

# LLM Parameters for Math Across Languages: Shared or Separate?


Behzad Shomali<sup>1,2†</sup> Luisa Victor<sup>2†</sup> Tim Selbach<sup>2†</sup> Ali Hamza Bashir<sup>1,3</sup>  
David Berghaus<sup>1,3</sup> Joachim Koehler<sup>1,3</sup> Mehdi Ali<sup>1,3</sup> Markus Frey<sup>1,2,3</sup>

<sup>1</sup>Lamarr Institute <sup>2</sup>University of Bonn <sup>3</sup>Fraunhofer IAIS <sup>†</sup>Equal contribution

## Abstract

Large language models (LLMs) exhibit substantial cross-lingual variation in mathematical reasoning performance, but it remains unclear whether these differences reflect language-specific parameters or a shared mechanism that manifests differently by language. We present a cross-lingual mechanistic analysis of mathematical reasoning in LLMs, enabling us to localize and compare model parameters that support mathematical reasoning across languages. We find that the extracted math-associated parameters exhibit partial cross-lingual overlap, with the strongest overlap concentrated in intermediate model layers. We further observe that English consistently produces the largest set of math-relevant parameters, whereas lower-resource languages reveal smaller sets of relevant parameters. These results suggest that math-related behavior in multilingual LLMs is neither fully language-invariant nor fully language-specific, but instead exhibits partial cross-lingual parameter overlap with systematic language-dependent differences.

 **Code:** <https://github.com/luisavictor/math-across-languages>

 **Translated datasets:** <https://huggingface.co/math-across-languages>

## 1 Introduction

Large language models (LLMs) have achieved remarkable success across a range of tasks, yet their performance on multilingual mathematical reasoning remains uneven (Chen et al., 2024). Understanding how these models process mathematical tasks across languages is critical for building robust multilingual systems.

Existing research on LLM capabilities can be broadly categorized into two streams. First, mechanistic studies localize task-specific skills to subsets of model parameters. Recent methods lever-

age forward-pass statistics (Sun et al., 2023). Notably, Christ et al. (2025) introduced *MathNeurosurgery*, isolating parameters critical to mathematical reasoning in LLMs. Second, multilingual reasoning studies primarily evaluate performance or representational properties. Datasets and strategies like MathOctopus (Chen et al., 2024) and the UST dataset (Ko et al., 2025) target cross-lingual generalization, while Liu et al. (2026) suggests that English-centric latent pathways may dominate multilingual reasoning. However, these approaches rarely identify whether the same parameters support mathematical reasoning across languages.

Despite these advances, several challenges remain. Mechanistic analyses are mostly limited to English, leaving open questions about cross-lingual parameter sharing. Similarly, performance-focused multilingual studies do not reveal the underlying computational substrates of reasoning. These gaps hinder our understanding of whether mathematical reasoning relies on a universal core or language-specific circuits. To address this gap, we investigate the localization and overlap of math-specific parameters in LLMs. Our contributions are threefold:

- **Mechanistic localization of multilingual math circuits:** We develop a framework to identify and compare model parameters supporting mathematical reasoning across languages.
- **Language-dependent variation in parameter allocation:** We show that English contains the largest set of math-specific parameters, while lower-resource languages have fewer.
- **Functional validation via weight intervention:** We demonstrate that mathematical reasoning emerges as a *collective* property of these localized parameters.

## 2 Related Work

**Mechanistic analysis of LLMs.** Several studies have explored localizing task-specific capabilities

Correspondence to: behzad.shomali@uni-bonn.de

within LLM parameters. Gradient-based attribution techniques identify parameters critical to specific tasks (Panigrahi et al., 2023), while forward-pass statistics offer a computationally efficient alternative (Sun et al., 2023). Christ et al. (2025) proposed *MathNeurosurgery*, which isolates math-specific parameters by comparing weight–activation products on math and non-math datasets. However, their analysis primarily focuses on English and does not examine cross-lingual parameter behavior.

**Multilingual LLMs.** Research on multilingual LLMs has explored dataset creation and evaluation. MathOctopus (Chen et al., 2024) and the UST dataset (Ko et al., 2025) target cross-lingual generalization, while other studies suggest that multilingual models often rely on English-centric latent pathways (Liu et al., 2026). These analyses typically evaluate performance or representations, leaving the underlying parameter-level mechanisms unexplored.

**Cross-lingual mechanistic studies.** Some recent work has investigated neuron activation patterns across translated inputs to identify shared functional units (Wang et al., 2024). Our study differs by using a parameter-importance framework, analyzing both attention and Feed-Forward Network (FFN) weights, to localize math-specific circuits and quantify their overlap across languages.

### 3 Methodology

We employ the *MathNeurosurgery* framework (Christ et al., 2025) to identify math-specific parameters and subsequently analyze their overlap, both at the level of the entire model and across individual layers, across multiple languages.

#### 3.1 Parameter Identification

Following (Christ et al., 2025), we first compute the task-specific  $S_{ij}$  for weight  $j$  in neuron  $i$ :

$$S_{ij} = \sum_{k=1}^N |W_{ij}| \cdot \left\| X_j^k \right\|_2 \quad \text{for } X \in \mathcal{D} \quad (1)$$

where  $W_{ij}$  represents the weight,  $|\cdot|$  denotes the absolute value operator, and  $\|X_j^k\|_2$  is the L2 norm of the  $j$ -th feature aggregated across input tokens to normalize the input  $X$  from dataset  $\mathcal{D}$  for sample  $k$ . Using this formulation, we compute scores for each parameter over math ( $\mathcal{D}_{\text{math}}$ ) and non-math datasets ( $\mathcal{D}_{\text{non-math}}$ ). Then we separately identify the top- $k$  proportion of parameters with the highest score for each task in each layer. Finally, we identify the

math-specific parameters by taking the subset of parameters that are crucial for the math task but not for the unrelated task.

#### 3.2 Similarity Metric

To measure similarity between two sets of extracted parameters, we first separately run the previous step for each language, and then measure their pairwise overlaps using the Jaccard coefficient. Let  $T_1$  and  $T_2$  denote the sets of important parameters representing task-specific parameters for the same task in two languages. The overlap between the two parameter sets is computed as:

$$J(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \quad (2)$$

## 4 Results

We begin with the experimental setup, followed by an analysis of the results.

#### 4.1 Experimental Setup

We evaluate Llama 3.2 1B IT (Llama 1B) (Meta AI, 2024b), Qwen3 4B IT-2507 (Qwen3 4B) (Qwen Team, 2025), and Llama 3.1 8B IT (Llama 8B) (Meta AI, 2024a) across English, German, French, and Hindi on GSM8K (Cobbe et al., 2021) for math reasoning ( $\mathcal{D}_{\text{math}}$ ), MMLU (Hendrycks et al., 2020) and RACE (Lai et al., 2017) as general language understanding datasets ( $\mathcal{D}_{\text{non-math}}$ ). We mainly focus on Llama 1B in the main text, while results for larger models are in the Appendices.

#### 4.2 Does math performance correlate with the model’s capacity for each language?

Figure 1 (left) illustrates that, for each top- $k$  value, English consistently exhibits the highest number of math-specific parameters, followed by German and French, and then Hindi. This pattern aligns with the observed reasoning performance across languages, as shown in Table 1. More specifically, the trend in the number of math-specific parameters aligns with the relative performance ranking across languages. For example, in Figure 1, at top- $k = 0.01$ , English has the highest parameter count, French and German have approximately equal counts, and Hindi has the lowest. Moreover, the gap between Hindi and French is smaller than that between Hindi and English. This pattern becomes even more pronounced with larger models (Figures F.4, F.5).

Model	English		German		Hindi		French	
	GSM8K	RACE	GSM8K	RACE	GSM8K	RACE	GSM8K	RACE
Llama 1B	0.340	0.379	0.235	0.321	0.145	0.299	0.185	0.348
Qwen3 4B	0.735	0.414	0.685	0.356	0.385	0.353	0.690	0.403
Llama 8B	0.765	0.448	0.585	0.396	0.415	0.388	0.580	0.413

Table 1: Pretraining accuracy (%) on GSM8K and RACE across English, German, Hindi, and French.

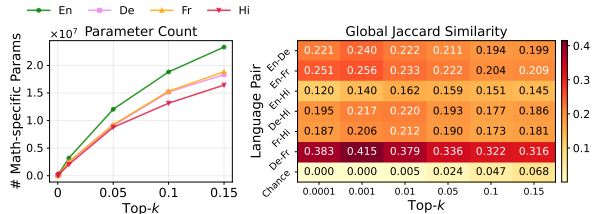


Figure 1: Math-specific parameters in Llama 1B. **Left:** Number of isolated parameters per language. **Right:** Jaccard similarity between language-specific parameter sets.

**Finding:** The distribution of math-specific parameters across languages tends to align with their relative reasoning performance in LLMs.

### 4.3 Is there a shared cross-lingual core for mathematical reasoning?

Figure 1 highlights the number of math-associated parameters in English (En), German (De), Hindi (Hi) and French (Fr), as well as the Jaccard similarity across language pairs for Llama 1B. Across all language pairs, the Jaccard similarity primarily peaks in the intermediate layers and decreases in deeper layers. The same behavior is also observed in larger models (Figures F.1-F.3). This pattern is consistent with prior findings that earlier layers capture more surface-level linguistic features, while later layers can become more specialized and language-dependent (Tang et al., 2024).

Figures 1 and 2 demonstrate that English consistently possesses the highest number of math-associated parameters, both in the entire model and in nearly every layer. This implies that the model possesses a more distributed and resilient internal circuit for mathematical computations in English.

Notably, the Jaccard similarity remains substantially above the random baseline across all pairings, except in the initial and final few layers. This suggests the presence of a partially shared parameter substrate associated with mathematical problem solving, especially in intermediate layers. Lower similarity in the initial and final layers, especially in larger models (Figures F.2, F.3), likely reflects two distinct phenomena. In early layers, low overlap is consistent with the surface-level processing

discussed above. In later layers, it instead points to increased language-specific specialization: with larger parameter budgets, models can mitigate the *curse of multilinguality* (Conneau et al., 2020) by allocating dedicated sub-networks for language-specific reasoning, rather than relying on shared parameters, consistent with other findings on layer specialization in LLMs (Frey et al., 2026).

While Figures 1 and 3 confirm that lexically similar, high-resource languages like French and German exhibit the highest overlap, the lower similarity observed in pairs involving Hindi, particularly En-Hi, suggests a structural barrier imposed by script and tokenization. Because English and Hindi share neither an alphabet nor substantial lexical overlap, the model appears to rely on more disjoint parameter subsets to map divergent input encodings to a shared internal representation. This pattern is also observed in Llama 8B, independent of model scale (Figure F.3). In contrast, the effect is less pronounced in Qwen3 4B (Figure F.2), potentially reflecting differences in multilingual pretraining coverage, vocabulary design, or tokenization. We emphasize that this interpretation is speculative, as our experiments do not isolate these factors.

Surprisingly, the De-Hi and Fr-Hi overlaps in many layers are notably higher than that of En-Hi. We consider two possible explanations. First, French and Hindi as well as German and Hindi have a comparable number of parameters compared to English and Hindi, which explains the lower Jaccard scores for En-Hi. The larger denominator in the similarity computation using Equation 2 is due to English’s larger set of math-associated parameters, which increases the size of the union and reduces the Jaccard similarity. Second, an alternative interpretation is the existence of a shared non-English subspace: a multilingual reasoning manifold that the model exploits when operating outside its primary English-centric pathways. Notably, the pattern in which En-Hi similarity is lower than De-Hi, Fr-Hi, and De-Fr persists across almost all layers and scales to larger models (Figures F.2, F.3). This consistent, surprising observation further motivates the hypothesis that non-English languages may rely on partially overlapping parameter subsets distinct from those used for English.

**Finding:** There exists a partially shared parameter substrate for math problem-solving, which is most prominent in the intermediate layers.

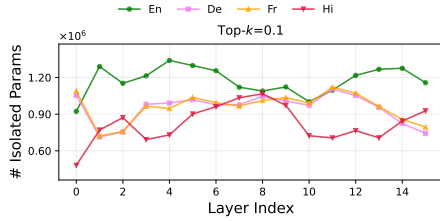


Figure 2: Layer-wise distribution of math-reasoning parameters per language for Llama 1B.

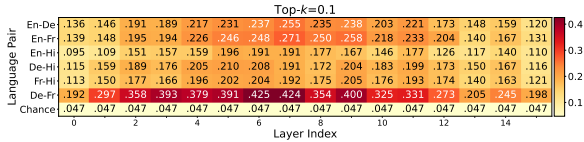


Figure 3: Layer-wise Jaccard similarity of math-specific parameters across language pairs for Llama 1B.

#### 4.4 How do weights scaling & pruning change the model’s behavior?

To validate and investigate the effectiveness of isolated parameters, following Christ et al. (2025), we apply two interventions: *pruning* and *scaling* (full results in Appendix D). In pruning, weights are set to zero, which ideally impairs mathematical performance more than non-mathematical performance. In scaling, weights are multiplied by a positive factor (1.1 for Llama 1B, 1.01 otherwise). While scaling may improve performance, the effect is less predictable than pruning, as outcomes depend sensitively on the factor choice. In both settings, the intended trend is observed: non-mathematical performance remains largely unchanged, whereas mathematical reasoning is either significantly reduced by pruning or enhanced by scaling.

Analyzing examples where scaling guides the model to the correct answer, we find that the reasoning logic remains largely unchanged, even though the reasoning may not be entirely accurate and may include some missing steps (Figure E.2). However, for all four languages, the gains mainly come from correcting errors in arithmetic computations, as shown in Figures E.1-E.4. This observation explains the trivial improvements observed in larger models, which already have strong arithmetic capabilities and are therefore less responsive to small parameter perturbations.

Tables D.9 and D.10 demonstrate an emerging pattern when pruning Qwen3 4B model with  $\text{top-}k=1\text{e-}5$ . Instead of dropping performance, performance on the English GSM8K task improved remarkably. Manual inspection revealed that suppressing such parameters mainly improved output formatting and better in-context learning (Fig-

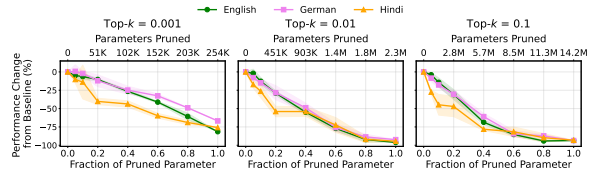


Figure 4: Effect of pruning isolated math-specific parameters on performance across languages for Llama 1B.

ure E.9). More details in Appendix D.

**Finding:** *Scaling math-specific parameters primarily corrects arithmetic execution rather than reasoning logic. In some cases, pruning may improve GSM8K accuracy by changing output format or few-shot imitation behavior, rather than by improving reasoning itself.*

#### 4.5 Math-specific parameters effect: collective or individual?

We investigate whether isolated math-specific parameters exhibit a collective effect, i.e. their impact arises from their aggregation, or an individual effect, where performance changes are driven by a small number of critical parameters. To this end, for the selected top- $k$  parameters, we sample random subsets at several intermediate proportions. For each proportion, we repeat the sampling process three times and prune the selected parameters.

Figure 4 depicts the results for Llama 1B across three languages. The observed trend is approximately linear, suggesting that the isolated parameters effect is predominantly *collective* rather than driven by individual parameters. Specifically, the linear relationship indicates that performance degradation scales with the proportion of important parameters removed. Moreover, the low standard deviation supports this conclusion, indicating that the results are not sensitive to the specific subset selected; instead, performance degradation is primarily determined by the proportion of parameters ablated. We observe similar trends for larger models (Figures D.1, D.2).

**Finding:** *Identified math-specific parameters have a collective influence on the outcome.*

## 5 Conclusion

Our mechanistic analysis suggests that math-associated parameters in multilingual LLMs exhibit partial cross-lingual overlap, with the highest overlap concentrated in intermediate layers. At the same time, deeper layers show more language-specific patterns. We find that English yields the largest

set of extracted math-associated parameters, consistent with its stronger performance in our experiments. Lower-resource languages and languages with different scripts, such as Hindi, show less overlap with English, while still sharing some overlap with other non-English languages. These results suggest a distinct pattern of overlap among non-English math-associated parameters. Finally, intervention experiments indicate that the effect of the identified parameters is distributed across the selected set, i.e. exhibiting a collective effect, rather than dominated by a small number of individual parameters.

## 6 Acknowledgment

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

## Limitations

Our work has several limitations. First, despite manual verification, the translated datasets may still lack full fluency and could benefit from stronger translation models. Second, we focus exclusively on mathematical reasoning, leaving open whether similar shared-core and language-dependent patterns extend to other reasoning forms. Third, our analysis considers a limited set of languages and model families, which may restrict the generality of our conclusions across scripts and tokenization schemes. We leave these directions for future work.

## References

- Ameen Ali, Shahar Katz, Lior Wolf, and Ivan Titov. 2025. Detecting and pruning prominent but detrimental neurons in large language models. *arXiv preprint arXiv:2507.09185*.
- Ameen Ali, Lior Wolf, and Ivan Titov. 2024. Mitigating copy bias in in-context learning through neuron pruning. *arXiv preprint arXiv:2410.01288*.
- Robert SP Beekes. 2011. *Comparative Indo-European Linguistics: An Introduction*, 2nd edition. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016.
- Bryan R Christ, Zachary Gottesman, Jonathan Kropko, and Thomas Hartvigsen. 2025. Math neurosurgery: Isolating language models’ math reasoning abilities using only forward passes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24803–24840.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Fangan Dong, Zuming Yan, Xuri Ge, Zhiwei Xu, Mengqi Zhang, Xuanang Chen, Ben He, Xin Xin, Zhumin Chen, and Ying Zhou. 2026. Identifying and transferring reasoning-critical neurons: Improving llm inference reliability via activation steering. *arXiv preprint arXiv:2601.19847*.
- Markus Frey, Behzad Shomali, Ali Hamza Bashir, David Berghaus, Joachim Koehler, and Mehdi Ali. 2026. Adaptive loops and memory in transformers: Think harder or know more? *arXiv preprint arXiv:2603.08391*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*.

- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.
- Yihong Liu, Raoyuan Zhao, Hinrich Schütze, and Michael A Hedderich. 2026. Large reasoning models are (not yet) multilingual latent reasoners. *arXiv preprint arXiv:2601.02996*.
- Meta AI. 2024a. [Introducing llama 3.1: Our most capable models to date](#). Accessed: 2026-01-26.
- Meta AI. 2024b. [Llama 3.2: Revolutionizing edge ai and vision with open, customizable models](#). Accessed: 2026-01-26.
- Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning*, pages 27011–27033. PMLR.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.
- Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

## A Experimental Setup

In this section, we describe the languages, the dataset curation, and the translation process employed throughout this study.

**Languages** To evaluate cross-lingual transferability of mathematical reasoning, we consider English (En), German (De), French (Fr), and Hindi (Hi). This selection enables comparisons across both closely related and more distantly related languages within the Indo-European family. English and German belong to the Germanic branch, while French is a Romance language, and Hindi belongs to the Indo-Aryan branch, following the classification outlined by [Beekes \(2011\)](#).

**Dataset curation:** Since MMLU includes a substantial number of math-oriented questions, unlike [Christ et al. \(2025\)](#), who include *all* available subjects, we apply category-based filtering and remove subject areas that are strongly quantitative, as these can confound our parameter extraction by making parameters associated with mathematical reasoning appear as general language parameters. Concretely, we exclude **mathematics-related categories** (abstract\_algebra, elementary\_mathematics, college\_mathematics, high\_school\_mathematics, high\_school\_statistics), **physics categories** (conceptual\_physics, college\_physics, high\_school\_physics, astronomy), **chemistry categories** (college\_chemistry, high\_school\_chemistry), **biology** (high\_school\_biology), **computer science** (college\_computer\_science, high\_school\_computer\_science), and **quantitatively oriented fields** (econometrics, professional\_accounting).

**Translations:** We translate all datasets into German, Hindi, and French using the pretrained multilingual machine translation model facebook/nllb-200-3.3B ([Costa-Jussà et al., 2022](#)).

For GSM8K, our mathematical reasoning benchmark, we employ a placeholder-based preprocessing step to preserve the integrity of mathematical expressions during translation. Specifically, dataset-specific markers such as  $\langle\langle \cdot \rangle\rangle$  and ##### are temporarily replaced with placeholders prior to translation and subsequently restored to their original form.

For RACE, we adopt a sentence-aware chunking strategy to accommodate the substantially longer reading passages compared to GSM8K and MMLU. The articles, questions, and answer options are

translated independently, after which we reconstruct the hierarchical reading comprehension structure while strictly preserving the original evaluation schema.

To ensure translation reliability, we manually inspected a random sample of approximately 10% of the translated instances.

## B Evaluation

To assess model performance before and after weight manipulation, we use the *EleutherAI Evaluation Harness* framework [Gao et al. \(2024\)](#). In addition to English, we evaluate on German, Hindi, and French by implementing custom `lm_eval` tasks for GSM8K, RACE, and MMLU that mirror the original English task specifications (i.e. matching dataset splits, preprocessing, and scoring). For GSM8K, we adopt the `gsm8k_cot` configuration, which uses an 8-shot chain-of-thought prompting format and following prior work [Christ et al. \(2025\)](#); [Brown et al. \(2024\)](#), we use a subset of 200 samples of GSM8K test set for efficacy. Across languages, we evaluate on the same fixed set of examples per task to ensure comparability of results. All evaluations are repeated for three different random seeds, and the mean and standard deviations are reported accordingly. To further facilitate a fair comparison across settings, we increase `max_new_tokens` from its default of 150 to 300 for evaluations on Hindi and French. This adjustment accounts for the possibly higher token counts needed to represent Hindi and French texts compared to math responses in other languages due to script differences. It is worth noting that employing varying values of `max_new_tokens` for each language does not undermine the objectives of our experiments. Our primary focus is on assessing the relative improvements or deteriorations observed in the intervention experiments for each language independently.

## C Jaccard Random Baseline

**Setup (fixed-size, independent masks)** Let  $U$  be a universe of  $N$  parameters. Each set  $S_i \subseteq U$  is sampled *independently and uniformly* from all subsets of size

$$|S_i| = m, \quad m = pN,$$

where  $p \in (0, 1)$  and  $m$  is assumed to be an integer. For any fixed parameter  $\theta \in U$ ,

$$\Pr(\theta \in S_i) = \frac{m}{N} = p, \quad \Pr(\theta \notin S_i) = 1 - p.$$

**Difference sets** Define the one-sided difference sets

$$A = S_1 \setminus S_2, \quad B = S_3 \setminus S_4,$$

where all  $S_i$  are independent.

A parameter  $\theta$  belongs to  $A$  if and only if  $\theta \in S_1$  and  $\theta \notin S_2$ . Since  $S_1$  and  $S_2$  are independent,

$$\Pr(\theta \in A) = \Pr(\theta \in S_1) \Pr(\theta \notin S_2) = p(1 - p).$$

The same holds for  $B$ :

$$\Pr(\theta \in B) = p(1 - p).$$

**Expected Jaccard similarity** The Jaccard similarity between  $A$  and  $B$  is

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

For any fixed  $\theta \in U$ , membership in  $A$  and  $B$  depends on disjoint pairs of sets and is therefore independent. Hence,

$$\Pr(\theta \in A \cap B) = [p(1 - p)]^2,$$

and

$$\begin{aligned} & \Pr(\theta \in A \cup B) \\ &= \Pr(\theta \in A) + \Pr(\theta \in B) - \Pr(\theta \in A \cap B) \\ &= 2p(1 - p) - [p(1 - p)]^2. \end{aligned}$$

For large  $N$ , the random set sizes concentrate around their expectations, and the expected Jaccard similarity is well approximated by the ratio of expectations:

$$\mathbb{E}[J(A, B)] \approx \frac{[p(1 - p)]^2}{2p(1 - p) - [p(1 - p)]^2}.$$

Factoring out  $p(1 - p)$  yields the baseline expression:

$$\mathbb{E}[J(A, B)] \approx \frac{p(1 - p)}{2 - p(1 - p)} = \frac{p(1 - p)}{p^2 - p + 2}.$$

It is worth noting that, the approximation arises from replacing  $\mathbb{E}[|A \cap B|/|A \cup B|]$  by  $\mathbb{E}|A \cap B|/\mathbb{E}|A \cup B|$ . This becomes accurate when  $N$  is large due to concentration of measure.

## D Intervention Results

In the following, we present results for scaling (Tables D.1-D.6) and pruning (Tables D.7-D.12) math-specific parameters across different models and multiple languages. These results serve as proof that the parameter identification for the task and language was performed correctly. In addition to accuracies on GSM8K, MMLU, and RACE, we also report the flexible-extract GSM8K score from the `lm_eval` (*lm-evaluation-harness*) framework. The harness computes GSM8K exact-match accuracy

after applying an answer-extraction step: a *strict* extractor that follows the dataset’s canonical output format (e.g., The answer is <answer>) and a more permissive *flexible* extractor that searches the generated text for a number-like substring. We include the flexible metric to reduce sensitivity to output formatting, as our analysis is not focused on instruction-following behavior.

### D.1 Robustness of parameter identification

Across all settings both in scaling and pruning experiments, the standard deviation across three random seeds remains low. We take this as evidence that the identified parameter subset captures math-reasoning capacity in a stable manner, rather than reflecting idiosyncrasies of a particular sample set. Moreover, all scaling experiments yield highly stable accuracies on MMLU and RACE, indicating that scaling the identified parameters does not measurably affect general language understanding. For pruning, the within-dataset variability is more sensitive to the choice of top- $k$ , but remains within a bounded and reasonably stable range. This confirms that we isolate parameter subsets that drive math reasoning while being largely irrelevant to general language understanding, which is of high importance to our parameter comparison and Jaccard similarity interpretations.

### D.2 Pruning results

Across all settings, increasing top- $k$  leads to a substantial decline in math-reasoning performance, while non-math performance remains comparatively stable on MMLU and RACE (Tables D.7-D.12). This selective degradation is the intended effect and indicates that our parameter identification is successful also across model sizes.

However, surprisingly, when pruning the Qwen3 model with a top- $k=1e-5$  threshold, corresponding to suppressing approximately 16K parameters, the performance on English GSM8K improves substantially, whereas only a marginal improvement is observed on English GSM8K-Flex shown in Table D.9, which may be a strong signal of wrong output formatting. This counter-intuitive gain from parameter reduction aligns with the framework proposed in (Ali et al., 2025), which identifies a class of *prominent but detrimental neurons*. These are highly activated units that, while capturing frequent training patterns, often enforce brittle heuristics that hinder the model’s ability to generalize to nuanced reasoning tasks. Manual inspection reveals

that this improvement is only partly due to corrections in arithmetic computation; more importantly, it stems from enhanced in-context instruction following. This manifests as shorter reasoning traces that are less likely to be truncated by the maximum generation length, and it also outputs the required output formats. This stabilization of the output path is consistent with the AdaRAS by Dong et al. (2026), which suggests that *over-active* reasoning circuits can lead to logical redundancy and verbosity. By pruning these circuits, we effectively guide the model toward more direct and stable reasoning paths, thereby better mimicking the reasoning style demonstrated in the provided examples.

Figure E.9 illustrates a representative example. Prior to pruning, the model persistently follows its own internal reasoning style, presumably acquired during post-training, rather than adhering to the few-shot exemplars. After suppressing approximately 16K of the most highly activated parameters, the model more consistently aligns its reasoning format with that of the provided demonstrations. This behavior provides empirical support for the theory of *Copy Bias* in In-Context Learning (Ali et al., 2024). According to this work, specific parameters act as *stylistic anchors* that prioritize internal priors over contextual cues; pruning them releases the model from these rigid post-training biases, allowing for greater *behavioral flexibility*.

### D.3 Scaling results

In contrast to our smallest model (Llama 1B), scaling does not yield significant consistent gains on larger models. Given the strong pruning results, we attribute this mainly to an insufficient scaling factor (1.01) rather than to low-quality parameter identification. Since optimizing this hyperparameter is beyond the scope of this work and does not affect our cross-language comparisons, we leave a systematic search over scaling factors to future work.

However, one can observe a relationship between the model’s general language understanding ability and the effect of scaling on performance, including indications of a saturation effect. For the small Llama 1B model, scaling yields performance gains (in English) of up to 9.7%. For larger models, performance also improves in some cases, but the changes are much smaller, ranging only from -1.4% to 2.4% compared to the baseline. This suggests that scaling is most effective when the model has not yet reached a level of saturated performance.

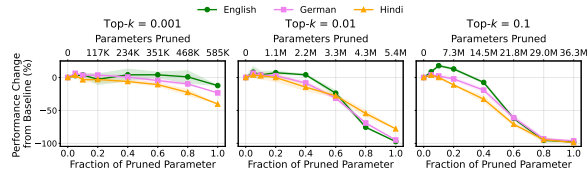


Figure D.1: Effect of pruning isolated math-specific parameters on performance across languages for Qwen3 4B at  $\text{top-}k \in \{0.001, 0.01, 0.1\}$ .

In cases where performance is already high, scaling may be limited because the task-relevant parameters have already been sufficiently optimized, rather than because the task-specific weights were poorly identified. For German, Hindi, and French, we frequently observe that baseline general language understanding performance (e.g., measured by MMLU or RACE) influences the gains obtained from scaling. When language understanding performance is too low, as is often the case for Hindi, scaling tends to produce only small positive or even negative effects. One possible explanation is that insufficient language knowledge prevents the model from identifying the optimal language-specific parameters required for mathematical reasoning. For instance, Qwen3 4B shows consistent improvements from scaling for both German and Hindi; however, the gains are larger for German, which has higher baseline language understanding scores on RACE and MMLU. Overall, these results suggest that a sufficient level of language understanding is beneficial for the scaling approach, while very high baseline performance can lead to saturation effects that limit further improvements through scaling.

### D.4 Collective vs. Individual Effects

We examine whether the identified math-specific parameters operate primarily through a collective mechanism or exert independent (individual) effects. For all experiments in this section, we set `max_new_tokens` to 300 across all languages and use RACE as the non-math dataset.

As discussed in Section 4.5 and shown in Figure 4, the Llama 1B model exhibits collective behavior. This trend persists in larger models: Figures D.1 and D.2 show that both Qwen3 4B and Llama 8B demonstrate similar collective dynamics, suggesting that this phenomenon is consistent across model scales.

Top- <i>k</i>	English			German			Hindi			French		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.340	0.345	0.379	0.235	0.230	0.321	0.145	0.160	0.299	0.185	0.195	0.348
0.000001	+0.02,9%	+0.02,9%	+0.50,0%	+0.90,0%	+1.34,3%	+0.00,0%	-9.06,9%	-7.56,2%	-0.30,0%	+1.10,0%	+0.05,1%	+0.00,0%
0.00001	+0.65,9%	+0.65,8%	+0.80,0%	-2.14,3%	-2.24,3%	+0.30,0%	-1.46,9%	-1.96,2%	-0.30,0%	+6.510,8%	+5.115,4%	-0.30,0%
0.0001	+2.48,8%	+2.38,7%	+0.30,0%	-2.14,3%	-0.94,3%	-0.30,0%	+1.46,9%	+0.06,2%	+0.30,0%	+6.55,4%	+5.15,1%	+0.00,0%
0.001	+6.82,9%	+6.72,9%	+1.10,0%	-5.58,5%	-4.313,0%	+0.00,0%	+4.86,9%	+3.16,2%	+0.30,0%	+1.15,4%	-1.55,1%	+0.00,0%
0.005	+6.52,9%	+6.42,9%	+0.30,0%	-3.00,0%	-2.20,0%	+0.60,0%	+0.013,8%	-4.42,5%	+0.70,0%	+8.10,0%	+6.75,1%	-0.30,0%
0.01	+2.90,0%	+3.50,0%	+0.50,0%	+1.34,3%	+3.54,3%	+0.00,0%	+2.16,9%	+3.16,2%	+0.70,0%	+10.80,0%	+8.75,1%	-0.30,0%
0.025	+5.02,9%	+4.92,9%	+0.30,0%	+3.04,3%	+5.24,3%	+0.30,0%	-4.86,9%	-4.46,2%	+0.30,0%	+2.75,4%	+0.05,1%	+0.00,0%
0.05	+5.30,0%	+4.90,0%	+0.50,0%	+3.44,3%	+5.20,0%	+0.30,0%	-3.40,0%	-6.36,2%	+0.30,0%	+7.05,4%	+6.75,1%	+0.00,0%
0.1	+3.52,9%	+3.52,9%	+0.30,0%	+0.94,3%	+2.20,0%	+0.90,0%	-5.56,9%	-3.16,2%	+0.30,0%	+8.15,4%	+8.75,1%	-0.30,0%
0.15	+3.52,9%	+3.52,9%	+0.00,0%	-1.38,5%	-1.38,7%	+0.30,0%	-4.80,0%	-3.16,2%	+0.70,0%	+5.45,4%	+3.65,1%	+0.00,0%

Table D.1: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for scaling with Llama 1B and RACE.

Top- <i>k</i>	English			German			Hindi			French		
	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU
0.0 (Pre-train)	0.340	0.345	0.458	0.235	0.230	0.363	0.145	0.160	0.322	0.185	0.195	0.360
0.000001	+3.52,9%	+3.52,9%	+0.20,0%	+3.04,3%	+3.54,3%	+0.30,0%	-1.46,9%	-3.16,2%	+0.30,0%	+4.35,4%	+1.55,1%	+0.00,0%
0.00001	+2.42,9%	+2.32,9%	+0.20,0%	-5.54,3%	-4.34,3%	+0.30,0%	+4.80,0%	+3.16,2%	+0.30,0%	+3.85,4%	+2.60,0%	+0.00,0%
0.0001	+1.52,9%	+1.42,9%	+0.40,0%	+3.04,3%	+3.04,3%	+0.00,0%	-4.80,0%	-3.16,2%	+0.30,0%	+5.45,4%	+3.65,1%	+0.00,0%
0.001	+1.55,9%	+1.45,8%	+0.70,0%	-5.18,5%	-5.28,7%	+0.00,0%	-1.40,0%	-1.30,0%	+0.30,0%	+0.05,4%	-1.05,1%	+0.00,0%
0.005	+3.55,9%	+3.55,8%	+0.20,0%	+1.30,0%	+2.20,0%	+0.00,0%	-2.16,9%	+0.06,2%	+0.30,0%	+8.10,0%	+6.25,1%	+0.00,0%
0.01	+6.52,9%	+6.42,9%	+0.40,0%	+4.34,3%	+4.34,3%	+0.30,0%	-4.86,9%	-3.16,2%	+0.60,0%	+8.10,0%	+6.75,1%	-0.30,0%
0.025	+5.32,9%	+5.82,9%	+0.20,0%	-5.14,3%	-3.04,3%	+0.60,0%	-11.76,9%	-10.60,0%	+0.30,0%	+3.85,4%	+2.610,3%	+0.00,0%
0.05	+3.52,9%	+3.52,9%	+0.40,0%	-0.94,3%	+0.04,3%	+0.30,0%	-1.40,0%	-3.16,2%	+0.30,0%	+3.85,4%	+4.110,3%	-0.60,0%
0.1	+6.82,9%	+6.72,9%	+0.20,0%	+1.30,0%	+2.20,0%	+0.30,0%	-9.06,9%	-9.46,2%	+0.30,0%	+16.25,4%	+15.40,0%	-0.30,0%
0.15	+9.75,9%	+9.35,8%	+0.20,0%	+1.34,3%	+1.34,3%	+0.30,0%	-4.86,9%	-4.46,2%	+0.60,0%	+9.25,4%	+7.75,1%	+0.00,0%

Table D.2: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for scaling with Llama 1B and MMLU.

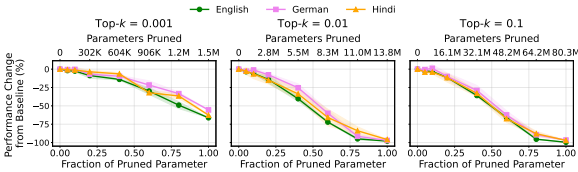


Figure D.2: Effect of pruning isolated math-specific parameters on performance across languages for Llama 8B at top-*k*  $\in$  {0.001, 0.01, 0.1}.

## E Qualitative Examples for Scaling & Pruning

In this section, we present several qualitative examples where scaling improves the model’s initial answers for mathematical reasoning across languages (Figures E.1-E.4). Furthermore, we show examples where pruning steers the model toward incorrect answers for mathematical reasoning (Figures E.5-E.8), alongside a notable exception in which pruning instead improves performance (Figure E.9), as discussed in Appendix D.

## F Similarities across Languages & Tasks

In this section, we present plots showing the number of isolated parameters across languages (English (En), German (De), Hindi (Hi), and French (Fr)), as well as Jaccard similarity plots for each language pair, across all top-*k* values and models.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.735	0.815	0.414	0.685	0.400	0.356	0.385	0.460	0.353	0.690	0.615	0.403
0.000001	+0.30,0%	+0.00,0%	-0.20,0%	+3.91,5%	+5.52,5%	+0.30,0%	-0.50,0%	-1.52,2%	-0.60,0%	+2.21,4%	+3.63,3%	+0.20,0%
0.00001	-0.31,4%	+2.10,0%	+0.00,0%	+3.91,5%	+13.75,0%	+0.80,0%	+2.60,0%	+1.52,2%	-0.80,0%	+1.71,4%	+2.91,6%	+0.70,0%
0.0001	+0.41,4%	+0.21,2%	-0.70,0%	+4.41,5%	+7.02,5%	+1.10,0%	+0.50,0%	-0.74,3%	-1.10,0%	+2.50,0%	+1.30,0%	+0.50,0%
0.001	+1.00,0%	+1.82,5%	-0.20,0%	+2.51,5%	+5.75,0%	+0.60,0%	+0.80,0%	-1.12,2%	-0.30,0%	+3.61,4%	+3.71,6%	+0.50,0%
0.005	+0.41,4%	+1.61,2%	+0.00,0%	+2.61,5%	+8.07,5%	+0.60,0%	+0.52,6%	-0.74,3%	-1.40,0%	+2.51,4%	+5.43,3%	+0.50,0%
0.01	+0.31,4%	+1.21,2%	+0.20,0%	+5.50,0%	+16.80,0%	+0.30,0%	+0.82,6%	-2.62,2%	-0.60,0%	+2.51,4%	+2.91,6%	+0.50,0%
0.025	+0.41,4%	-0.42,5%	+0.00,0%	+3.91,5%	+13.20,0%	+0.30,0%	+0.02,6%	-1.72,2%	-0.60,0%	+3.61,4%	+3.31,6%	+0.20,0%
0.05	-0.41,4%	+0.01,2%	-0.70,0%	+4.81,5%	+8.02,5%	+0.80,0%	+1.82,6%	+1.70,0%	-0.60,0%	+1.91,4%	+2.01,6%	+0.20,0%
0.1	+0.31,4%	+0.42,5%	-0.50,0%	+2.91,5%	+8.27,5%	-0.30,0%	-0.80,0%	-1.52,2%	-0.80,0%	+2.91,4%	+3.61,6%	+0.70,0%
0.15	-1.41,4%	+0.42,5%	-0.50,0%	+4.11,5%	+8.07,5%	+0.30,0%	+0.02,6%	+0.44,3%	+0.00,0%	+1.91,4%	+1.31,6%	+0.70,0%

Table D.3: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for scaling with Qwen3 4B and RACE.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU
0.0 (Pre-train)	0.735	0.815	0.710	0.685	0.400	0.585	0.385	0.460	0.471	0.690	0.615	0.548
0.000001	+1.81,4%	+1.21,2%	+0.00,0%	+2.61,5%	+9.27,5%	-0.20,0%	+0.00,0%	-2.22,2%	+0.20,0%	+1.70,0%	+2.01,6%	+0.20,0%
0.00001	+0.31,4%	+0.00,0%	+0.00,0%	+3.41,5%	+2.05,0%	-0.20,0%	+1.32,6%	+1.12,2%	+0.20,0%	+2.91,4%	+3.31,6%	+0.40,0%
0.0001	+1.11,4%	-1.80,0%	+0.00,0%	+4.11,5%	+10.05,0%	-0.20,0%	+0.50,0%	-1.72,2%	+0.20,0%	+2.21,4%	+0.01,6%	+0.20,0%
0.001	+2.01,4%	+0.90,0%	+0.00,0%	+2.91,5%	+6.72,5%	-0.20,0%	+0.50,0%	+0.02,2%	+0.20,0%	+2.50,0%	+2.41,6%	+0.20,0%
0.005	+0.71,4%	-1.82,5%	+0.10,0%	+3.41,5%	+8.05,0%	-0.20,0%	+1.32,6%	+2.24,3%	+0.20,0%	+1.21,4%	+2.91,6%	+0.20,0%
0.01	+0.31,4%	+1.50,0%	+0.00,0%	+4.81,5%	+8.22,5%	-0.20,0%	+1.32,6%	-0.44,3%	+0.20,0%	+3.21,4%	+2.90,0%	+0.20,0%
0.025	+1.02,7%	-1.61,2%	+0.00,0%	+4.41,5%	+13.05,0%	-0.20,0%	+0.52,6%	-1.72,2%	+0.20,0%	+2.21,4%	+2.41,6%	+0.20,0%
0.05	-1.01,4%	-1.00,0%	+0.00,0%	+4.11,5%	+7.55,0%	-0.20,0%	+0.52,6%	-1.12,2%	+0.40,0%	+2.21,4%	+1.61,6%	+0.20,0%
0.1	-0.41,4%	-0.91,2%	+0.00,0%	+3.21,5%	+11.25,0%	-0.20,0%	+0.50,0%	-0.44,3%	+0.20,0%	+1.41,4%	-0.30,0%	+0.20,0%
0.15	-0.31,4%	-0.92,5%	+0.00,0%	+4.11,5%	+8.25,0%	-0.20,0%	-0.52,6%	+0.02,2%	+0.20,0%	+1.91,4%	+0.81,6%	+0.00,0%

Table D.4: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for scaling with Qwen3 4B and MMLU.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.765	0.775	0.448	0.585	0.590	0.396	0.415	0.490	0.388	0.580	0.585	0.413
0.000001	+1.00,0%	+0.01,3%	+0.00,0%	+0.31,7%	+0.81,7%	-0.50,0%	+3.62,4%	+1.60,0%	+0.00,0%	+2.11,7%	+2.11,7%	+0.20,0%
0.00001	+2.40,0%	+1.90,0%	+0.20,0%	+1.21,7%	+2.21,7%	-0.30,0%	+3.10,0%	+0.02,0%	+0.00,0%	+0.51,1%	+0.33,4%	+0.20,0%
0.0001	+0.01,3%	-0.31,3%	+0.00,0%	+2.90,0%	+3.70,0%	-0.30,0%	+0.50,0%	-2.70,0%	+0.30,0%	+0.33,4%	+0.33,4%	+0.50,0%
0.001	+1.01,3%	+1.01,3%	-0.40,0%	+0.91,7%	+2.51,7%	-0.30,0%	-0.54,8%	-2.02,0%	+0.00,0%	+1.71,7%	+1.71,7%	+0.20,0%
0.005	+1.61,3%	+0.61,3%	+0.20,0%	+0.51,7%	+0.50,0%	+0.00,0%	+5.30,0%	+1.60,0%	+0.30,0%	+0.33,4%	-0.33,4%	+0.50,0%
0.01	+0.71,3%	+0.02,6%	+0.20,0%	+0.51,7%	+0.81,7%	-0.30,0%	+0.74,8%	-1.44,1%	+0.00,0%	+0.51,7%	+0.31,7%	+0.20,0%
0.025	+1.60,0%	+0.60,0%	-0.40,0%	+1.71,7%	+2.91,7%	-0.50,0%	+1.92,4%	+0.60,0%	+0.30,0%	+1.21,7%	+1.21,7%	+0.20,0%
0.05	+2.21,3%	+1.50,0%	-0.40,0%	+2.60,0%	+3.90,0%	+0.00,0%	+2.92,4%	+0.42,0%	+0.30,0%	+0.50,0%	+0.50,0%	+0.50,0%
0.1	+0.31,3%	-0.41,3%	+0.00,0%	-0.31,7%	+0.31,7%	+0.00,0%	+3.10,0%	+0.62,0%	+0.30,0%	+0.01,7%	-0.31,7%	+0.00,0%
0.15	+0.71,3%	+0.32,6%	-0.20,0%	+0.01,7%	+0.50,0%	+0.00,0%	+1.70,0%	-2.02,0%	+0.00,0%	+0.30,0%	+0.30,0%	+0.20,0%

Table D.5: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for scaling with Llama 8B and RACE.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU
0.0 (Pre-train)	0.765	0.775	0.679	0.585	0.590	0.538	0.415	0.490	0.450	0.580	0.585	0.512
0.000001	+1.60,0%	+1.31,3%	-0.10,0%	+1.21,7%	+1.71,7%	+0.20,0%	-0.52,4%	-3.12,0%	+1.10,0%	+0.50,0%	+0.30,0%	+0.20,0%
0.00001	+0.71,3%	+0.41,3%	-0.10,0%	+0.01,7%	+0.81,7%	+0.20,0%	+0.52,4%	+0.02,0%	+1.10,0%	-0.31,7%	-0.31,7%	+0.20,0%
0.0001	+1.00,0%	+1.00,0%	-0.10,0%	+2.20,0%	+2.91,7%	+0.00,0%	+1.92,4%	+1.02,0%	+1.10,0%	+0.31,7%	+0.31,7%	+0.20,0%
0.001	+1.61,3%	+0.41,3%	-0.10,0%	+0.91,7%	+2.51,7%	+0.20,0%	-0.72,4%	-3.12,0%	+1.10,0%	+0.00,0%	+0.00,0%	+0.20,0%
0.005	+2.01,3%	+1.91,3%	-0.10,0%	+0.01,7%	+1.71,7%	+0.20,0%	+1.22,4%	-0.42,0%	+0.90,0%	-1.73,4%	-2.23,4%	+0.20,0%
0.01	+2.21,3%	+1.31,3%	-0.10,0%	+0.51,7%	+0.81,7%	+0.00,0%	+2.42,4%	-0.44,1%	+0.90,0%	+0.03,4%	-0.33,4%	+0.00,0%
0.025	+0.71,3%	-0.41,3%	-0.10,0%	+0.51,7%	+1.41,7%	+0.20,0%	+1.72,4%	-1.02,0%	+1.10,0%	-1.21,7%	-1.21,7%	+0.00,0%
0.05	+1.02,6%	+0.91,3%	-0.10,0%	+0.91,7%	+1.21,7%	+0.20,0%	+3.60,0%	+0.40,0%	+1.10,0%	-2.61,7%	-2.61,7%	+0.00,0%
0.1	-0.31,3%	-0.91,3%	+0.00,0%	+1.71,7%	+2.91,7%	+0.20,0%	+1.94,8%	+0.02,0%	+0.90,0%	+0.30,0%	+0.30,0%	+0.20,0%
0.15	+1.01,3%	+0.62,6%	-0.10,0%	+0.01,7%	+0.51,7%	+0.20,0%	-0.52,4%	-3.12,0%	+0.90,0%	-0.53,4%	-1.21,7%	+0.20,0%

Table D.6: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for scaling with Llama 8B and MMLU.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.340	0.345	0.379	0.235	0.230	0.321	0.145	0.160	0.299	0.185	0.195	0.348
0.000001	+9.72.9%	+8.72.9%	+0.30.0%	-9.44.3%	-6.54.3%	+1.20.0%	-9.06.9%	-8.112.5%	+0.70.0%	-3.85.4%	-4.15.1%	-0.30.0%
0.00001	-7.92.9%	-9.32.9%	+1.62.6%	-18.312.8%	-16.113.0%	+0.00.0%	-31.020.7%	-29.418.8%	+2.00.0%	-7.05.4%	-8.70.0%	-1.42.9%
0.0001	-25.02.9%	-26.12.9%	-0.50.0%	-31.14.3%	-29.64.3%	+2.83.1%	-43.46.9%	-31.218.8%	-1.00.0%	-17.35.4%	-19.05.1%	-0.90.0%
0.001	-81.55.9%	-80.65.8%	-0.80.0%	-66.88.5%	-65.28.7%	+0.30.0%	-75.96.9%	-70.612.5%	-3.70.0%	-59.510.8%	-60.010.3%	-5.20.0%
0.005	-94.72.9%	-94.22.9%	-2.12.6%	-89.44.3%	-91.33.7%	-2.83.1%	-87.66.9%	-84.46.2%	-5.73.3%	-90.35.4%	-92.35.1%	-6.00.0%
0.01	-96.22.9%	-96.22.9%	-2.12.6%	-92.30.0%	-90.44.3%	-1.20.0%	-94.56.9%	-92.56.2%	-6.40.0%	-90.80.0%	-90.80.0%	-6.30.0%
0.025	-92.60.0%	-92.20.0%	-9.00.0%	-90.60.0%	-90.44.3%	-3.13.1%	-96.66.9%	-90.66.2%	-8.00.0%	-88.15.4%	-88.25.1%	-6.60.0%
0.05	-93.55.9%	-93.65.8%	-20.15.3%	-89.44.3%	-91.34.3%	-7.20.0%	-95.26.9%	-91.90.0%	-8.03.3%	-88.15.4%	-88.75.1%	-8.60.0%
0.1	-93.50.0%	-93.60.0%	-12.42.6%	-93.64.3%	-90.44.3%	-10.33.1%	-93.16.9%	-92.50.0%	-9.03.3%	-94.65.4%	-91.30.0%	-12.92.9%
0.15	-96.52.9%	-95.70.0%	-13.72.6%	-92.38.5%	-90.44.3%	-10.33.1%	-97.96.9%	-88.86.2%	-8.03.3%	-88.15.4%	-88.710.3%	-10.10.0%

Table D.7: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for pruning with Llama 1B and RACE.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU
0.0 (Pre-train)	0.340	0.345	0.458	0.235	0.230	0.363	0.145	0.160	0.322	0.185	0.195	0.360
0.000001	+3.88.8%	+2.98.7%	-0.20.0%	-9.44.3%	-8.74.3%	-0.60.0%	-15.96.9%	-18.86.2%	-1.20.0%	+0.05.4%	-3.65.1%	-1.10.0%
0.00001	-14.72.9%	-15.40.0%	-3.70.0%	-18.312.8%	-15.213.0%	-0.80.0%	-35.96.9%	-30.06.2%	-0.30.0%	-18.95.4%	-21.55.1%	-4.72.8%
0.0001	-19.78.8%	-20.98.7%	-2.40.0%	-34.04.3%	-33.54.3%	-0.60.0%	-35.913.8%	-29.46.2%	-1.90.0%	-11.95.4%	-15.45.1%	-0.60.0%
0.001	-80.32.9%	-80.32.9%	-7.62.2%	-67.212.8%	-67.48.7%	-5.52.8%	-46.213.8%	-48.812.5%	-3.40.0%	-55.710.8%	-57.410.3%	-4.70.0%
0.005	-90.62.9%	-89.92.9%	-6.62.2%	-91.54.3%	-91.34.3%	-8.30.0%	-88.36.9%	-83.16.2%	-4.73.1%	-87.65.4%	-88.75.1%	-9.20.0%
0.01	-92.62.9%	-92.22.9%	-12.22.2%	-86.44.3%	-85.70.0%	-11.80.0%	-91.06.9%	-91.96.2%	-9.30.0%	-90.30.0%	-88.70.0%	-15.32.8%
0.025	-92.15.9%	-92.25.8%	-25.36.6%	-95.74.3%	-92.24.3%	-14.92.8%	-98.60.0%	-93.80.0%	-14.93.1%	-88.15.4%	-88.75.1%	-13.12.8%
0.05	-91.82.9%	-91.32.9%	-27.54.4%	-92.84.3%	-90.48.7%	-14.32.8%	-96.66.9%	-93.86.2%	-15.53.1%	-90.30.0%	-90.80.0%	-17.25.6%
0.1	-89.12.9%	-88.42.9%	-15.90.0%	-94.54.3%	-92.64.3%	-17.40.0%	-95.20.0%	-92.56.2%	-14.36.2%	-91.90.0%	-90.80.0%	-11.90.0%
0.15	-91.82.9%	-89.92.9%	-13.82.2%	-90.64.3%	-90.40.0%	-15.22.8%	-94.56.9%	-92.56.2%	-16.83.1%	-90.35.4%	-84.65.1%	-10.82.8%

Table D.8: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for pruning with Llama 1B and MMLU.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.735	0.815	0.414	0.685	0.400	0.356	0.385	0.460	0.353	0.690	0.615	0.403
0.000001	-21.89.5%	-6.13.7%	+1.00.0%	-0.32.9%	+4.52.5%	-0.30.0%	+8.62.6%	-7.019.6%	-3.70.0%	+4.12.9%	+11.43.3%	-0.52.5%
0.00001	+19.34.1%	+4.72.5%	-0.70.0%	-3.21.5%	+24.227.5%	-1.10.0%	-1.87.8%	-24.310.9%	-3.70.0%	+2.91.4%	+3.69.8%	-5.20.0%
0.0001	+1.46.8%	-15.78.6%	-2.70.0%	-22.64.4%	-28.027.5%	-0.30.0%	-19.07.8%	-31.523.9%	-7.42.8%	-16.45.8%	-32.826.0%	-3.72.5%
0.001	-12.28.2%	-55.58.6%	-3.62.4%	-23.45.8%	+1.35.0%	-0.62.8%	-40.32.6%	-13.54.3%	-3.42.8%	-28.01.4%	-59.717.9%	-3.72.5%
0.005	-77.16.8%	-90.46.1%	-1.00.0%	-62.51.5%	-45.87.5%	-0.80.0%	-50.17.8%	-46.38.7%	-5.10.0%	-67.41.4%	-93.04.9%	-4.00.0%
0.01	-97.31.4%	-97.31.2%	-2.92.4%	-94.91.5%	-93.22.5%	+3.12.8%	-77.97.8%	-75.74.3%	-5.40.0%	-95.71.4%	-94.81.6%	-0.70.0%
0.025	-97.60.0%	-97.51.2%	-4.30.0%	-97.51.5%	-93.02.5%	-1.72.8%	-97.42.6%	-95.22.2%	-5.92.8%	-98.31.4%	-96.40.0%	-3.50.0%
0.05	-97.60.0%	-97.80.0%	-3.62.4%	-96.81.5%	-94.52.5%	-1.12.8%	-97.40.0%	-96.12.2%	-7.40.0%	-98.61.4%	-95.61.6%	-3.20.0%
0.1	-98.01.4%	-96.72.5%	-4.12.4%	-96.11.5%	-93.20.0%	+0.30.0%	-98.72.6%	-96.12.2%	-7.12.8%	-98.61.4%	-97.91.6%	-4.20.0%
0.15	-97.71.4%	-97.31.2%	-3.12.4%	-95.31.5%	-94.52.5%	-1.40.0%	-97.42.6%	-96.72.2%	-7.90.0%	-97.41.4%	-97.91.6%	-6.52.5%

Table D.9: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for pruning with Qwen3 4B and RACE.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU
0.0 (Pre-train)	0.735	0.815	0.710	0.685	0.400	0.585	0.385	0.460	0.471	0.690	0.615	0.548
0.000001	-6.85.4%	+0.22.5%	-0.70.0%	+3.92.9%	+38.812.5%	-0.70.0%	+7.32.6%	-5.44.3%	-3.40.0%	+2.61.4%	+7.03.3%	+0.00.0%
0.00001	+8.84.1%	-13.38.6%	-0.80.0%	+0.71.5%	-0.515.0%	-2.41.7%	+7.05.2%	-53.92.2%	-2.50.0%	+6.11.4%	+10.63.3%	-0.20.0%
0.0001	-14.76.8%	-39.03.7%	-1.00.0%	-11.40.0%	-18.315.0%	-2.20.0%	-15.12.6%	+5.04.3%	-9.32.1%	-7.51.4%	-17.416.3%	-0.40.0%
0.001	-15.41.4%	-27.07.4%	-2.50.0%	-18.52.9%	-32.512.5%	-6.30.0%	-21.35.2%	-7.02.2%	-5.12.1%	-13.32.9%	-17.98.1%	-3.60.0%
0.005	-79.22.7%	-90.61.2%	-3.90.0%	-68.21.5%	-64.27.5%	-6.00.0%	-61.85.2%	-52.24.3%	-8.72.1%	-64.52.9%	-75.91.6%	-2.20.0%
0.01	-96.61.4%	-97.30.0%	-6.80.0%	-94.61.5%	-89.52.5%	-6.70.0%	-78.75.2%	-76.72.2%	-8.92.1%	-92.02.9%	-89.93.3%	-2.90.0%
0.025	-97.71.4%	-97.81.2%	-8.20.0%	-97.51.5%	-97.02.5%	-8.20.0%	-94.32.6%	-93.02.2%	-9.10.0%	-95.72.9%	-94.31.6%	-9.31.8%
0.05	-98.91.4%	-98.41.2%	-9.40.0%	-95.61.5%	-93.80.0%	-10.10.0%	-98.72.6%	-96.72.2%	-10.82.1%	-98.80.0%	-97.11.6%	-10.01.8%
0.1	-98.91.4%	-96.61.2%	-8.30.0%	-97.51.5%	-95.82.5%	-8.00.0%	-97.92.6%	-96.14.3%	-7.90.0%	-97.81.4%	-97.11.6%	-6.91.8%
0.15	-95.61.4%	-95.31.2%	-7.90.0%	-96.40.0%	-94.20.0%	-7.41.7%	-96.12.6%	-95.02.2%	-8.10.0%	-96.72.9%	-96.41.6%	-5.30.0%

Table D.10: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for pruning with Qwen3 4B and MMLU.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.765	0.775	0.448	0.585	0.590	0.396	0.415	0.490	0.388	0.580	0.585	0.413
0.000001	-2.43.9%	-3.23.9%	+0.20.0%	+2.11.7%	+1.71.7%	+0.30.0%	-8.914.5%	-2.46.1%	+1.00.0%	-7.46.9%	-10.38.5%	+0.20.0%
0.00001	+0.31.3%	-0.62.6%	+1.80.0%	-4.81.7%	-5.91.7%	+0.80.0%	-10.821.7%	-9.82.0%	-2.30.0%	-14.18.6%	-19.712.0%	+0.50.0%
0.0001	-15.42.6%	-15.52.6%	-0.40.0%	-22.61.7%	-22.41.7%	+0.30.0%	-22.49.6%	-24.14.1%	-2.62.6%	-22.98.6%	-30.418.8%	-3.10.0%
0.001	-66.33.9%	-66.52.6%	-2.92.2%	-55.65.1%	-54.63.4%	-3.52.5%	-62.77.2%	-60.86.1%	-5.20.0%	-60.05.2%	-59.85.1%	-3.42.4%
0.005	-94.81.3%	-93.82.6%	-4.00.0%	-93.21.7%	-92.71.7%	-4.52.5%	-94.72.4%	-94.32.0%	-8.02.6%	-92.81.7%	-92.31.7%	-3.92.4%
0.01	-98.01.3%	-97.71.3%	-4.70.0%	-97.11.7%	-97.11.7%	-4.00.0%	-95.92.4%	-94.94.1%	-6.70.0%	-97.90.0%	-97.11.7%	-7.52.4%
0.025	-97.81.3%	-96.80.0%	-6.50.0%	-98.31.7%	-97.11.7%	-5.60.0%	-99.50.0%	-99.02.0%	-12.60.0%	-97.11.7%	-96.10.0%	-5.82.4%
0.05	-98.41.3%	-98.71.3%	-2.00.0%	-97.91.7%	-97.81.7%	-9.10.0%	-98.32.4%	-94.90.0%	-12.90.0%	-96.90.0%	-96.61.7%	-8.02.4%
0.1	-99.60.0%	-98.11.3%	-7.60.0%	-96.60.0%	-94.91.7%	-10.92.5%	-97.12.4%	-96.32.0%	-12.92.6%	-97.91.7%	-96.91.7%	-7.04.8%
0.15	-96.70.0%	-94.81.3%	-6.02.2%	-96.11.7%	-95.41.7%	-7.10.0%	-97.62.4%	-95.94.1%	-12.12.6%	-98.31.7%	-97.13.4%	-7.52.4%

Table D.11: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for pruning with Llama 8B and RACE.

Top-k	English			German			Hindi			French		
	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU	GSM8K	GSM8K flex	MMLU
0.0 (Pre-train)	0.765	0.775	0.679	0.585	0.590	0.538	0.415	0.490	0.450	0.580	0.585	0.512
0.000001	-12.06.5%	-13.25.2%	-0.90.0%	+1.20.0%	+0.81.7%	+0.00.0%	-27.74.8%	-6.74.1%	+0.90.0%	-11.65.2%	-14.08.5%	-0.20.0%
0.00001	-4.82.6%	-5.42.6%	-0.40.0%	+0.01.7%	-0.50.0%	-0.40.0%	-15.719.3%	-7.10.0%	+0.40.0%	-11.20.0%	-11.61.7%	-0.40.0%
0.0001	-14.01.3%	-14.81.3%	-2.51.5%	-14.23.4%	-14.45.1%	-0.60.0%	-18.62.4%	-20.44.1%	-0.70.0%	-26.43.4%	-32.010.3%	-1.40.0%
0.001	-81.45.2%	-79.02.6%	-6.61.5%	-60.71.7%	-60.50.0%	-3.20.0%	-56.17.2%	-61.24.1%	-8.70.0%	-63.33.4%	-62.93.4%	-2.70.0%
0.005	-94.41.3%	-94.21.3%	-6.50.0%	-94.01.7%	-92.71.7%	-5.80.0%	-89.22.4%	-89.22.0%	-4.90.0%	-91.41.7%	-91.11.7%	-5.72.0%
0.01	-99.10.0%	-97.81.3%	-8.11.5%	-95.43.4%	-94.93.4%	-8.91.9%	-96.92.4%	-94.92.0%	-9.30.0%	-97.90.0%	-97.40.0%	-4.90.0%
0.025	-98.41.3%	-98.11.3%	-9.41.5%	-97.90.0%	-98.00.0%	-11.50.0%	-99.50.0%	-98.02.0%	-6.90.0%	-96.61.7%	-95.21.7%	-8.60.0%
0.05	-98.30.0%	-96.41.3%	-10.91.5%	-96.61.7%	-96.31.7%	-15.21.9%	-98.10.0%	-96.92.0%	-15.34.4%	-95.33.4%	-94.51.7%	-7.02.0%
0.1	-97.61.3%	-96.42.6%	-17.15.9%	-96.10.0%	-95.30.0%	-20.81.9%	-96.92.4%	-95.30.0%	-21.64.4%	-97.81.7%	-97.10.0%	-10.42.0%
0.15	-92.82.6%	-91.92.6%	-9.60.0%	-94.01.7%	-94.11.7%	-13.01.9%	-93.52.4%	-92.22.0%	-11.62.2%	-94.01.7%	-93.71.7%	-9.02.0%

Table D.12: Relative change in accuracy (%), averaged over 3 seeds (mean  $\pm$  std), for pruning with Llama 8B and MMLU.

**Input**

**Q:** Toula went to the bakery and bought various types of pastries. She bought 3 dozen donuts which cost \$68 per dozen, 2 dozen mini cupcakes which cost \$80 per dozen, and 6 dozen mini cheesecakes for \$55 per dozen. **A:**

**Output before scaling (wrong)**

3 dozen donuts cost  $3 \times 68 = 204$  dollars.  
2 dozen mini cupcakes cost  $2 \times 80 = 160$  dollars.  
6 dozen mini cheesecakes cost  $6 \times 55 = 330$  dollars.  
**204 + 160 + 330 is 734. The answer is 734.**

**Output after scaling (correct)**

3 dozen donuts cost  $3 \times 68 = 204$  dollars.  
2 dozen mini cupcakes cost  $2 \times 80 = 160$  dollars.  
6 dozen mini cheesecakes cost  $6 \times 55 = 330$  dollars.  
**204 + 160 + 330 is 694. The answer is 694.**

Figure E.1: Illustration of arithmetic inconsistency before scaling and corrected output after scaling for English reasoning using Llama 1B.

### Input

**Q:** Gunter versucht, die Marmeladebohnen in einem Glas zu zählen. Er fragt seine Freunde, wie viele in dem Glas sind. Einer sagt 80. Ein anderer sagt 20 mehr als die Hälfte des ersten. Ein Drittel sagt 25% mehr als der erste. Wie hoch ist ihre durchschnittliche Vermutung? **A:**

*(English translation: Gunter is trying to count the jam beans in a jar. He asks his friends how many are in the jar. One says 80. Another says 20, more than half the first. A third say 25% more than the first. What is their average guess?)*

### Output before scaling (wrong)

Der erste sagt 80. Ein anderer sagt 20 mehr als die Hälfte des ersten, also  $80 + 20 = 100$ . Ein Drittel sagt 25% mehr als der erste, also  $80 + (80 * 0,25) = 100$ . Die durchschnittliche Vermutung ist 100. Die Antwort ist 100.

*(English translation: The first person says 80. Another says 20 more than half of the first, so  $80 + 20 = 100$ . One third says 25% more than the first, so  $80 + (80 * 0.25) = 100$ . The average guess is 100. The answer is 100.)*

### Output after scaling (correct)

Der erste sagt 80. Ein anderer sagt 20 mehr als die Hälfte des ersten, also  $80 + 20 = 100$ . Ein Drittel sagt 25% mehr als der erste, also  $80 + (80 * 0,25) = 100 + 20 = 120$ . Die durchschnittliche Vermutung ist 80. Die Antwort ist 80.

*(English translation: The first person says 80. Another says 20 more than half of the first, so  $80 + 20 = 100$ . One third says 25% more than the first, so  $80 + (80 * 0.25) = 100 + 20 = 120$ . The average guess is 80. The answer is 80.)*

Figure E.2: Illustration of arithmetic inconsistency before scaling and corrected output after scaling for German reasoning using Llama 1B.

### Input

**Q:** मैरी ने एक चिकन भोजन का ऑर्डर दिया जिसकी कीमत 12 डॉलर है, 5 पैकेट दूध की कीमत 3 डॉलर है, 4 सेबों की कीमत 1.50 डॉलर है, और पिज्जा के कुछ डिब्बे हैं। मैरी ने कुल 50 डॉलर का भुगतान किया। अगर प्रत्येक बॉक्स की कीमत 8.50 डॉलर है तो मैरी ने पिज्जा के कितने डिब्बे मंगवाए? **A:**

*(English translation: Mary ordered a chicken meal that costs \$12, 5 packets of milk that cost \$3, 4 apples that cost \$1.50, and some boxes of pizza. Mary paid a total of \$50. How many boxes of pizza did Mary order if each box costs \$8.50?)*

### Output before scaling (wrong)

मैरी ने एक चिकन की कीमत 12 डॉलर के लिए 1 बॉक्स खरीदा। उसके पास  $50 - 12 = 38$  डॉलर बचे हैं। 5 पैकेट दूध की कीमत 3 डॉलर है, जिसकी कुल कीमत  $5 * 3 = 15$  डॉलर है। उसके पास  $38 - 15 = 23$  डॉलर बचे हैं। 4 सेबों की कीमत 1.50 डॉलर है, जिसकी कुल कीमत  $4 * 1.50 = 6$  डॉलर है। उसके पास  $23 - 6 = 17$  डॉलर बचे हैं। पिज्जा के कुछ डिब्बे की कीमत 8.50 डॉलर है, जिसकी कुल कीमत  $17 / 8.50 = 1.96$  डॉलर है। उसके पास 1.96 डॉलर बचे हैं।  $1.96 / 8.50 = 0.23$ । उसके पास 0.23 डिब्बे हैं। उत्तर 0.23 है।

*(English translation: Mary bought a box of chicken for \$12. She has  $50 - 12 = \$38$  left. 5 packets of milk cost \$3, for a total cost of  $5 * 3 = \$15$ . She has  $38 - 15 = \$23$  left. 4 apples cost \$1.50, for a total cost of  $4 * 1.50 = \$6$ . She has  $23 - 6 = \$17$  left. Some boxes of pizza cost \$8.50, for a total cost of  $17 / 8.50 = \$1.96$ . She has \$1.96 left.  $1.96 / 8.50 = 0.23$ . She has 0.23 boxes. The answer is 0.23.)*

### Output after scaling (correct)

मैरी ने एक चिकन की कीमत 12 डॉलर के लिए 1 बॉक्स खरीदा। उसके पास  $50 - 12 = 38$  डॉलर बचे हैं। 5 पैकेट दूध की कीमत 3 डॉलर है, जिसकी कुल कीमत  $5 * 3 = 15$  डॉलर है। उसके पास  $38 - 15 = 23$  डॉलर बचे हैं। 4 सेबों की कीमत 1.50 डॉलर है, जिसकी कुल कीमत  $4 * 1.50 = 6$  डॉलर है। उसके पास  $23 - 6 = 17$  डॉलर बचे हैं। पिज्जा के कुछ डिब्बे की कीमत 8.50 डॉलर है, जिसकी कुल कीमत  $17 / 8.50 = 2$  डिब्बे हैं। उत्तर 2 है।

*(English translation: Mary bought a box of chicken for \$12. She has  $50 - 12 = \$38$  left. 5 packets of milk cost \$3, for a total cost of  $5 * 3 = \$15$ . She has  $38 - 15 = \$23$  left. 4 apples cost \$1.50, for a total cost of  $4 * 1.50 = \$6$ . She has  $23 - 6 = \$17$  left. Some boxes of pizza cost \$8.50, for a total cost of  $17 / 8.50 = 2$  boxes. The answer is 2.)*

Figure E.3: Illustration of arithmetic inconsistency before scaling and corrected output after scaling for Hindi reasoning using Llama 1B.

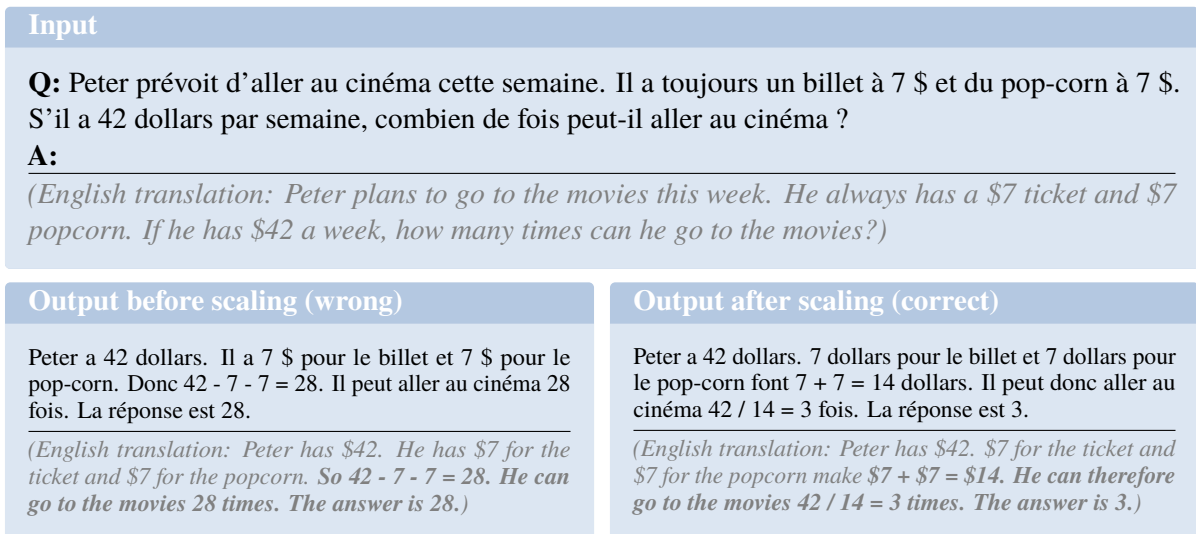


Figure E.4: Illustration of arithmetic inconsistency before scaling and corrected output after scaling for French reasoning using Llama 1B.

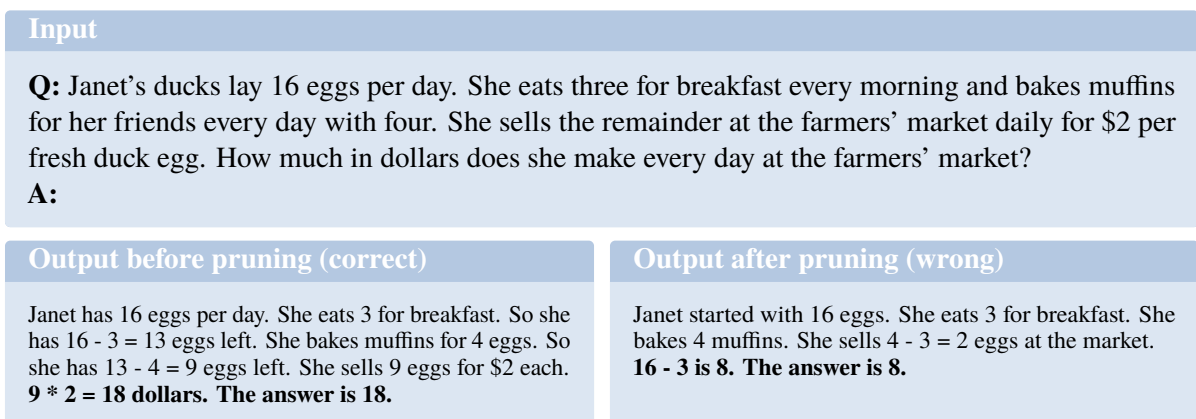


Figure E.5: Illustration of correct output before pruning and wrong output after pruning for English reasoning (top- $k=0.01$ ) using Llama 1B.

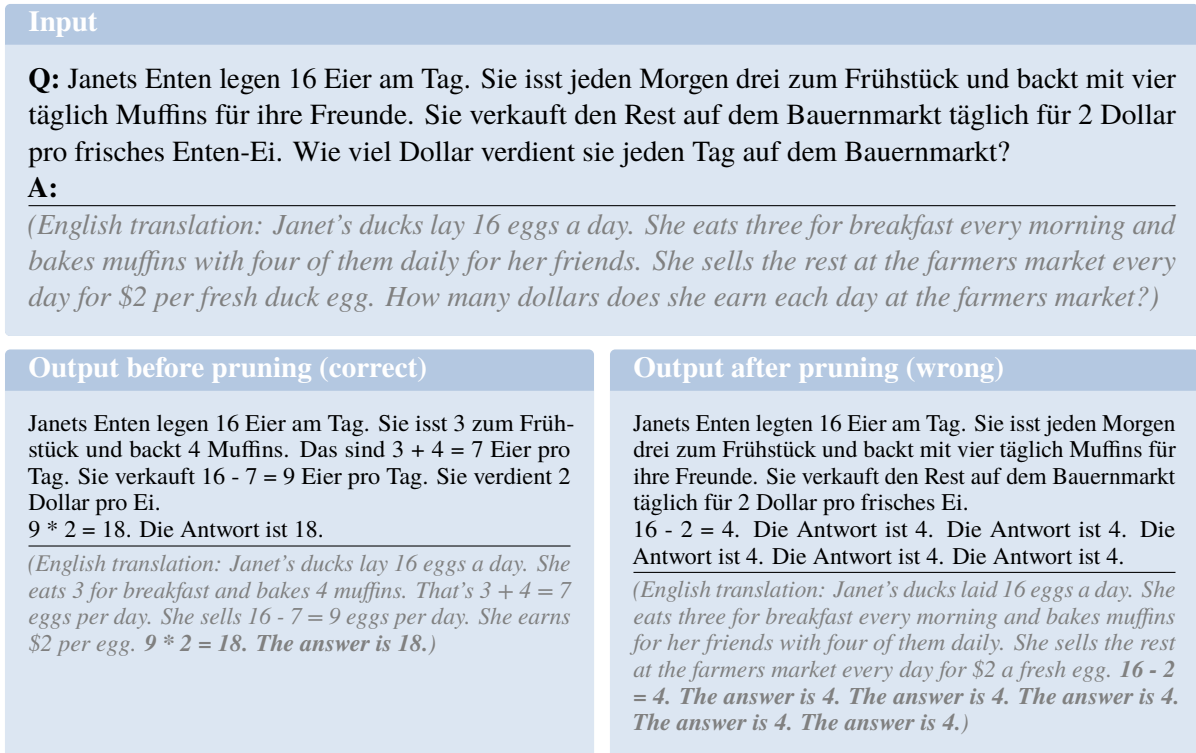


Figure E.6: Illustration of correct output before pruning and wrong output after pruning for German reasoning (top- $k=0.01$ ) using Llama 1B.

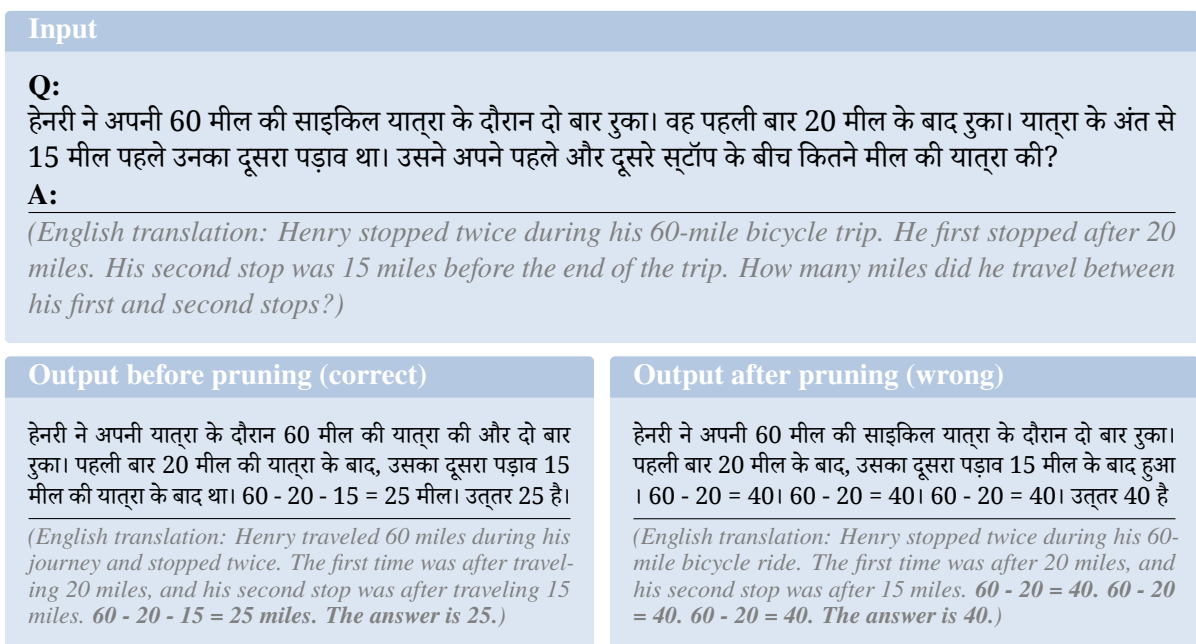


Figure E.7: Illustration of correct output before pruning and wrong output after pruning for Hindi reasoning (top- $k=0.01$ ) using Llama 1B.

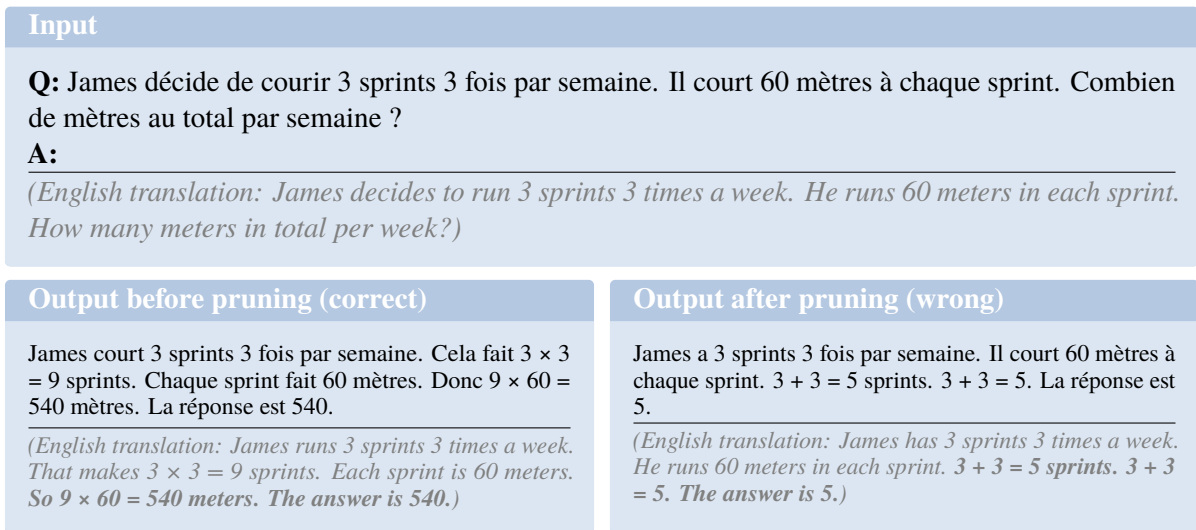


Figure E.8: Illustration of correct output before pruning and wrong output after pruning for French reasoning (top- $k=0.01$ ) using Llama 1B.

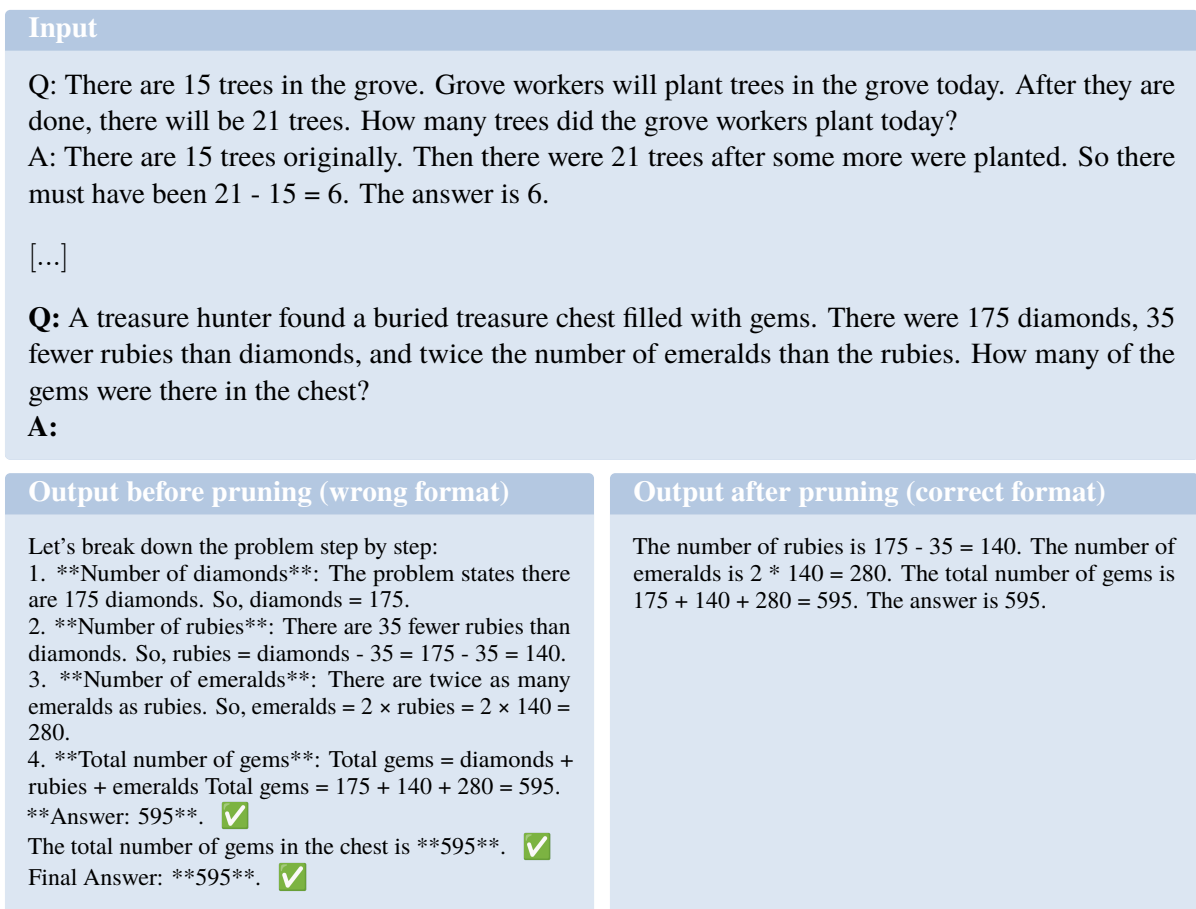


Figure E.9: Illustration of wrong output formatting and poor in-context learning before pruning and correct output after pruning for English reasoning (top- $k=1e-5$ ) using Qwen3 4B.

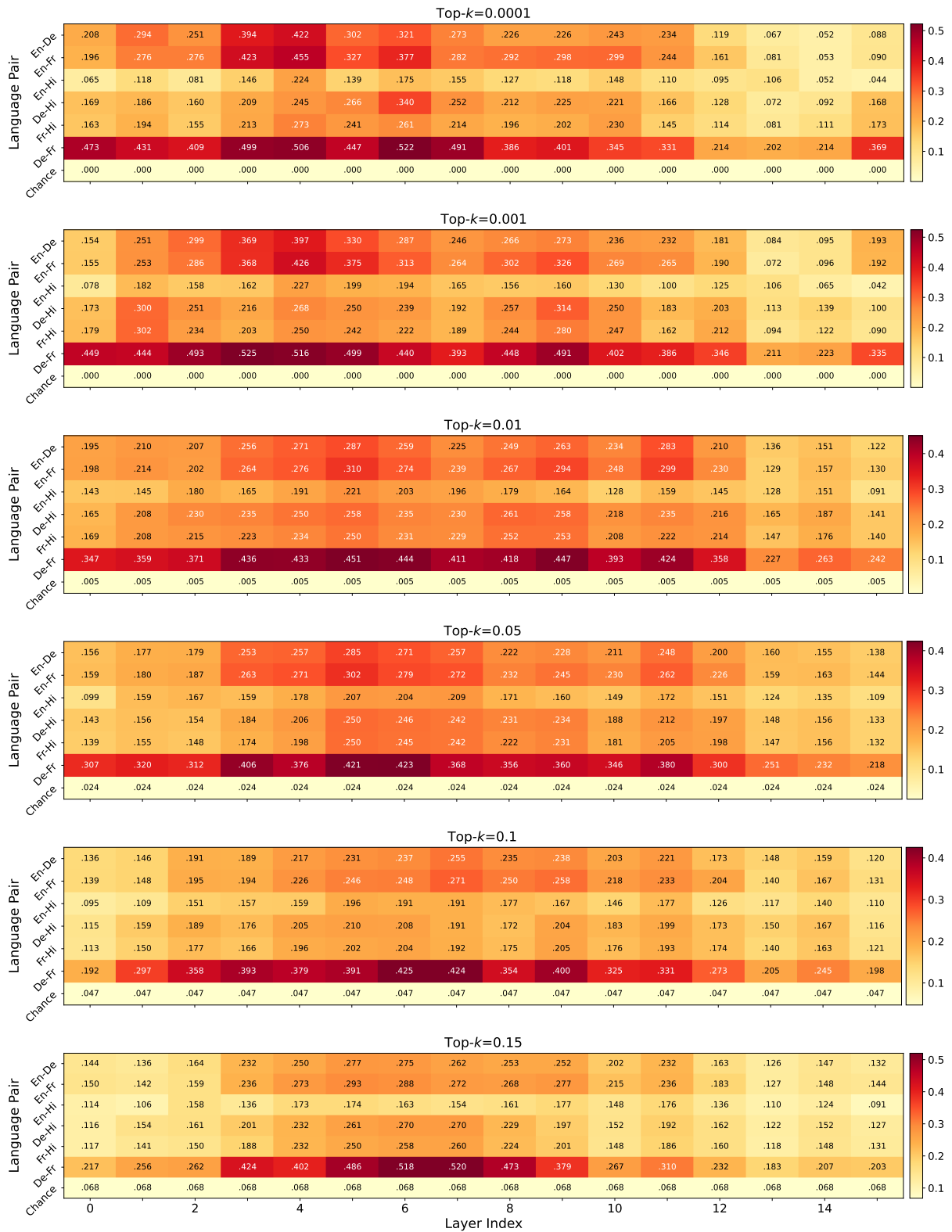


Figure F.1: Layer-wise Jaccard similarity of math-reasoning parameters across top- $k$  values in Llama 1B, computed for all language pairs (English (En), German (De), French (Fr), and Hindi (Hi)) and the random baseline (Chance) on RACE.

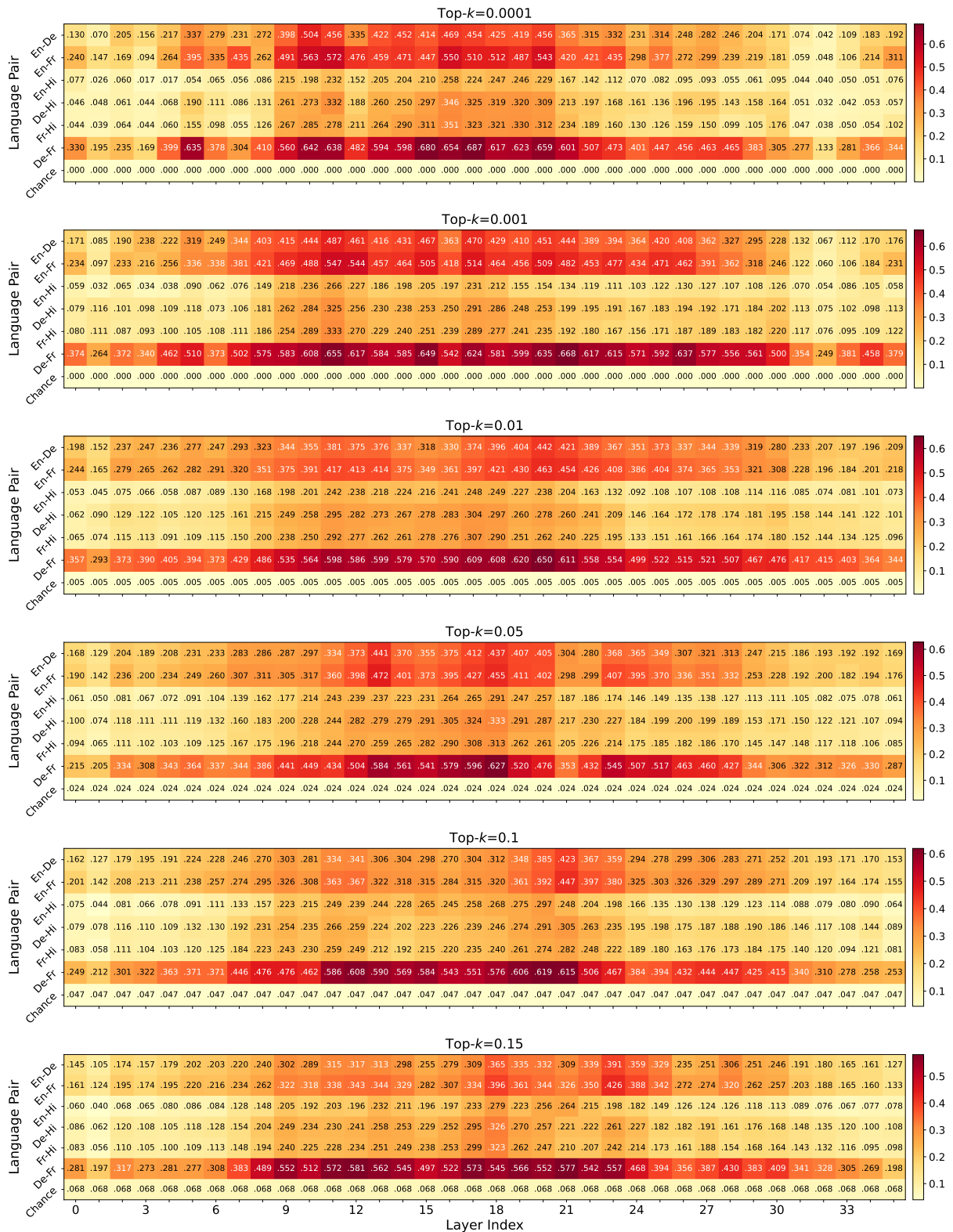


Figure F.2: Layer-wise Jaccard similarity of math-reasoning parameters across top- $k$  values in Qwen3 4B, computed for all language pairs (English (En), German (De), French (Fr), and Hindi (Hi)) and the random baseline (Chance) on RACE.

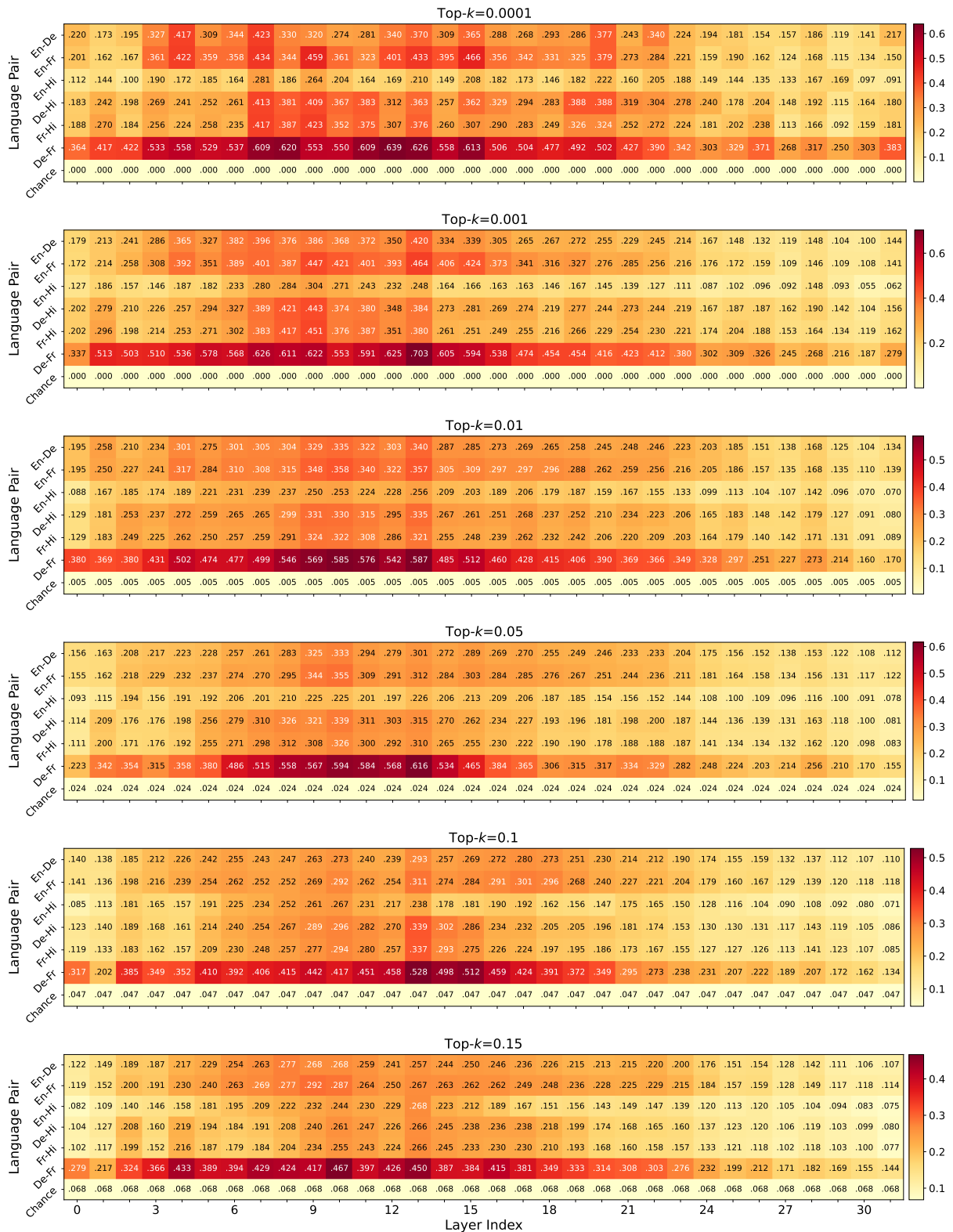


Figure F.3: Layer-wise Jaccard similarity of math-reasoning parameters across top- $k$  values in Llama 8B, computed for all language pairs (English (En), German (De), French (Fr), and Hindi (Hi)) and the random baseline (Chance) on RACE.

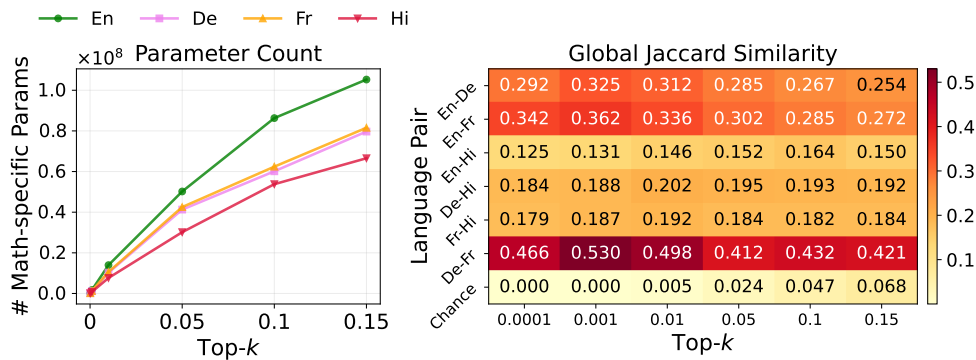


Figure F.4: Math-specific parameters in Qwen3 4B. **Left:** Number of isolated parameters per language (English (En), German (De), French (Fr), and Hindi (Hi)). **Right:** Jaccard similarity between language-specific parameter sets across all language pairs, compared to a random baseline using RACE as comparison dataset.

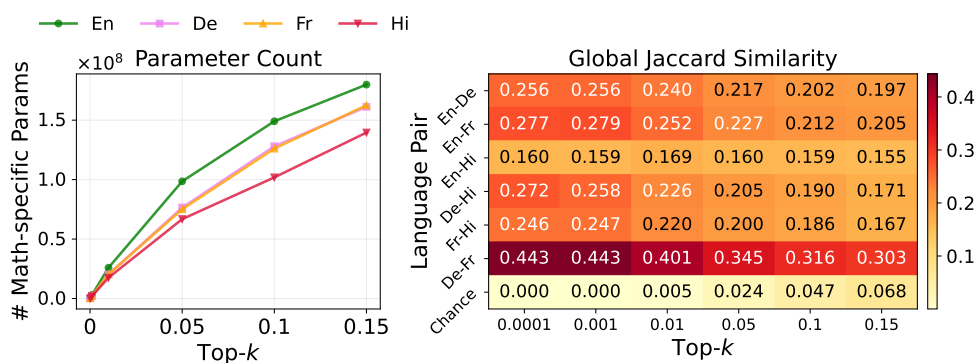


Figure F.5: Math-specific parameters in Llama 8B. **Left:** Number of isolated parameters per language (English (En), German (De), French (Fr), and Hindi (Hi)). **Right:** Jaccard similarity between language-specific parameter sets across all language pairs, compared to a random baseline using RACE as comparison dataset.

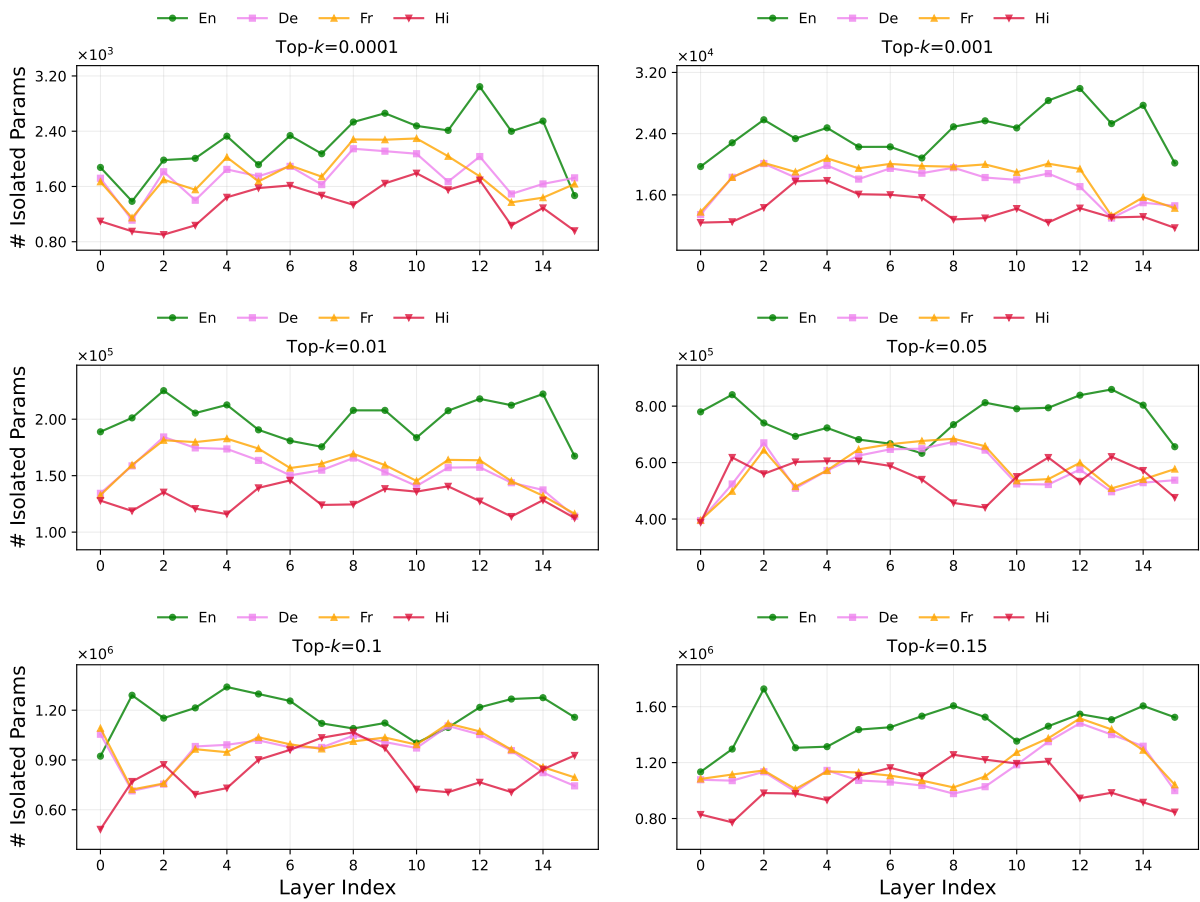


Figure F.6: Layer-wise distribution of math-reasoning parameters across top- $k$  values in Llama 1B, using RACE as comparison dataset.

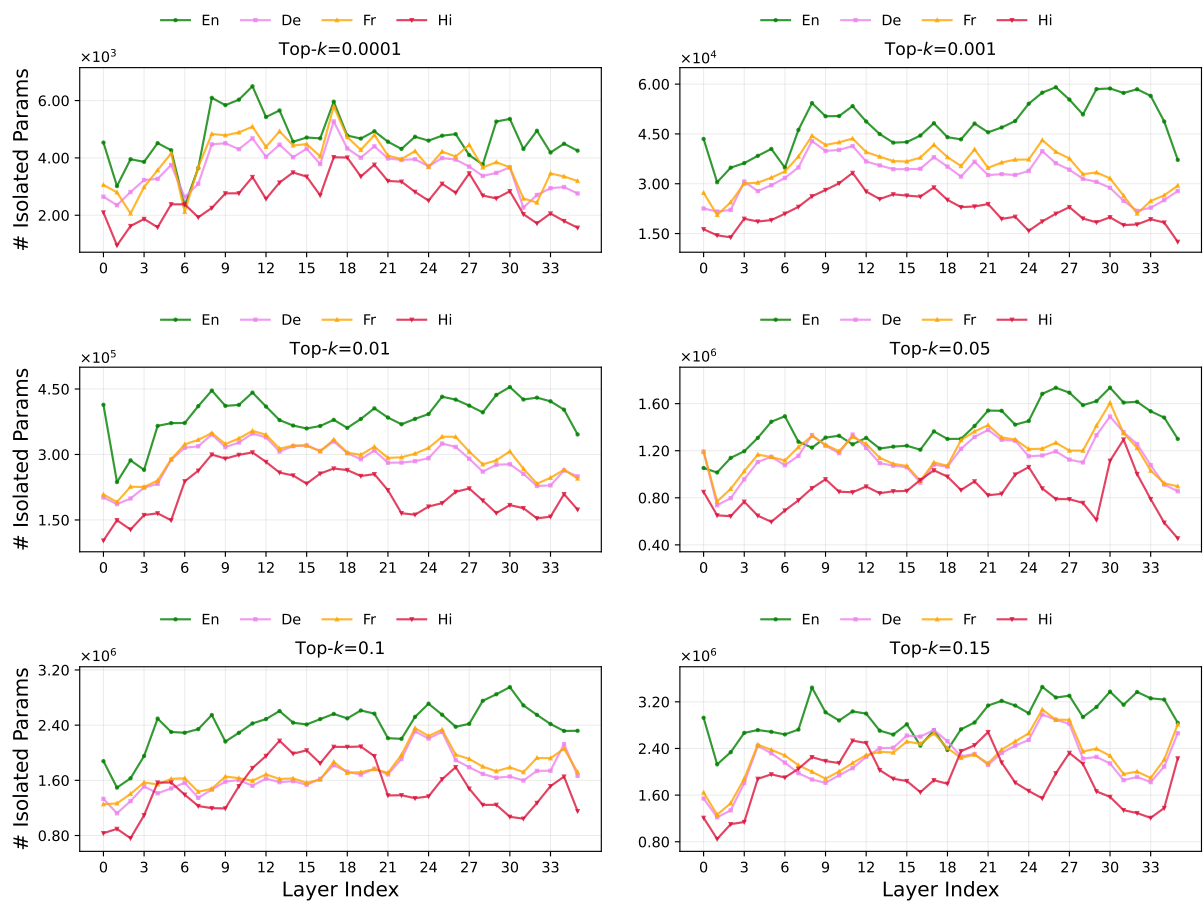


Figure F.7: Layer-wise distribution of math-reasoning parameters across top- $k$  values in Qwen3 4B, using RACE as comparison dataset.

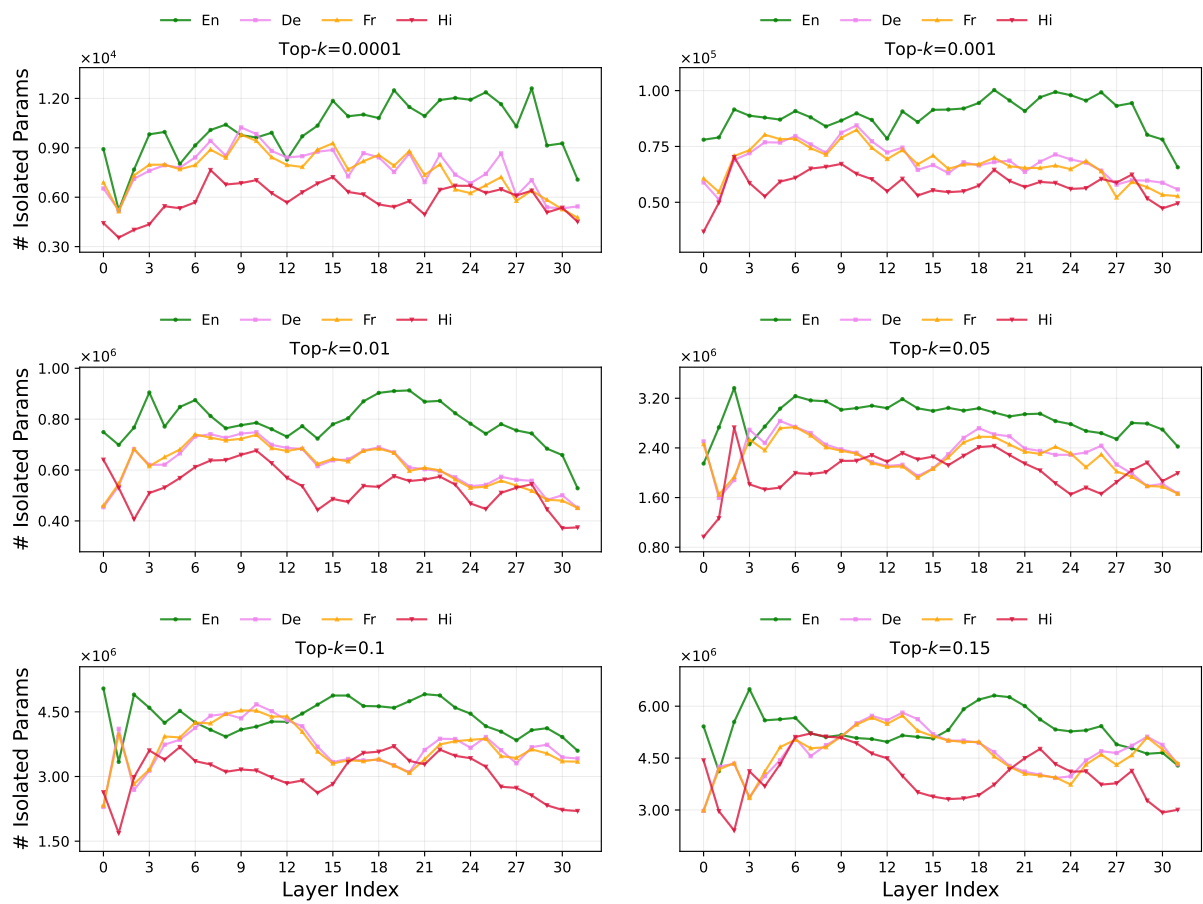


Figure F.8: Layer-wise distribution of math-reasoning parameters across top- $k$  values in Llama 8B, using RACE as comparison dataset.