reduction in training time, attributed to the substantially fewer parameter adjustments required (less than $1\%$). It results in a time reduction of at least threefold on LLaMA2-7B and LLaMA2-13B. Overall, Precise SFT offers an effective direction for boosting mathematical abilities for LLMs.

**Ablative studies.** The key issue with Precise SFT lies in determining the quantity and specific set of components to adjust. To demonstrate this, we experimented with varying numbers of heads and MLPs, with the results laid out in Table 2. We discovered that fine-tuning 32 heads yields the best average improvement across different numbers of involved heads. We also compared experiments with the introduction of MLPs. We observed that as more MLPs are added, the mathematical capability improves by $2.1\%$, but the general performance will decrease by $1.5\%$ (results in Appendix G). Overall, the top-3 MLPs yielded the best comprehensive results. However, even the introduction of a single MLP can reduce computational efficiency by $15\%$. How to more precisely fine-tune MLPs will be explored in our future work.

**More discussions.** The above results underscore the potential of employing interpretability tools to analyze the inner mechanism of LLMs and to enhance their specific capabilities. However, there are several areas that require deeper investigation: (i) Our primary experiments and discussions center around the LLaMA2 series. The results presented in Appendix C demonstrate the potential for generalization across different LLMs, such as Mistral-7B (Jiang et al., 2023). For more rigorous considerations, it's necessary to perform specific adaptations on a broader range of LLMs. (ii) This work mainly focuses on interpreting the fundamental ability of "arithmetic calculation", since it's universally shared across various levels of complexity for mathematical problems. The results in Appendix E reveal that solving the math word problems requires a synergy of multiple skills including "text comprehension" and "arithmetic calculation", which is aligned with the findings in recent research (Opedal et al., 2024). It's imperative for continued research to investigate more complex mathematical problems. (iii) The potential of generalizing to more complex mathematical tasks like exponentiation (*e.g.*, "{A} to the power of {B} equals _") has been validated in Appendix D. An intriguing research direction would be to investigate the shared and distinct mechanisms across various mathematical tasks.

## 6. Conclusion

In this study, we have identified, analyzed, and fine-tuned the internal components responsible for the mathematical calculation capability of LLMs. The language models frequently involve sparse heads to particularly attend to operands and operators, and subsequent MLPs to work out answers. We apply the precise tuning on the calculation-related heads/MLPs for better mathematical capabilities, with less impact on non-mathematical tasks compared with tuning all parameters. These findings contribute to a better understanding of the inner mechanism of LLMs.

## Acknowledgements

This work was supported in part by NSFC No. 62222117. YGZ and YMC were supported in part by NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant: N_HKBU214/21; in part by RGC Senior Research Fellow Scheme under Grant: SRFS2324-2S02.

## Impact Statement

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

## References

Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J. L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Barak, B., Edelman, B. L., Goel, S., Kakade, S., Malach, E., and Zhang, C. Hidden progress in deep learning: Sgd learns parities near the computational limit. *arXiv preprint arXiv:2207.08799*, 2022.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Viégas, F. B., and Wattenberg, M. An interpretability illusion for BERT. *CoRR*, abs/2104.07143, 2021.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.