LLEMMA

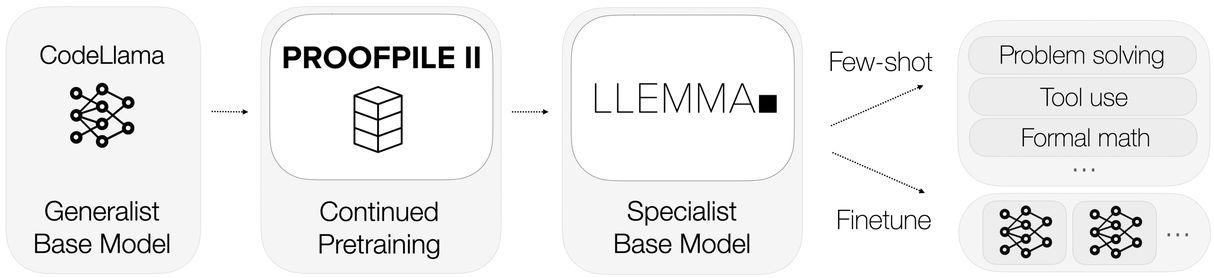
Llemma: An Open Language Model for Mathematics

Zhangir Azerbayev[1,2] Hailey Schoelkopf[2] Keiran Paster[3] Marco Dos Santos[4]
Stephen McAleer[5] Albert Q. Jiang[4] Jia Deng[1] Stella Biderman[2] Sean Welleck[5]

[1]Princeton [2]Eleuther AI [3]Toronto [4]Cambridge [5]CMU

Llemma

Key idea: math foundation model via continued pretraining on Proofpile II



We initialize from Code Llama and train Llemma-7b for 200B tokens, and Llemma-34b for 50B tokens.

Data: Proofpile II

1. **Mathematical Code**: `AlgebraicStack` (11B tokens): Code from GitHub and Stack, 17 languages.

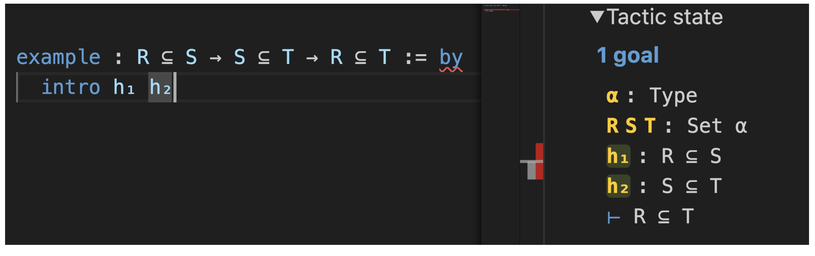


Figure 1. The `AlgebraicStack` includes extracted proof states from Lean and Isabelle; in total 1.5B tokens of formal math.

2. **Mathematical web data**: `OpenWebMath` (15B tokens): Filtered CommonCrawl web pages

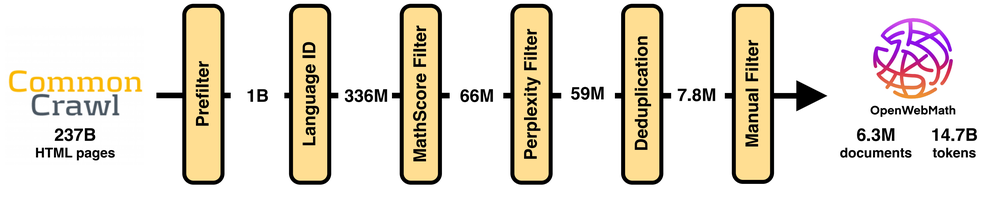
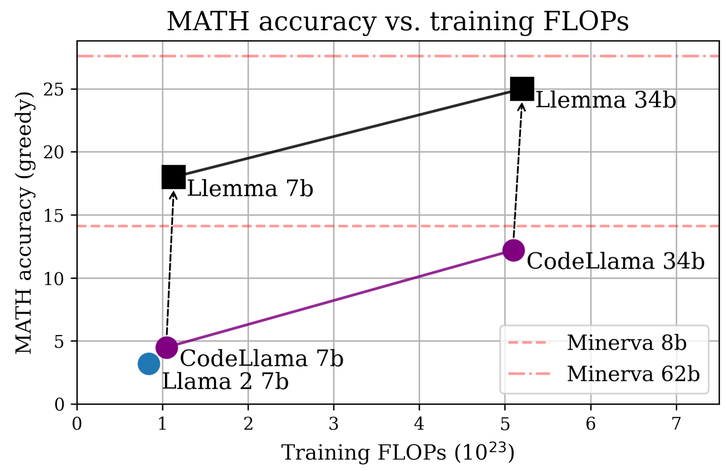


Figure 2. *OpenWebMath: An Open Dataset of High-Quality Mathematical Web Text* at ICLR!

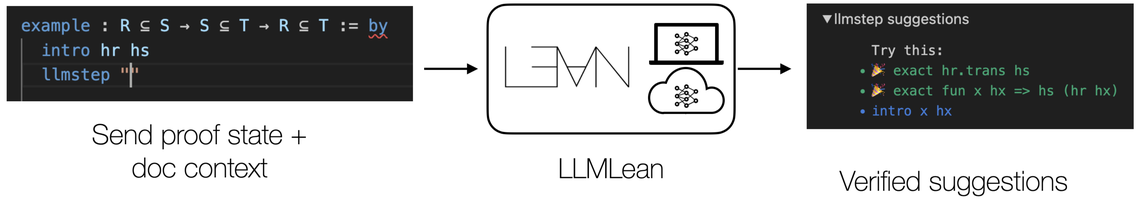
3. **Scientific Papers**: arXiv (29B tokens)

Continued Pretraining on Proofpile II improves mathematical capabilities



Problem solving, tool use, formal-to-formal and informal-to-formal theorem proving.

Llemma + Lean theorem prover



`LLMlean`: `https://github.com/cmu-l3/llmlean`.
Verified proof suggestions from Llemma running on your laptop or in the cloud.

Everything open: models, data, code

- Code: `https://github.com/EleutherAI/math-lm`
- Models: `https://huggingface.co/EleutherAI/llemma_7b`
- Data: `https://huggingface.co/datasets/EleutherAI/proof-pile-2`