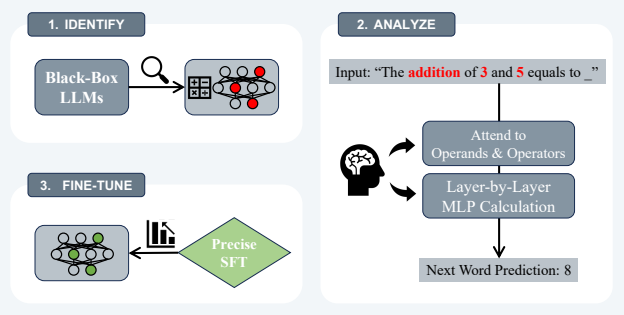

Figure 1: The pipeline involves three steps: 1) *identify* the key components attributed to arithmetic calculations in black-box LLMs, 2) *analyze* the working mechanism of the key components towards human-understandable explanations, 3) *fine-tune* the key components to precisely improve the mathematical capability of LLMs.

tervention (Pearl, 2009) on the transformer attention heads and multi-layer perceptrons (MLPs) to observe their effects on the predicted logits[1]. Our findings reveal that only a small percentage ($< 5\%$) of the attention heads and the MLPs after these heads significantly impact the model's performance. Namely, LLMs frequently involve these attention heads and the subsequent MLPs when completing the calculations. Subsequently, we knock out these frequently-involved heads/MLPs to validate their faithfulness. We find that the model performance decreases sharply when those pivotal heads/MLPs are knocked out, resulting in a decrease of around 70% in accuracy.

To interpret the working mechanism of identified heads/MLPs towards human-understandable explanations, we gain a deeper analysis of their operational "behaviors". Specifically, we investigate the attention patterns of the crucial heads, and find that these attention heads exhibit a strong focus on the tokens representing operands and operators within mathematical sentences, demonstrating a relative insensitivity to other non-relevant tokens. For the analysis of MLPs, we compare the correlations between the embeddings of MLPs' input/output and the embeddings of number tokens (*i.e.*, operands and answers). It reveals that the MLPs, guided by these number-attended heads, take operands as input, and mirror the attributes of tokens corresponding to correct answers more closely. These observations lead us to hypothesize that *LLMs may initially employ a set of heads to pinpoint arithmetic operands from text, subsequently engaging MLPs to work out the answers.* Additionally, the observed behaviors of these heads/MLPs exhibit a high degree of transferability, analogous to adversarial examples being *transferable across models* (Szegedy et al., 2014). Namely, the key heads/MLPs identified on one dataset are also effective for other datasets. For instance, their impact is noticeable on the publicly available math datasets (*e.g.*, SVAMP (Patel et al., 2021)), as well as varied data formats involving multi-digit integers, rational numbers, etc. This empirical observation underscores the crucial role of key heads/MLPs in mathematical calculations.

In addition to uncovering the internal mechanisms, we have devised an effective strategy that involves targeted fine-tuning of the specific attention heads and MLPs closely tied to mathematical computations, thereby enhancing the model's mathematical prowess. The experimental results are compelling: with fine-tuning as few as 32 attention heads (with a total of 1024 heads), we observe a remarkable improvement in the model's mathematical capabilities. This precise tuning methodology not only matches but can surpass the enhancements achieved through full-model fine-tuning. Moreover, this fine-grained strategy of adjustment has a distinct advantage—it leaves most of the model's parameters unchanged, avoiding the performance trade-offs in non-mathematical domains commonly observed with full-model fine-tuning.

In summary, this work aims to delve into the inner mechanism of LLMs through mathematical calculation tasks, along the pipeline of "identify-analyze-finetune" shown in Figure 1. Our findings reveal a sparsity in the attention heads of LLMs, with less than 5% of heads exhibiting close correlations. These heads particularly attend to the operands and operators, while the subsequent MLPs gradually deduce the correct answers. The discovered mechanism shows strong cross-dataset transferability and inspires us to precisely finetune the calculation-related heads/MLPs for better mathematical capability. We empirically find that precise tuning brings in much less impact on non-mathematical tasks when improving the targeted ability of LLMs.

2. Related Works

Interpretability Methods. Interpreting the inner mechanism of large language models (LLMs) has become increasingly urgent in recent years (Madsen et al., 2022; Rauker et al., 2023), especially when LLMs are applied in high-stakes decision-making domains such as healthcare, criminal justice, and finance (Obermeyer et al., 2019; Rudin, 2019; Bender et al., 2021). Vig et al. (2020) adapted the approach of *causal mediation analysis* (CMA) (Pearl, 2001) for interpreting the deep language models, and it has been applied for various tasks, such as subject-verb agreement (Finlayson et al., 2021), natural language inference (Geiger et al., 2021), retention of factual associations (Meng et al., 2022; Geva et al., 2023). Furthermore, *path patching* extends the concept of CMA by measuring how a treatment

[1]Here, doing a hard intervention is equal to replacing the value of attention heads and MLPs, while performing a soft intervention means modifying the modules for calculating the attention and MLP values (Pearl, 2009).