MUSTARD: Mastering Uniform Synthesis of Theorem and Proof Data

Yinya Huang[1,6], Xiaohan Lin[2], Zhengying Liu[3], Qingxing Cao[2], Huajian Xin[2], Haiming Wang[2], Zhenguo Li[3], Linqi Song[1,6], Xiaodan Liang[2,4,5]

[1]City University of Hong Kong [2]Shenzhen Campus of Sun Yat-sen University

[3]Huawei Noah's Ark Lab [4]DarkMatter AI Research [5]MBZUAI [6]CityUSRI

Contact: yinya.huang@hotmail.com

Intermediate Step Validation

- Correct intermediate steps are crucial for LLMs to perform complex reasoning.
- High-quality step-wise annotations are hard to obtain.
- Current data synthesis trades off between corectness and reusability.

✓ MUSTARD synthesizes **correct, scalable, and reusable** mathematical data by combining the advantages of LLMs in verbalization and formal theorem provers in rigorous data validation.

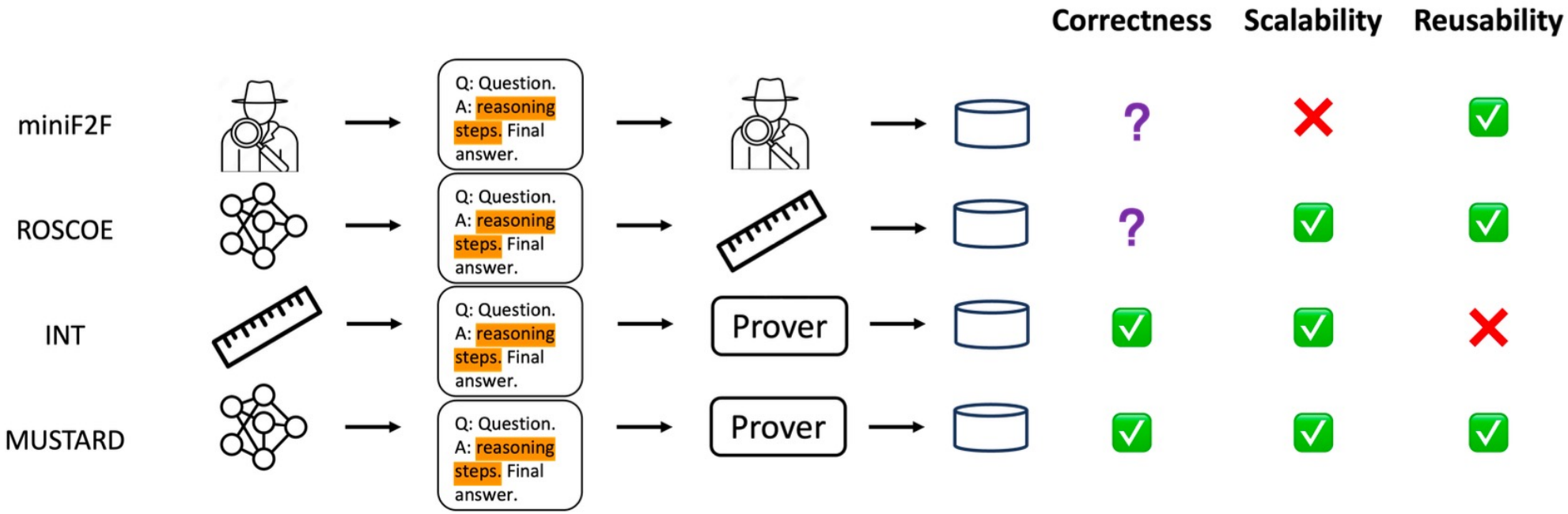
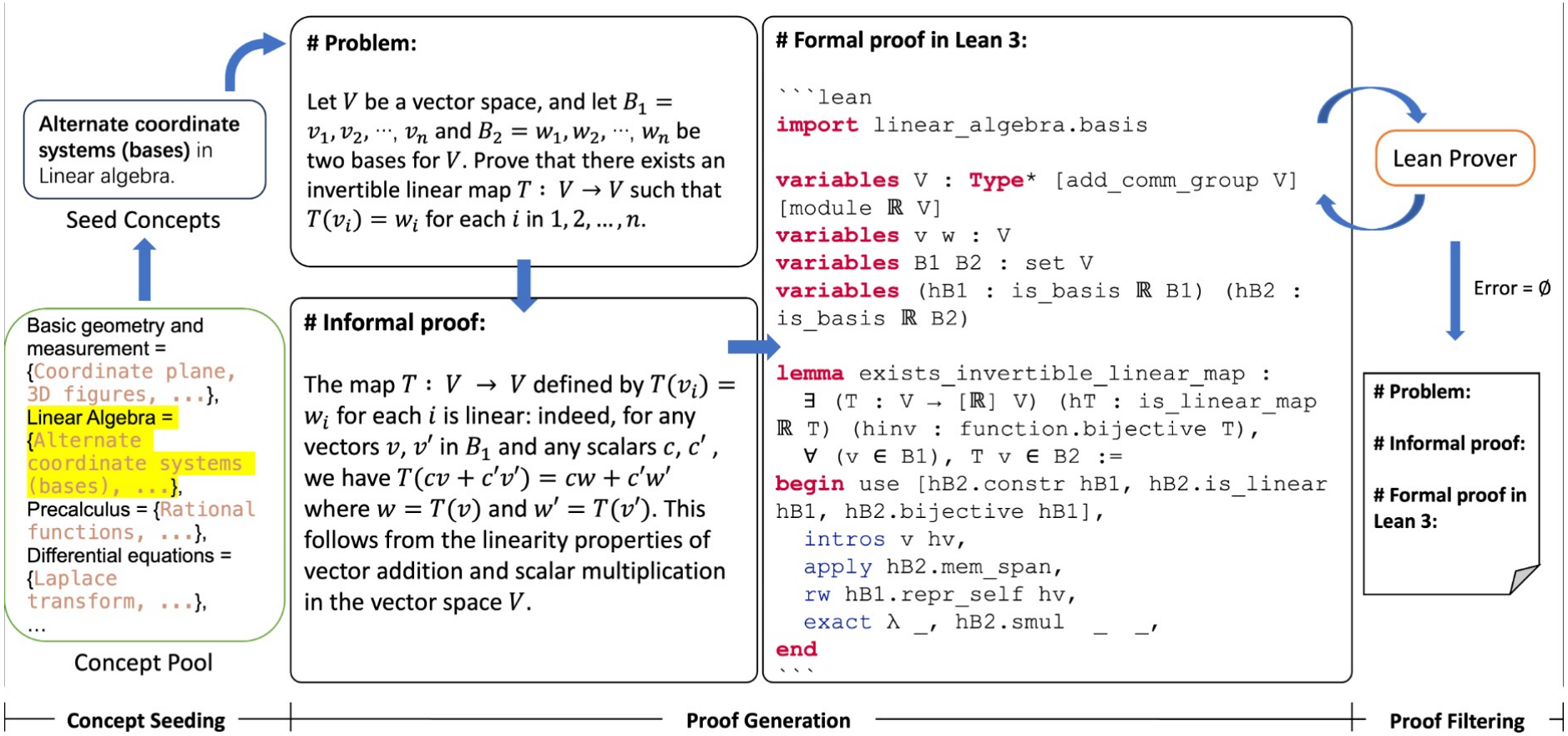


Figure 1: A comparison of methods of synthesizing and validating intermediate reasoning steps.



- Sampling math concepts from concept pool
- Prompting an LLM to generate problem, informal solution, and formal solution in Lean 3.
- Collecting data by Lean Prover validation.

→ **MUSTARDSAUCE** (MUSTARD resource)

MUSTARDSAUCE: Data Difficulty

- ✓ 5,866 MWP & ATP data points w/ {informal, formal} {statement, solution}.
- ✓ Proof length increases over educational levels.
- ✓ The most challenging problems require around 30 proof steps w/ 20 Lean tactics.

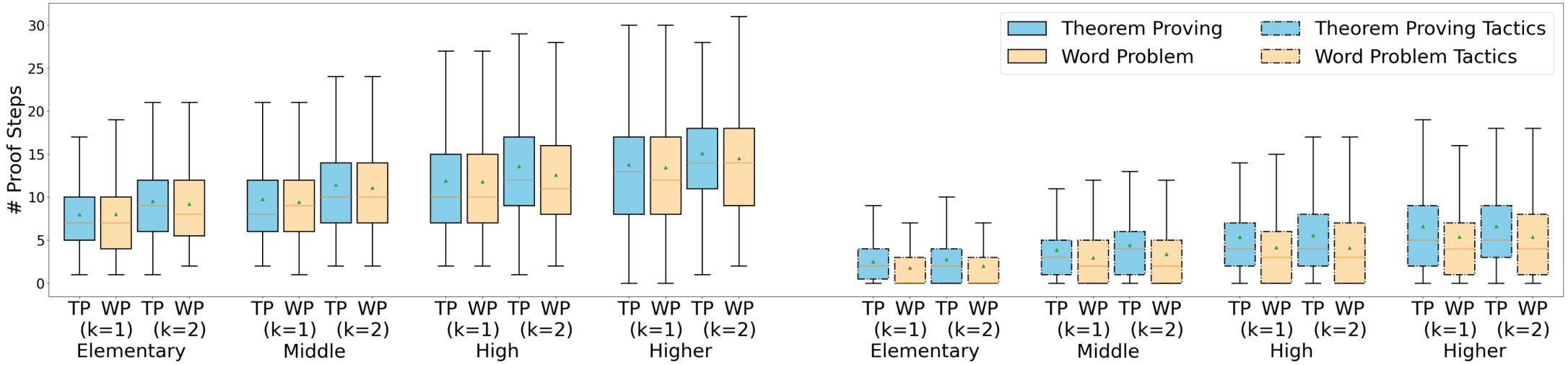


Figure 4: Distributions of formal proof lengths.

MUSTARDSAUCE: Data Quality

- We randomly select 200 data points from the generated data, 100 of which pass the Lean Prover (Group Valid) and 100 of which do not (Group Invalid).
- 6 inspection dimensions including factuality check and consistency check on informal/formal statement/proof.

✓ **(D1)** shows significance. → High quality math questions.

✓ **(D4)** and **(D6)** show significant differences between Group Valid & Invalid. → High quality autoformalization.

Table 3: Inspection dimensions and requirements in human evaluation. IS: Informal Statement. IP: Informal Proof. FS: Formal Statement. FP: Formal Proof. RT: Reasoning Type. Significant $p < 0.005$ are marked with **bold**.

Inspection Dimension	Requirement	Valid	Invalid	p-value
(D1) IS Correctness	*Whether the informal statement is factually correct.*	93.50	83.50	**0.00167**
(D2) IS Relevance	*Whether the informal statement is relevant to each seed concept.*	87.50	92.50	0.09604
(D3) RT Classification	*Whether the informal statement is of the required question type.*	67.00	68.50	0.74903
(D4) IP Correctness	*Whether the informal proof correctly solves the informal statement.*	88.50	73.50	**0.00012**
(D5) IS-FS Alignment	*Whether the informal statement and the formal statement describe the same problem and are aligned with each other.*	74.00	66.50	0.10138
(D6) IP-FP Alignment	*Whether the informal proof and the formal proof describe the same solution and have aligned proof steps.*	72.00	54.00	**0.00018**

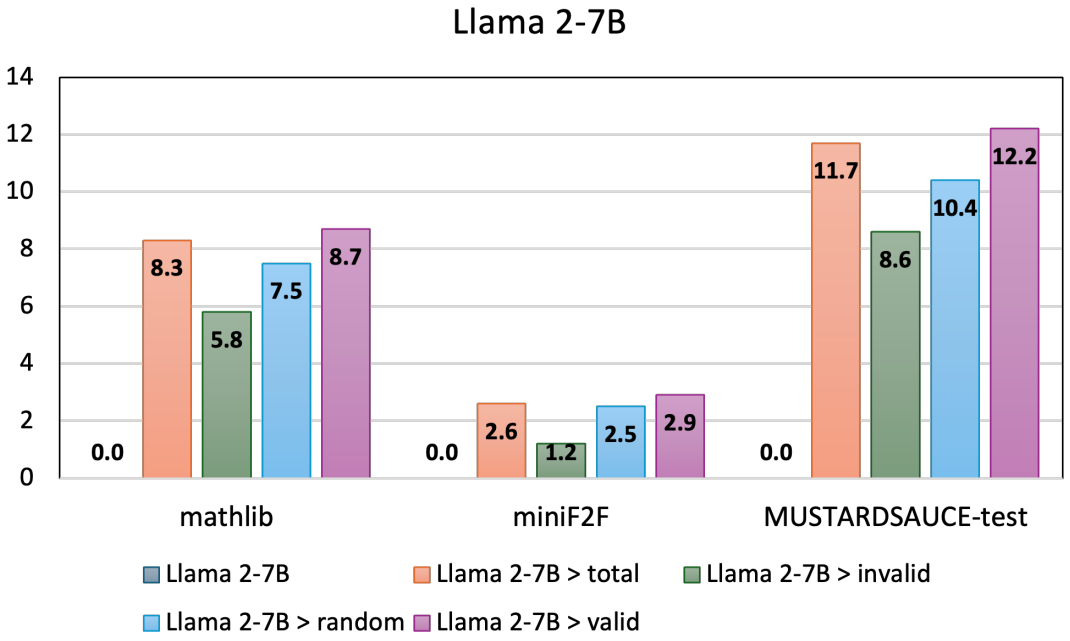
Data Benefits: Automated Theorem Proving

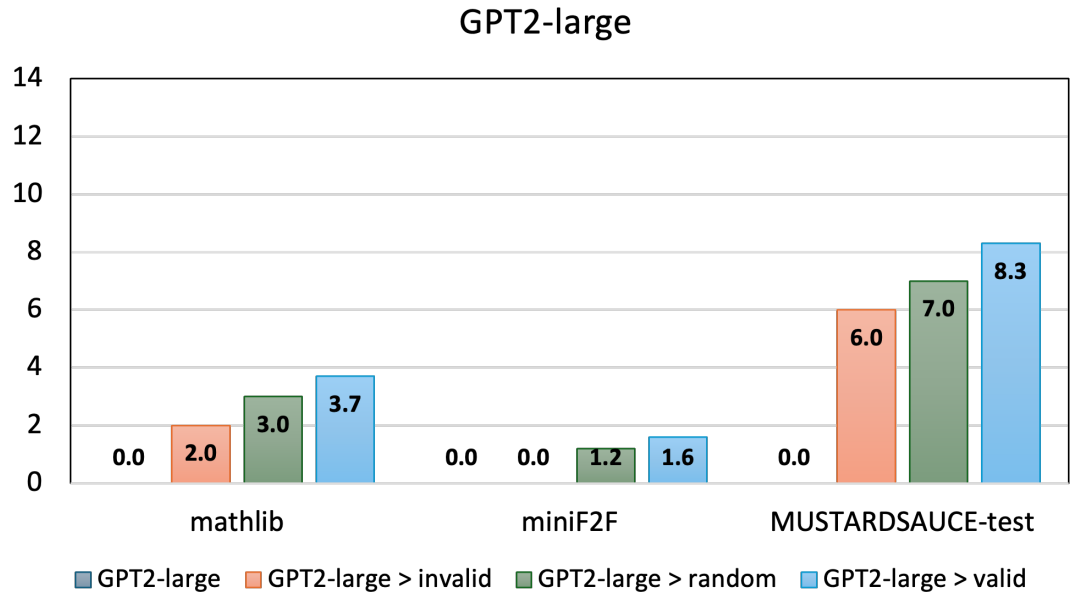
- Using the generated MUSTARDSAUCE to fine-tune LMs and then validate on automated theorem proving (ATP).

✓ Average of **18.15%** relative performance gain.

✓ Llama 2-7B achieves **16.00%** gain on mathlib, **16.00%** gain on miniF2F, and **17.31%** gain on MUSTARDSAUCE-test.

Fig. Pass@1 results on automated theorem proving (ATP) tasks. > denotes a fine-tuning step. test: MUSTARDSAUCE-test. Note that the reported results on MUSTARDSAUCE-test are obtained by only fine-tuning on the MUSTARDSAUCE-valid training split.





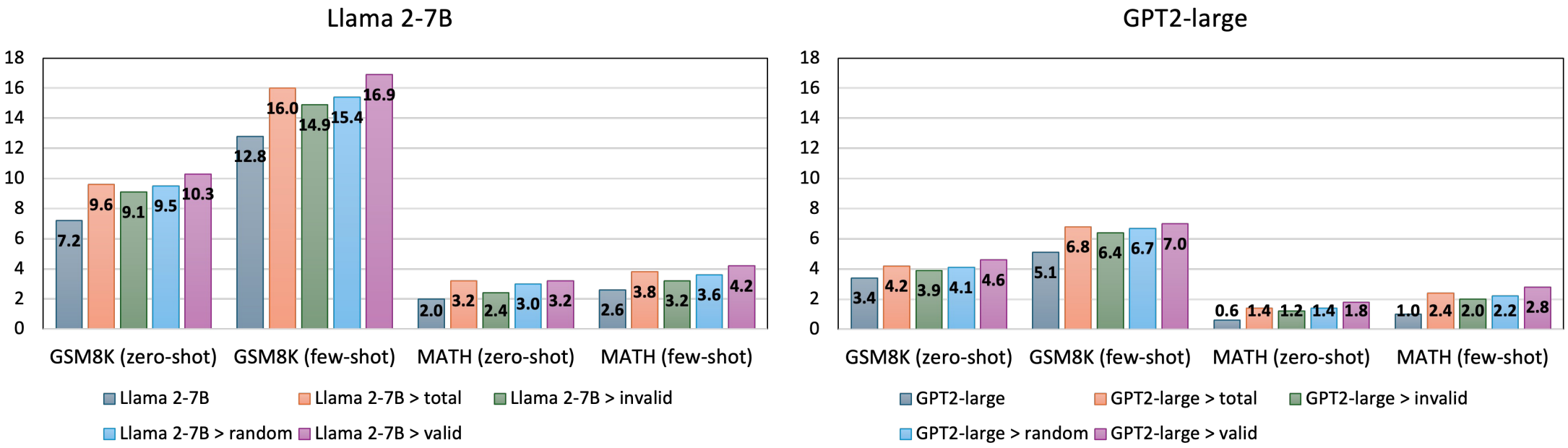
Data Benefits: Math Word Problem

- Using the generated MUSTARDSAUCE to Fine-tune LMs and then validate on math word problems (MWP).

✓ Average of **11.01%** relative performance gain.

✓ GPT2-large achieves **28.57%** gain on MATH, **12.20%** gain on GSM8K, zero-shot setting.

Fig. Maj1@1 results on GSM8K (G) and MATH (M). Zero: Zero-shot. Few: Few-shot. > denotes a fine-tuning step.



MUSTARD!

ICLR

Poster: Halle B, Tue 7 May 4:30— 6:30 p.m. CEST

1. We propose a novel framework **MUSTARD** that can generate high-quality mathematical data (both informal and formal) with an interplay between generative language model and theorem prover assistants.
2. We release the **MUSTARDSAUCE**, which contains both math word problems and theorem-proving problems spanning elementary to higher educational levels. Each sample has corresponding informal and formal solutions.
3. We conduct extensive analysis and experiments on the generated data, demonstrating their quality, diversity, and effectiveness in improving language models' mathematical reasoning performance.

Contact: yinya.huang@hotmail.com

👆 Read our paper

👆 Try our code

👆 Take on the exciting challenges!