

Interpreting and Improving Large Language Models in Arithmetic Calculation

Wei Zhang^{*1,2} Chaoqun Wan² Yonggang Zhang^{†3} Yiu-ming Cheung³ Xinmei Tian^{1,4} Xu Shen^{†2}
Jieping Ye²

Abstract

Large language models (LLMs) have demonstrated remarkable potential across numerous applications and have shown an emergent ability to tackle complex reasoning tasks, such as mathematical computations. However, even for the simplest arithmetic calculations, the intrinsic mechanisms behind LLMs remain mysterious, making it challenging to ensure reliability. In this work, we delve into uncovering a specific mechanism by which LLMs execute calculations. Through comprehensive experiments, we find that LLMs frequently involve a small fraction (< 5%) of attention heads, which play a pivotal role in focusing on operands and operators during calculation processes. Subsequently, the information from these operands is processed through multi-layer perceptrons (MLPs), progressively leading to the final solution. These pivotal heads/MLPs, though identified on a specific dataset, exhibit transferability across different datasets and even distinct tasks. This insight prompted us to investigate the potential benefits of selectively fine-tuning these essential heads/MLPs to boost the LLMs' computational performance. We empirically find that such precise tuning can yield notable enhancements on mathematical prowess, without compromising the performance on non-mathematical tasks. Our work serves as a preliminary exploration into the arithmetic calculation abilities inherent in LLMs, laying a solid foundation to reveal more intricate mathematical tasks.

^{*}This work was done when the author was visiting Alibaba Cloud as a research intern. [†]University of Science and Technology of China ²Alibaba Cloud ³Hong Kong Baptist University ⁴Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. Correspondence to: Xu Shen[†] <shenxu.sx@alibaba-inc.com>, Yonggang Zhang[†] <csygzhang@comp.hkbu.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Large language models (LLMs) have experienced rapid advancements and shown impressive language understanding capabilities (Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2022). Notably, LLMs exhibit emergent abilities (Wei et al., 2022b) that enable them to solve intricate reasoning tasks akin to humans, such as mathematical computations (Frieder et al., 2023; Jie et al., 2022), chain-of-thought reasoning (Wei et al., 2022c; Kojima et al., 2022), few-shot prompting (Brown et al., 2020; Alayrac et al., 2022), etc. Despite these impressive characteristics, the complex inner processes governing LLMs' functionality have yet to be fully illuminated, due to the complex and intricate non-linear interactions within densely-connected layers. Comprehending these underlying mechanisms could contribute to predicting how the LLMs behave beyond their training data (Mu & Andreas, 2020), gaining insights into the emergence of certain behaviors (Nanda & Lieberman, 2022; Barak et al., 2022; Wei et al., 2022a), as well as identifying and rectifying errors present in the specific models (Hernandez et al., 2021; Vig et al., 2020).

In this work, we take the first attempt to interpret the inner process of LLMs through the lens of mathematical computation problems, which are conducted on publicly available LLMs (*e.g.*, LLaMA2 series (Touvron et al., 2023b)). Unlike typical language comprehension tasks, mathematical computation tasks involve concise problem statements with definitive correct answers, requiring a process of reasoning and calculation rather than direct copying to derive the solutions. These characteristics enable us to gain insights into the models' reasoning capabilities without interference from unrelated factors. Specifically, we focus on tasks involving the arithmetic calculation with two operands, *i.e.*, addition, subtraction, multiplication, and division, which are fundamentals of mathematical computation. To this end, we create datasets of various types of sentences that involve the calculation logic, such as “The addition of 3 and 5 equals to _” in Figure 1. The LLMs could provide answers with high confidence scores of over 80% on average.

To unveil how these models correctly complete the task (*e.g.*, “3 + 5 = 8”), we begin by identifying the task-related internal components in LLMs. We do a hard in-