ai4fm.cs.luc.edu

ChatTLA+ - Can LLMs Write Correct TLA+ Specifications?

Eric Spencer

LOYOLA
UNIVERSITY CHICAGO

The issue:

Software should be mathematically correct

Cruise control, elevator doors, bank software, etc.

The solution:

Model checking programming languages

verify a system's edge cases by ensuring a bad action can never be true

LOYOLA
UNIVERSITY CHICAGO

TLA+: Temporal Language of Actions

Invariants must never happen

$$\wedge\ radiation' = OFF$$

Figure 6. Scenario leading to an unsafe condition: The microwave is initially radiating with the door closed and several seconds of time remaining. When the user opens the door, harmful radiation comes out.

Läufer, Konstantin; Mertin, Gunda; Thiruvathukal, George K. (2024). WIP: An Engaging Undergraduate Intro to Model Checking in Software Engineering Using TLA+IEEE Frointiers in Education, 2024. https://doi.org/10.6084/m9.figshare.27226500.v1

https://arxiv.org/abs/2407.21152

LOYOLA UNIVERSITY CHICAGO

Large Language Models

The language problem

- Only **~200** verified examples of natural language -> model checking in TLA+

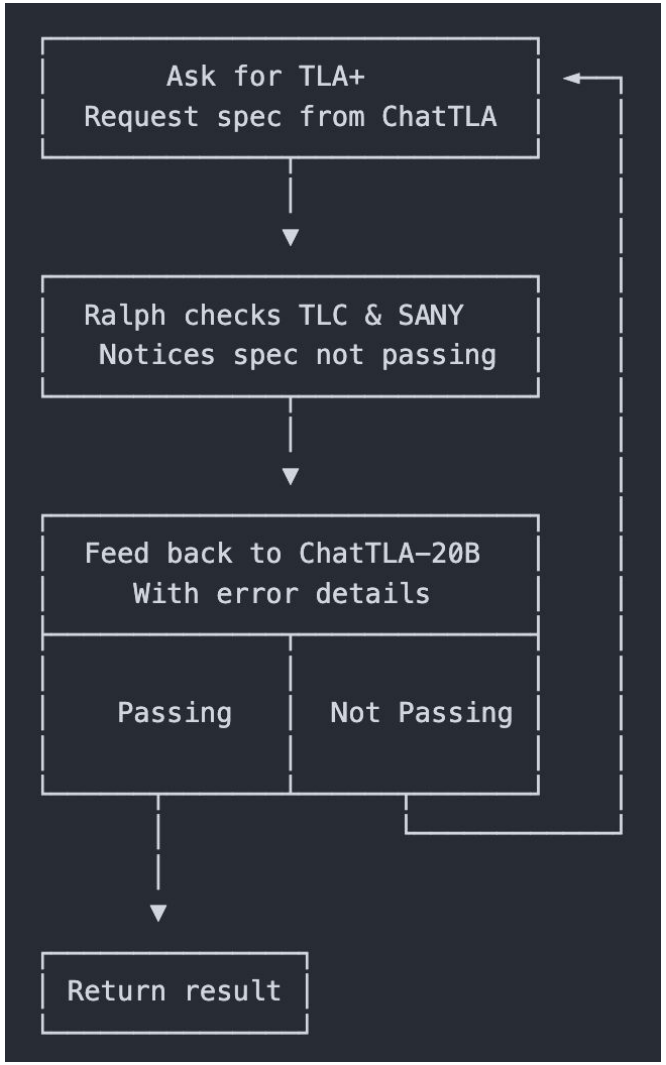
- Existing LLMs only output correct syntax **26.6%** of the time and output a verifiable model **8.6%** of the time (B. Ortiz et. al, 2026).

https://ai4fm.cs.luc.edu/papers/llm-tla-evaluation/

The context-repair solution

- Let the model try once
  - It might fail, give it the error output
- Let it try again

- This results in a much higher verified result

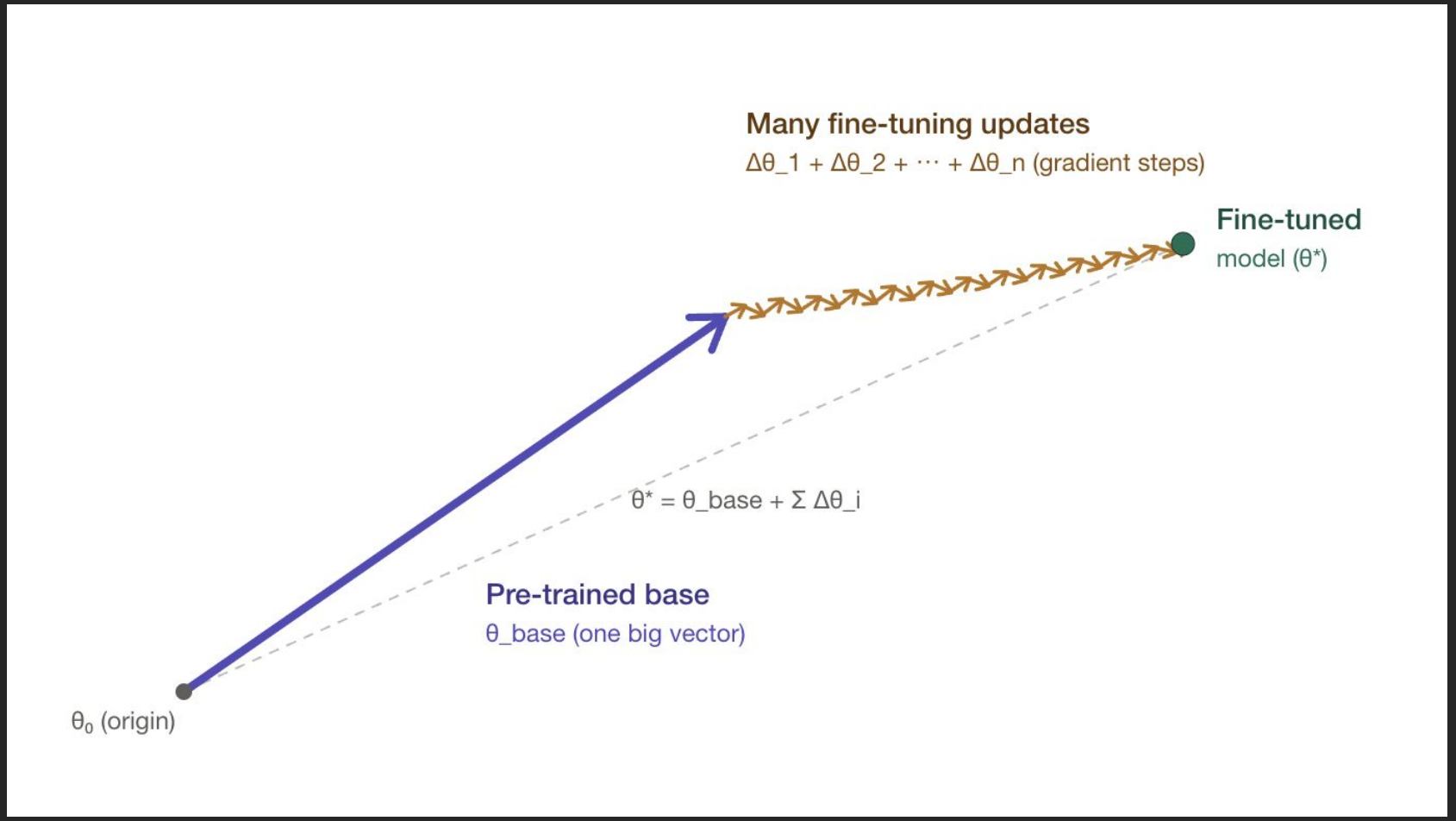


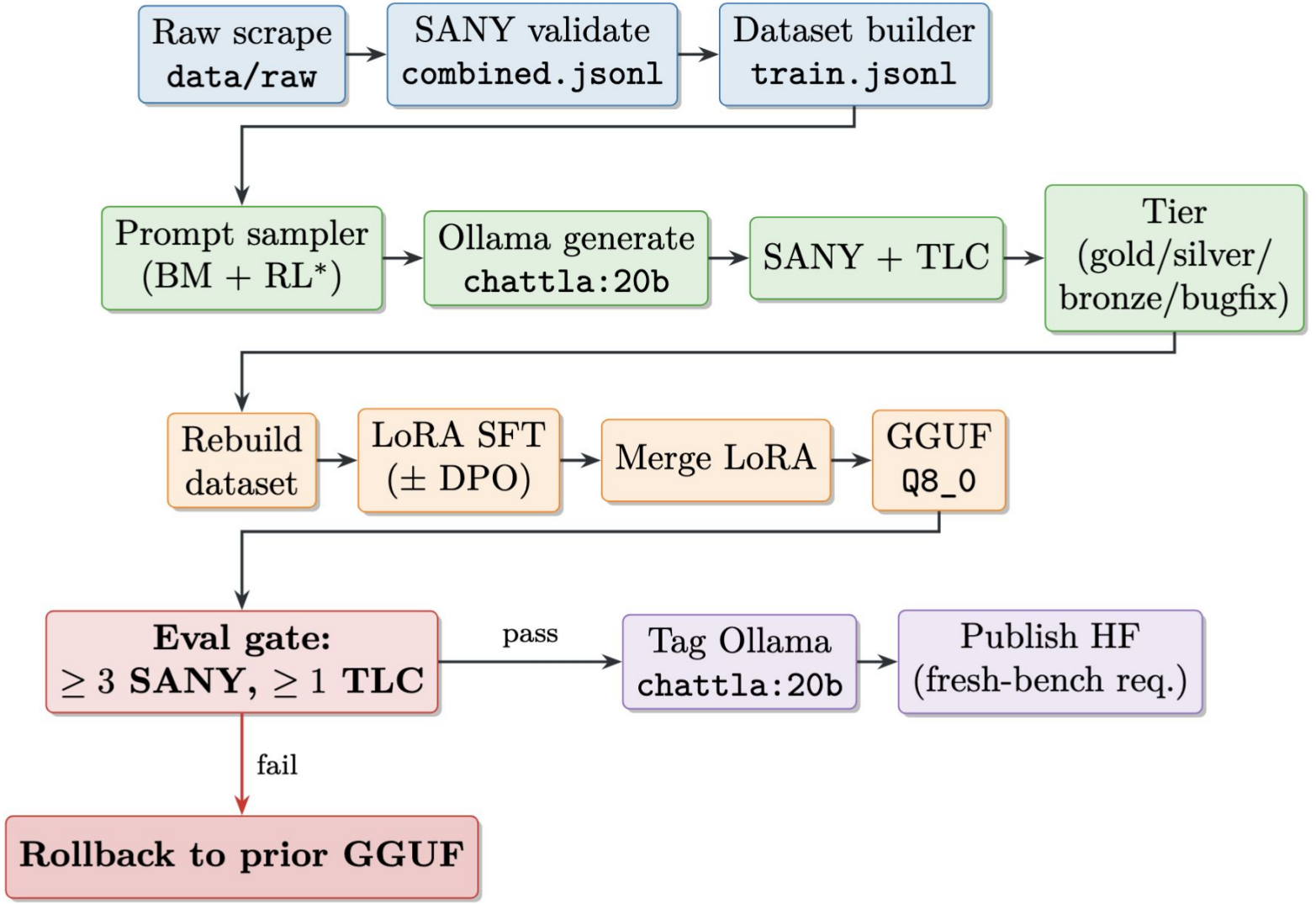
https://github.com/LUC-AI4FM/ralph-tla

Applicable to other languages
(E. First et. al, 2023).

https://people.cs.umass.edu/~brun/pubs/pubs/First23fse.pdf

Reinforcement learning




Raw scrape
data/raw
SANY validate
combined.jsonl
Dataset builder
train.jsonl
Prompt sampler
(BM + RL*)
Ollama generate
chattla:20b
SANY + TLC
Tier
(gold/silver/
bronze/bugfix)
Rebuild
dataset
LoRA SFT
(± DPO)
Merge LoRA
GGUF
Q8_0
Eval gate:
≥ 3 SANY, ≥ 1 TLC
pass
Tag Ollama
chattla:20b
Publish HF
(fresh-bench req.)
fail
Rollback to prior GGUF

Results

- The fine-tuned model achieved better results than base models
- At it’s best: **80%** syntax, **25%** semantic performance on 30 results

Metric	gpt-oss:20b	ChatTLA v9	Gain
SANY	69.2%	80%	**+10.8 pp (1.16×)**
TLC	23.1%	25%	**+1.9 pp (1.08×)**

Strategy	SANY	TLC
Few-Shot	**18/26 (69.2%)**	0/26 (0%)
Half-Completion	3/26 (11.5%)	0/26 (0%)
Progressive	9/26 (34.6%)	**6/26 (23.1%)**
FIM	5/26 (19.2%)	0/26 (0%)

ai4fm.cs.luc.edu

Thank you!

Eric Spencer - AI for Formal Methods in Software Engineering

LOYOLA
UNIVERSITY CHICAGO