UNIVERSITY OF TARTU
Institute of Computer Science

GREEN AI

DO LARGER AI MODELS REALLY DEBUG BETTER?

A Performance and Sustainability Analysis of LLMs for Python Bug Fixing

Author: Syed Fakhar Abbas Naqvi | MSc Software Engineering | Supervisor: Hina Anwar

Web Link: https://github.com/syedfakhar25/empirical-llm-bugfix-benchmark

WHY IT MATTERS?

The rapid adoption of AI in software debugging highlights the need for models that are not only accurate, but also fast, efficient, and sustainable.

The Problem

- Large Language Models (LLMs) are widely used for code generation, debugging, and program repair.
- Larger models require more computational resources and consume significantly more energy during inference.
- It is still unclear whether bigger models always deliver better performance for real-world software engineering tasks.

This study investigates the relationships among **model size, functional correctness, inference time, and energy consumption in real-world Python bug-fixing tasks.**

Methodology

DATASET
- 40 real-world Python bug-fixing tasks from BugsInPy
- 9 diverse open-source projects

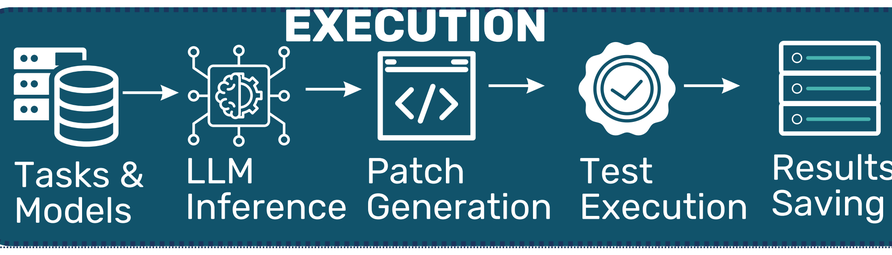
MODELS EVALUATED
- Qwen1.5B
- CodeLlama13B
- DeepSeek6.7B
- StarCoder7B
- StableCode3B
- StarCoder15B

(Size range: 1.5B to 15B parameters)

EVALUATION METRICS
- Functional Correctness:Pass@1,Pass@10
- Inference Time (seconds)
- GPU Energy Consumption (Joules)

SETUP
- UT-HPC, NVIDIA Tesla V100
- 32GB, 4CPU per job, Exclusive Environment

EXECUTION



SMALLER MODELS OUT PERFORMED LARGER MODELS

ON REAL-WORLD PYTHON BUG-FIXING TASKS

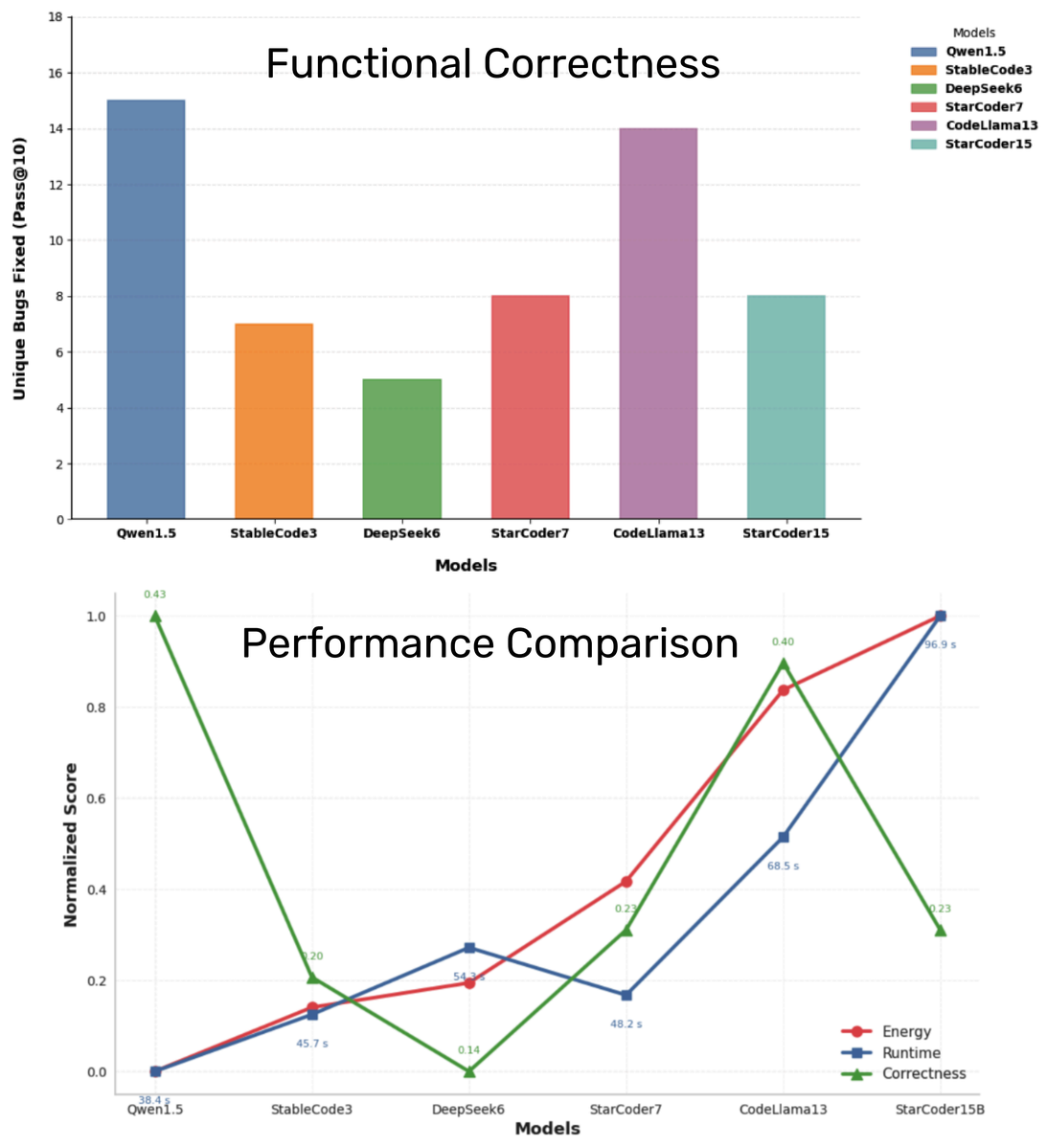
15
BUGS FIXED
Qwen1.5 fixed the most unique tasks

6X
LESS ENERGY
Qwen1.5 consumed 526J/inference vs. 3203J/inference by StarCoder15

38.7s
FASTEST INFERENCE
Qwen1.5 achieved the shortest average inference time

BIGGER IS NOT ALWAYS BETTER.
SMALLER MODELS CAN BE MORE ACCURATE, FASTER AND FAR MORE ENERGY-EFFICIENT

Results



Smaller Models
- Simpler, test-compliant fixes
- Fewer structural errors
- Better on localized bugs

Larger Models
- More verbose and complex patches
- Extra logic caused test failures

Key Contributions

- ✓ Real-world comparison of open-source LLMs across model scales
- ✓ Shows that smaller models can achieve strong accuracy with lower energy use
- ✓ Provides practical insights for sustainable AI-assisted debugging

Implications

- ✓ Smaller models reduce inference cost and energy consumption
- ✓ Efficient models can still deliver reliable bug-fixing performance
- ✓ Sustainability should be considered alongside correctness in AI selection

Conclusion

- ✓ Smaller models offer the best balance between performance & sustainability
- ✓ Efficient AI can support practical and environmentally responsible software engineering
- ✓ Bigger models are not always the best choice for real-world debugging tasks