0.7418**	0.5495	0.6145
Mixture	(22, 16)	0.3206	0.3422	0.2763	0.2388
of FOLIO	(22, 26)	**0.6736**	**0.7418**	0.5495	0.6145
and	(18, 1)	0.5292	0.5340	0.5155	0.4933
ProntoQA	(21, 9)	0.6058	0.6158	0.5990	0.6225
-OOD	(24, 11)	0.3370	0.4014	0.4000	0.4610
Baseline	-	0.4969	0.5212	0.4523	0.4717

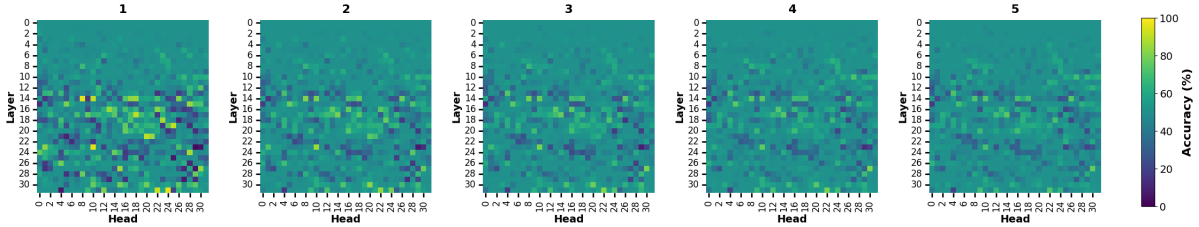
Table 8: Performance of QK-score on top 5 heads from DeepSeek R1-Distill-Qwen-7B, selected on ProntoQA-OOD and FOLIO in cross-domain evaluation on PARARULE Plus dataset. Best results are highlighted in **bold**. Those results that are better than baseline are underlined

			No	# of heads pruned			
Ruleset	Depth	Distractors	pruning	10 best	10 random	20 best	20 random
Modus	1	0	86.32	81.44	84.66 ± 2.14	82.08	82.41 ± 3.83
Ponens	2	0	76.36	76.17	78.04 ± 3.56	73.89	75.99 ± 3.47
	3	0	66.13	70.60	67.70 ± 4.67	58.04	69.97 ± 4.62
	4	0	61.99	61.10	62.29 ± 3.01	55.84	62.14 ± 6.38
	5	0	60.97	61.07	62.64 ± 3.27	58.74	61.41 ± 5.06
	1	1	73.98	73.37	83.80 ± 2.00	79.55	81.34 ± 2.74
	1	2	66.12	63.72	68.17 ± 2.69	71.37	69.28 ± 4.38
	1	3	64.53	63.25	65.46 ± 3.17	66.57	62.34 ± 6.25
	1	4	61.07	58.63	62.23 ± 3.43	65.15	60.83 ± 4.91
	1	5	61.14	64.54	60.37 ± 2.39	64.74	66.49 ± 5.21
All rules	1	0	60.23	63.77	59.94 ± 1.25	62.31	57.95 ± 5.20
	2	0	57.91	57.55	58.42 ± 1.18	57.22	57.03 ± 1.45
	3	0	54.13	57.14	55.83 ± 0.83	54.04	55.07 ± 1.25
	4	0	50.94	54.29	52.46 ± 0.75	49.77	54.00 ± 1.01
	5	0	50.97	50.56	50.65 ± 0.93	50.18	51.78 ± 0.74

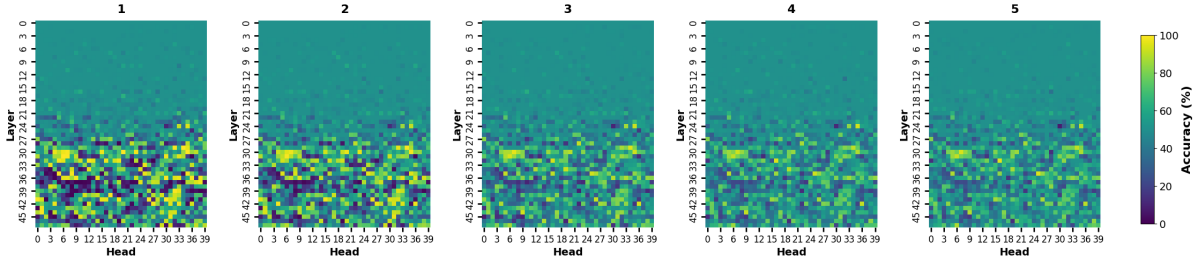
Table 9: Performance of LLaMA-3-8B (base) model on various setups from PRONTOQA-OOD after pruning a number of attention heads; averaging done over 7 restarts.



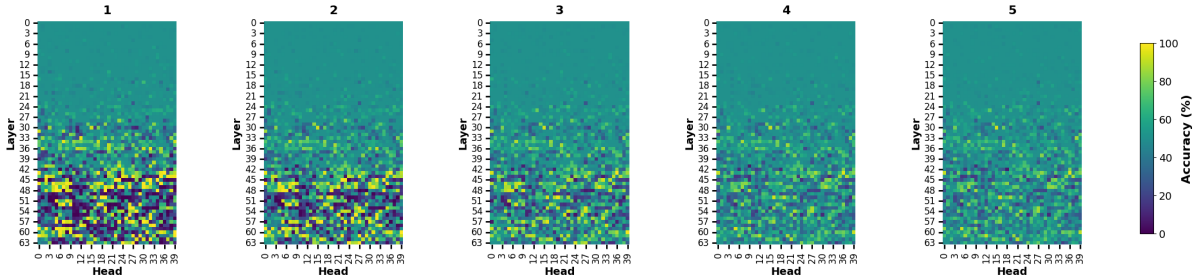
(a) DeepSeek-Qwen 7B



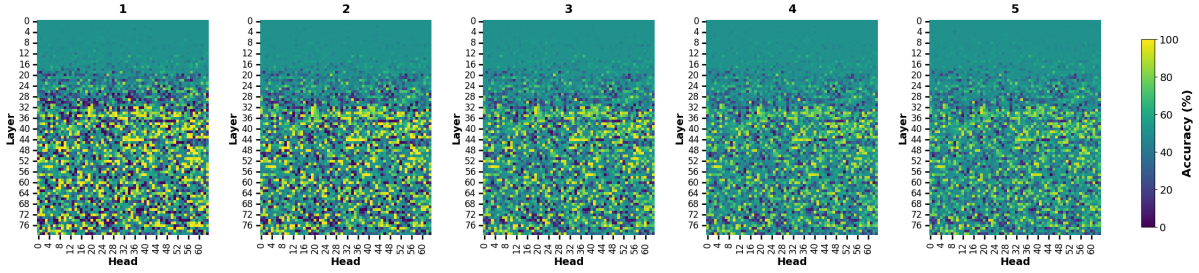
(b) Llama-3.1-8B



(c) Qwen2.5-14B-Instruct

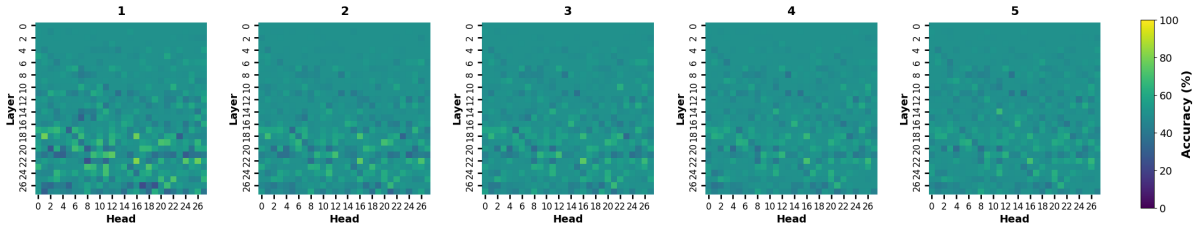


(d) Qwen2.5-32B-Instruct

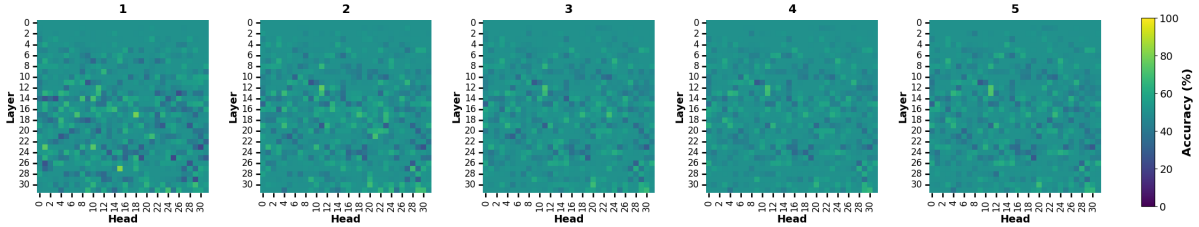


(e) Llama-3.1-70B-Instruct

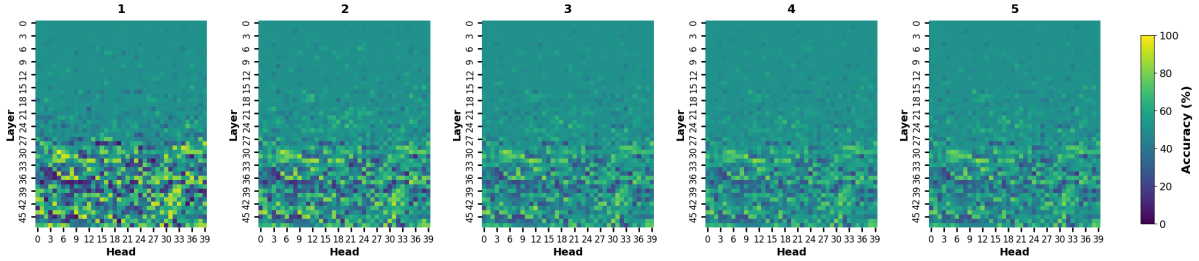
Figure 8: Accuracy of the QK-score across attention heads in models of varying sizes on Modus Ponens tasks with 1–5 distractors.



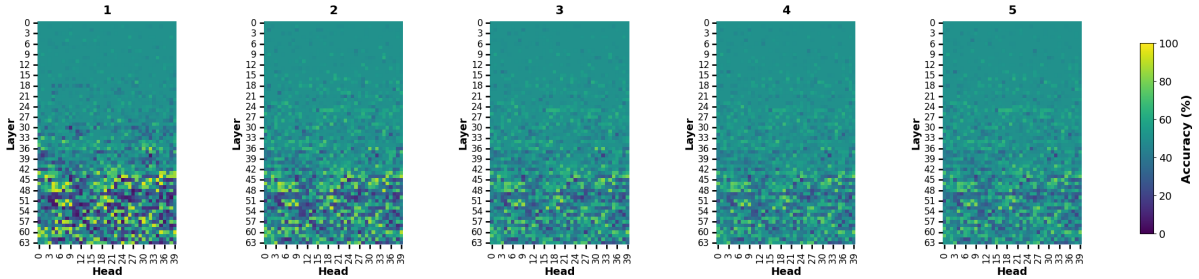
(a) DeepSeek-Qwen 7B



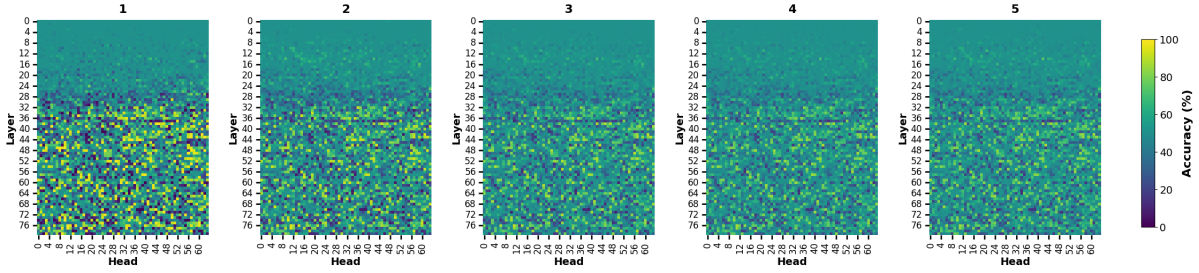
(b) Llama-3.1-8B



(c) Qwen2.5-14B-Instruct



(d) Qwen2.5-32B-Instruct



(e) Meta-Llama-3.1-70B-Instruct

Figure 9: Accuracy of the QK-score across attention heads in models of varying sizes on tasks composed from several logical rules with 1 - 5 reasoning depth.