上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

It Ain't That Bad: Understanding the Mysterious Performance Drop in OOD Generalization for Generative Transformer Models

Xingcheng Xu, Zihao Pan, Haipeng Zhang, Yanqing Yang

xingcheng.xu18@gmail.com, {panzh,zhanghp}@shanghaitech.edu.cn, yanqingyang@fudan.edu.cn

1. Abstract

Large language models (LLMs) have achieved remarkable proficiency on solving diverse problems. However, their generalization ability is not always satisfying and the generalization problem is common for generative transformer models in general. Researchers take basic mathematical tasks like n-digit addition or multiplication as important perspectives for investigating their generalization behaviors. It is observed that when training models on n-digit operations (e.g., additions) in which both input operands are n-digit in length, models generalize successfully on unseen n-digit inputs (in-distribution (ID) generalization), but fail miserably on longer, unseen cases (out-of-distribution (OOD) generalization). We bring this unexplained performance drop into attention and ask whether there is systematic OOD generalization. Towards understanding LLMs, we train various smaller language models which may share the same underlying mechanism. We discover that the strong ID generalization stems from structured representations, while behind the unsatisfying OOD performance, the models still exhibit clear learned algebraic structures. Specifically, these models map unseen OOD inputs to outputs with learned equivalence relations in the ID domain, which we call the *equivalence generalization*. These findings deepen our knowledge regarding the generalizability of generative models including LLMs, and provide insights into potential avenues for improvement.

2. Introduction

LLMs such as GPT-4, Claude, Gemini and Llama have exhibited remarkable advancements across diverse domains, prominently in natural language processing (NLP). These models have demonstrated exceptional versatility, tackling tasks ranging from natural language challenges to code translation and mathematical reasoning. However, despite these accomplishments, the generalization ability of LLMs is not fully understood and is sometimes inadequate, particularly in areas like natural language understanding and mathematical reasoning. The black-box nature of these models has led researchers to explore basic mathematical tasks as a means to gain insights into their generalization behaviors. Observations have revealed a distinct difference between ID generalization, where models perform well on familiar inputs, and OOD generalization, where they struggle with longer, unseen cases. This study aims to bridge the generalization gap by investigating the mechanistic perspectives behind these behaviors, using smaller models to uncover insights that could apply to larger LLMs.

Main Contributions:

- **Showcasing the power of mechanistic empirical evaluation for LLM generalization:** We train small generative language models (e.g., NanoGPT, MinGPT) on arithmetic tasks to directly investigate ID vs. OOD generalization, rather than resorting to workarounds. As a result, our approach provides macroscopic insights.
- **Discovering learned structure for OOD generalization:** The discernible algebraic structure and the equivalence generalization would hopefully guide robust essential solutions for strong OOD generalization.
- **Understanding the role of representations in generalization:** We show that representation learning enables strong ID performance, while unanticipated extension of representations to OOD inputs leads to systematic errors.

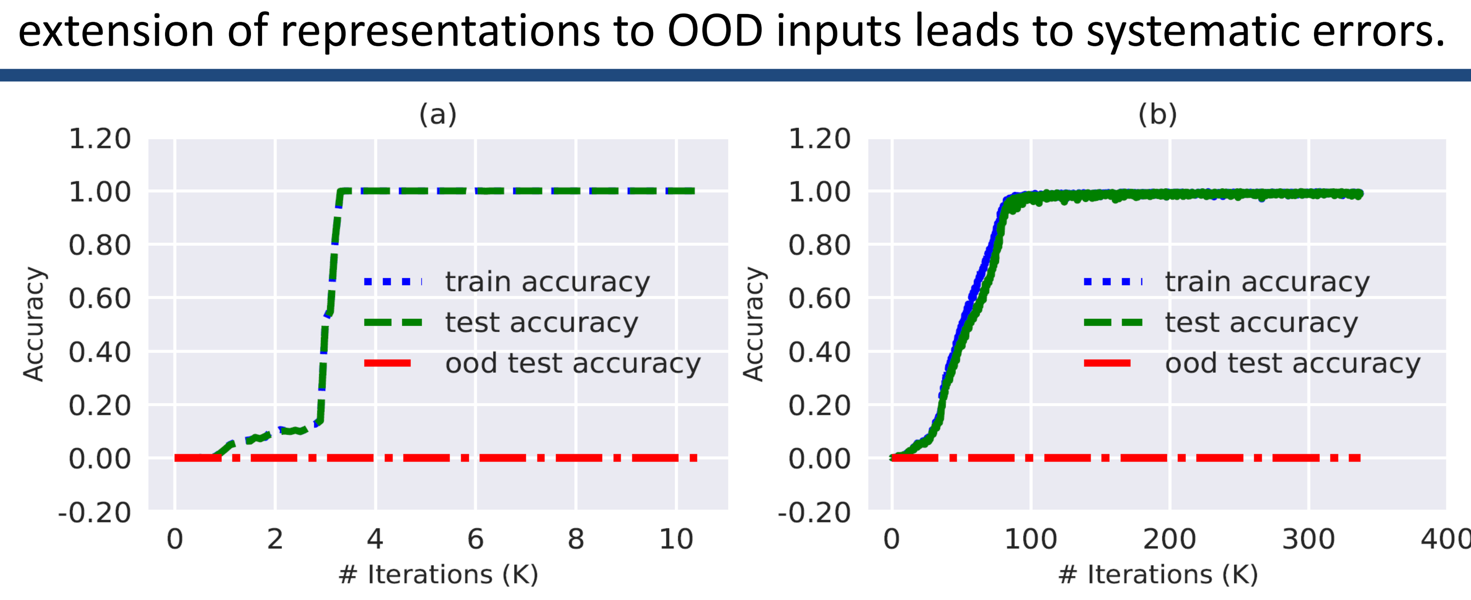

Figure 1 Training curves in addition and multiplication operations

Operands	Output Result	Correct Result
349 + 705	1,054	1,054
1,349 + 2,705	1,054	4,054
128 × 256	32,768	32,768
3,128 × 4,256	32,768	13,312,768

Table 1 Examples on models' outputs for addition and multiplication

3. Experiments and Results

- **Algebraic Structure**

The models map unseen OOD inputs to outputs with equivalence relations in the ID domain. Equivalence Classes:

$$[(a,b)]_p := \{(x,y) \in \mathbb{N}^2 \mid x \equiv a \pmod p,\ y \equiv b \pmod p\} \quad \mathbf{Z}_p^2 = \mathbf{Z}_p \times \mathbf{Z}_p = \{[(a,b)]_p \mid (a,b) \in \mathbb{N}^2\} \quad \mathbf{Z}_p = \mathbb{Z}/p\mathbb{Z}$$

- **Probability Structure**

We introduce perturbations to the thousands digit of a and b, enabling us to compare the variations in probabilities before and after the perturbations occur.

- **Representation Structure**

The representations gradually transition from disorderly to structured throughout the learning process, and enable powerful ID generalization, but the extrapolation to OOD inputs gives rise to systematic, rather than random, errors.

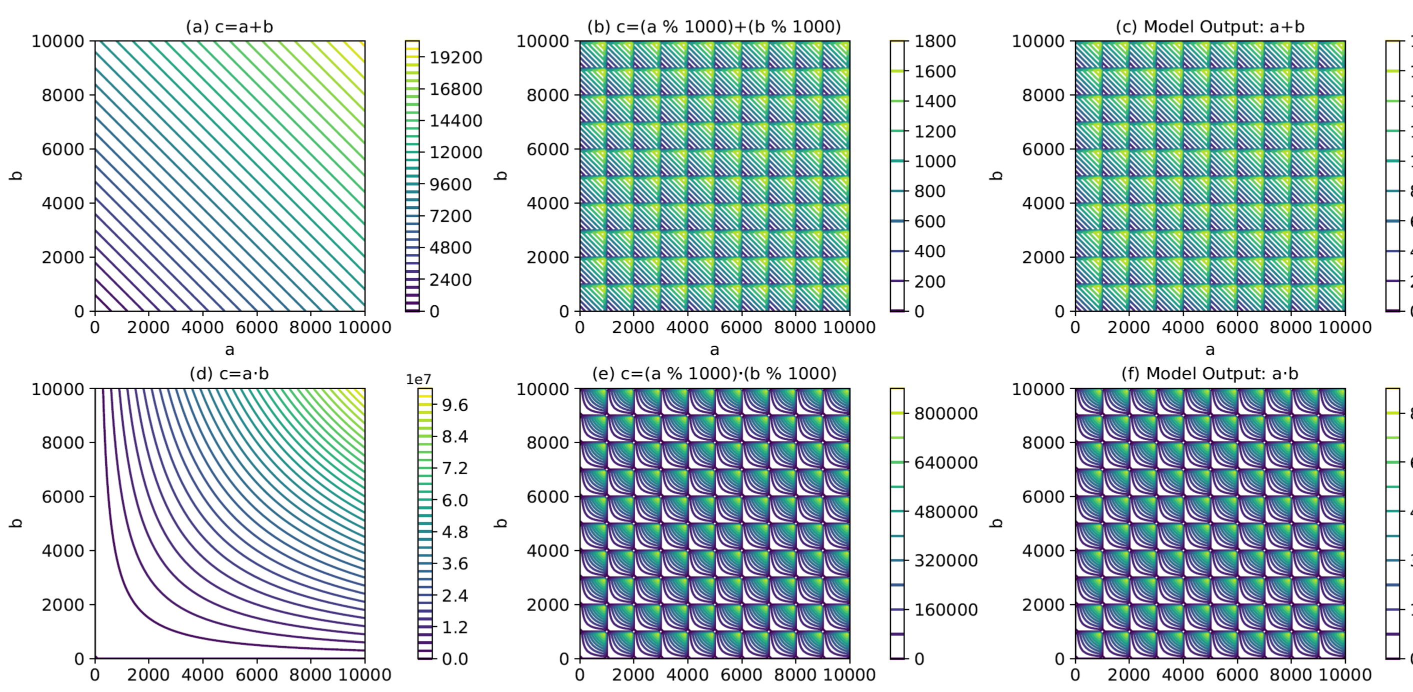

Figure 2 Contour plots for addition and multiplication operations

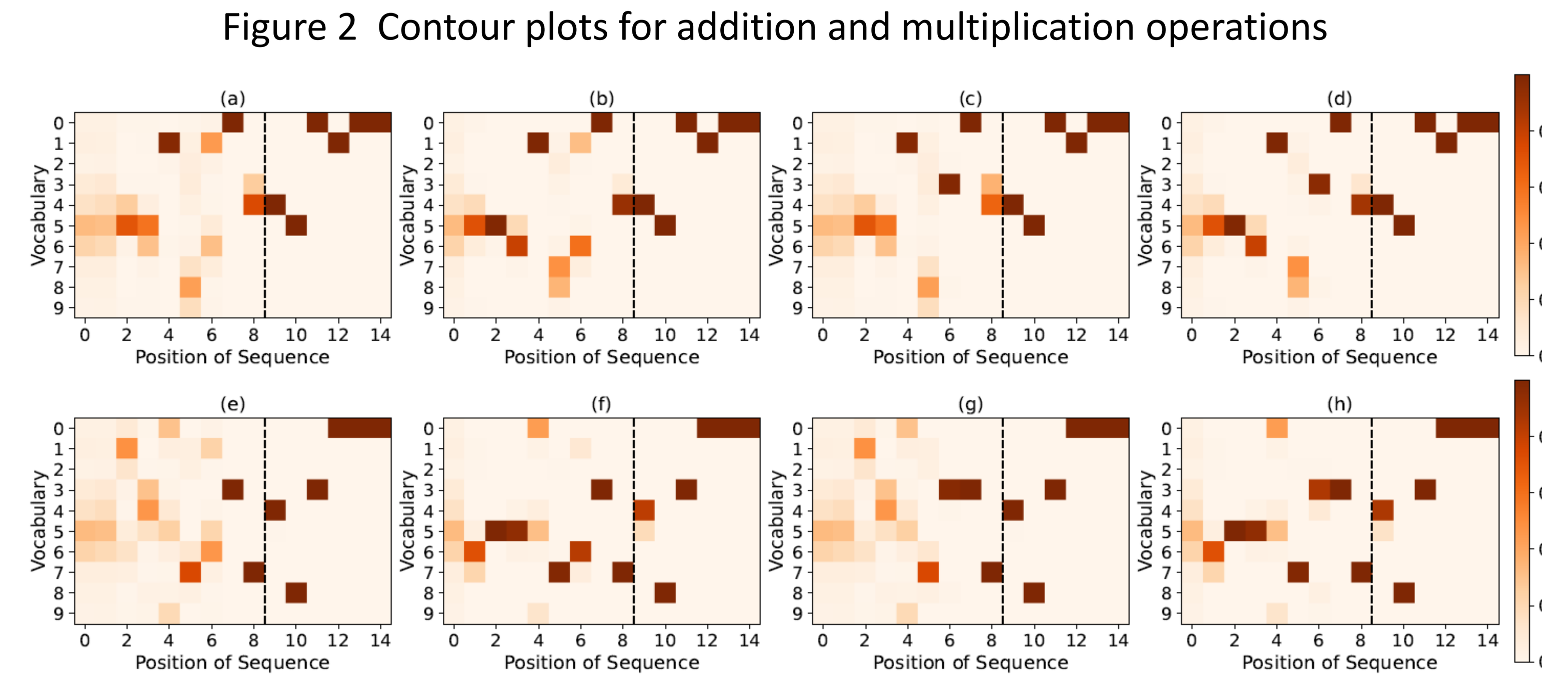

Figure 3 The probability distribution of each digit of the sequence in addition

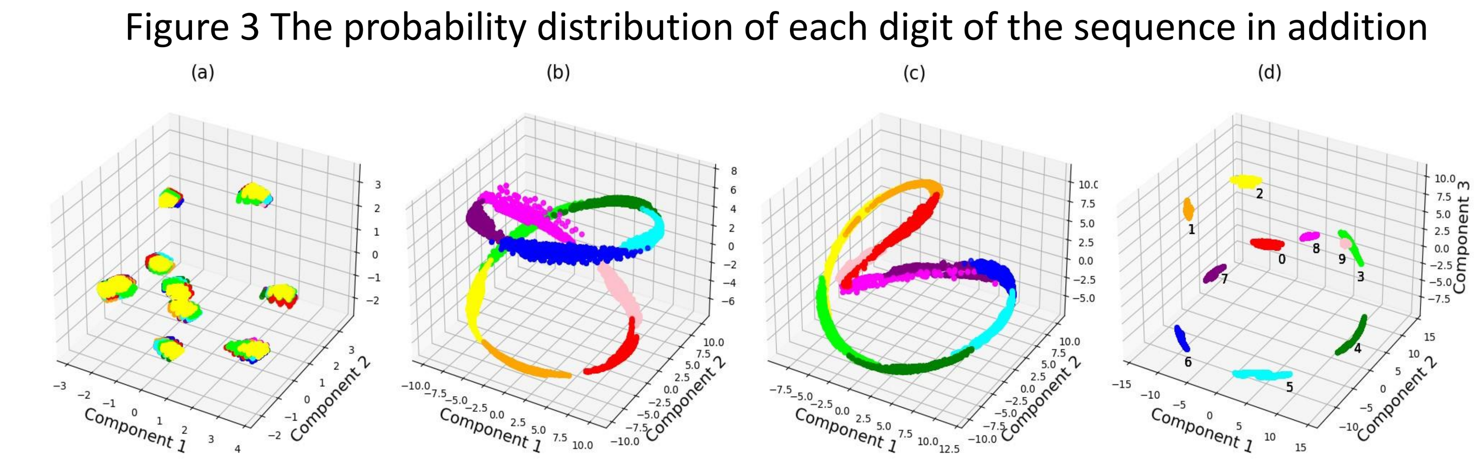

Figure 4 3D representation structure of the first three principal components in the addition operation. The learning process: random initialization→ well-trained model.

4. Conclusion

We investigate the length generalization problem in arithmetic tasks for generative language models. We perform mechanistic analysis on smaller models and reveal that these models have strong generalization within the trained distribution. However, our investigation also uncovers an underlying algebraic structure that contributes to the models' unsatisfactory performance on OOD inputs. The models attempt to map OOD inputs using equivalence relations within the ID domain (we call "*equivalence generalization*"), leading to errors and a lack of robustness in OOD scenarios. The representation plays a crucial role in enabling both ID and OOD generalization. The observation that length generalization ability does not vary with model scale, helps us extend our conclusion to LLMs.

Despite challenges in OOD generalization, our findings suggest that these models hold valuable information for improved generalization. However, due to the inherent subjectivity of natural language, much more efforts are needed to establish equivalence in NLP tasks for LLMs. In addition, the finding of equivalence generalization may serve as helpful prior knowledge, guiding the training process of LLMs regarding generalizability. For example, we may stop training once these equivalence classes are formed, reducing the extensive data needed for generalizability. Besides, in domain adaptation, people often finetune existing models, to adapt to OOD data and similarity metrics of equivalence classes may facilitate this process.

Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)

IJCAI
International Joint Conferences on Artificial Intelligence Organization