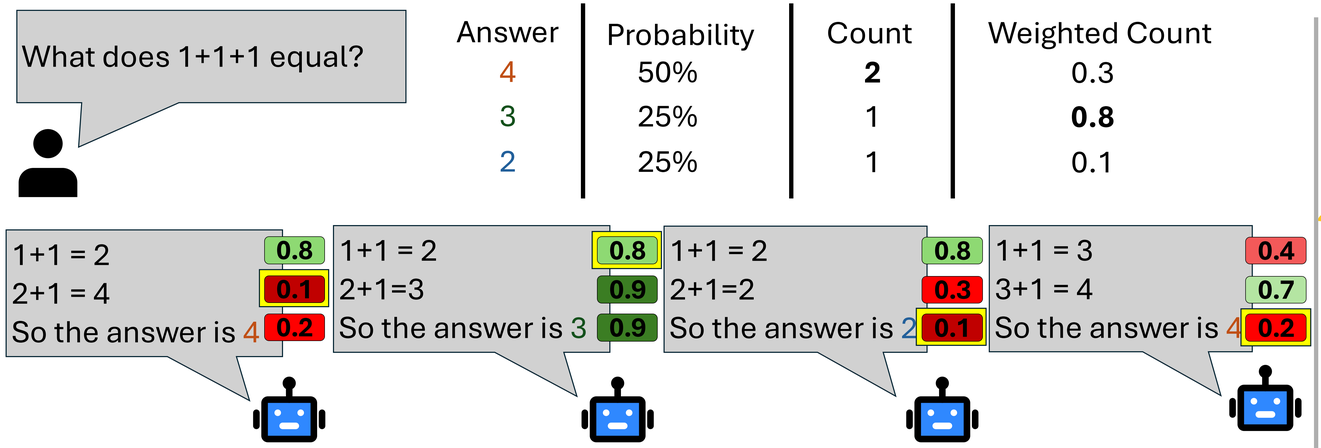
VersaPRM: Multi-Domain Process Reward Model via Synthetic Reasoning Data

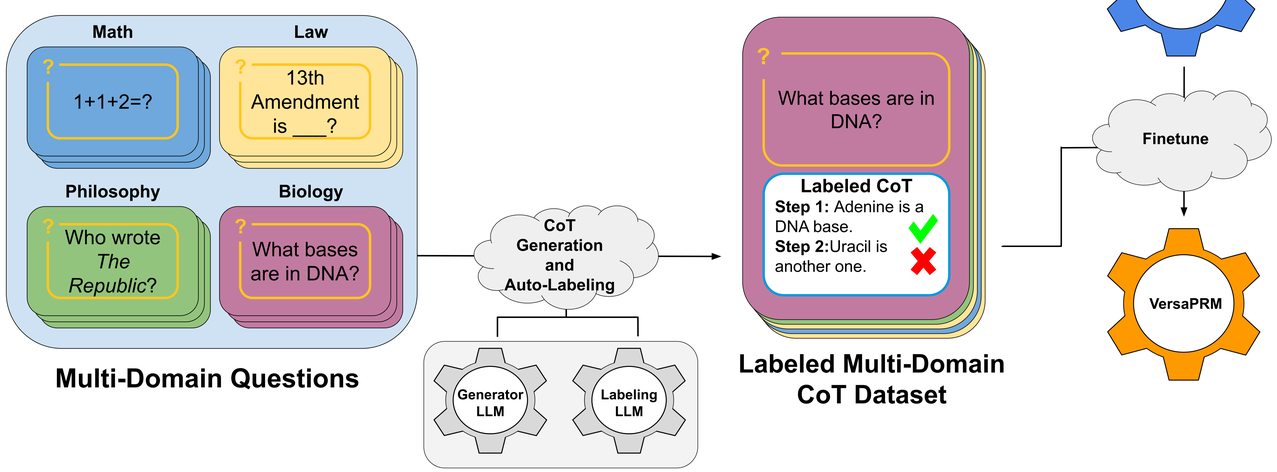
Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, Kangwook Lee

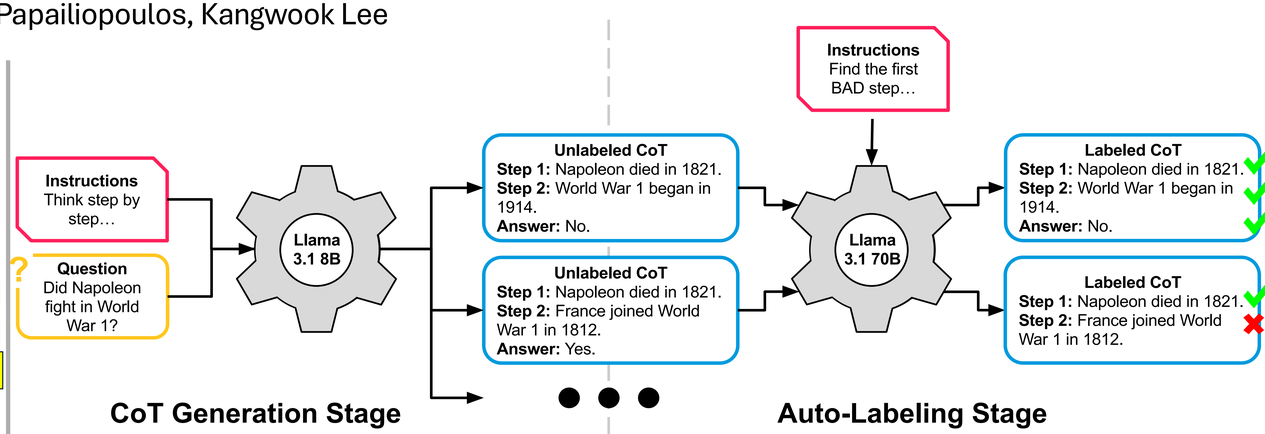


Can we create a PRM that generalizes to domains beyond math?

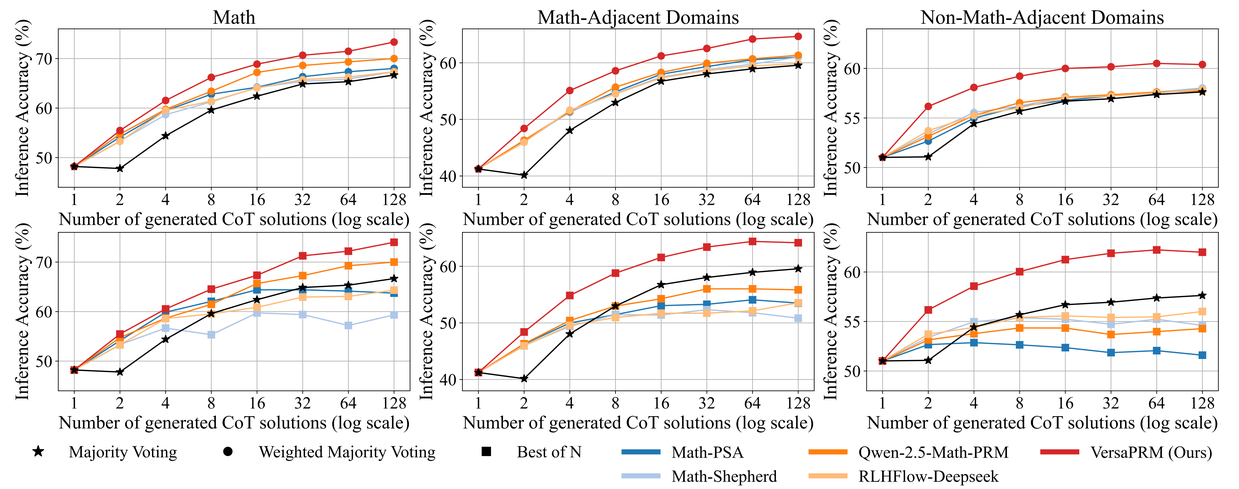
To create dataset to finetune VersaPRM:

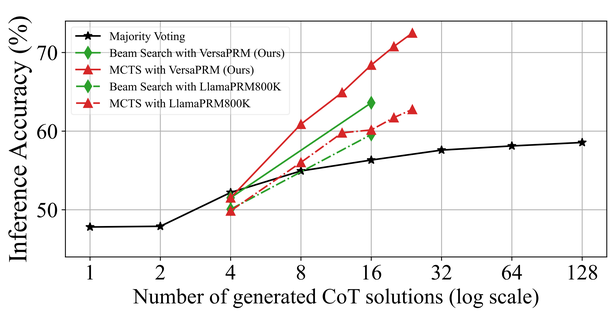
- We source multi-domain multiple choice questions from MMLU-Pro
- Use Llama70B to annotate correctness of reasoning step
- *Crucially*, provide the correct final answer in the system prompt used for annotation





Experimental Results





Findings:

- Finetuning with synthetic multi-domain data can enhance PRM performance in non-math domains
- Improvement is present across all tested domains
- The key ingredient is **diverse** data that we **synthetically** generate