UNIVERSITÄT
LEIPZIG

Mathematics and Machine Learning

BENCHMARKING PROGRAM-AIDED VS. NATURAL LANGUAGE MATHEMATICAL REASONING IN LLMS

Mohamed El Maghari, Ralf König, Fritz Körner, Khalid Sabih

Leipzig, 18.02.2026

MOTIVATION

The Paradox of Modern LLMs in Mathematics

- **AlphaProof** (DeepMind, 2025) — silver medal at IMO 2024
  - Solved P6, the hardest problem, solved by only 5 of 609 humans
  - Uses Lean 4 formal verification → zero hallucination possible
- **Yet top models fail at this:**
  - *"Alice has N brothers and M sisters. How many sisters does Alice's brother have?"*
  - State-of-the-art models answer this **incorrectly**
- **So which is it — can LLMs do math or not?**

→ **Research Goal:** Compare Pure LLM vs. Python vs. Lean 4 on the same problems, same model, same conditions

HYPOTHESES

- **H1:** Easy problems solvable across all three modalities
- **H2:** Hard problems benefit from Python code generation
- **H3:** Lean 4 zero-shot accuracy will be very low
- **H4:** Code-specialized LLMs outperform general-purpose LLMs as program-aided models

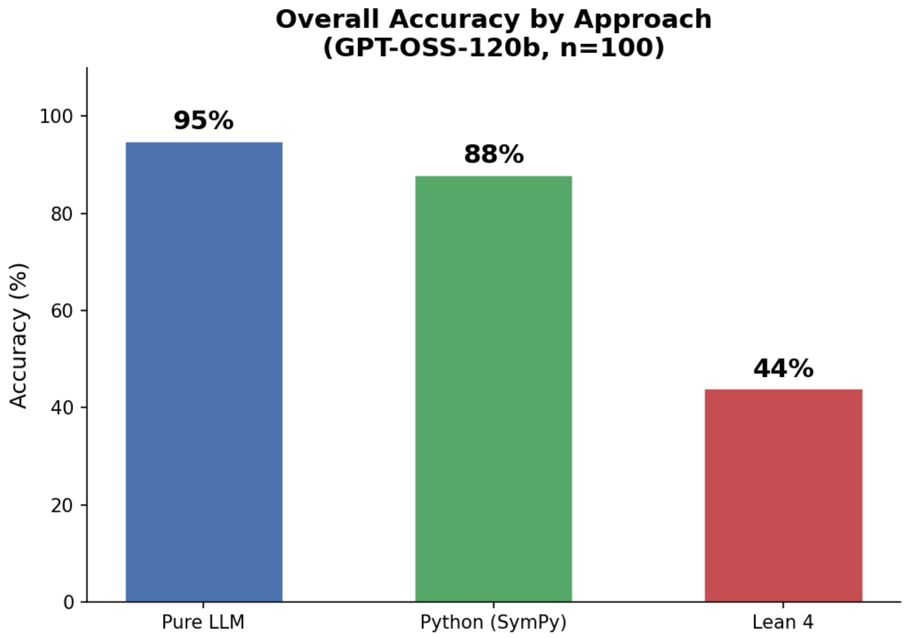
METHODOLOGY / DATASET

- Dataset: MATH-500 benchmark — 100 problems sampled
- 7 categories × 5 difficulty levels (L1–L5)
- Model tested: GPT-OSS-120b via Blablador (Helmholtz)
- Zero-shot prompting — no examples, no feedback loop
- Python sandbox with SymPy, NumPy, Math libraries
- Lean 4 via Kimina-LEAN-Server for remote proof verification
- Evaluation: LLM-as-Judge

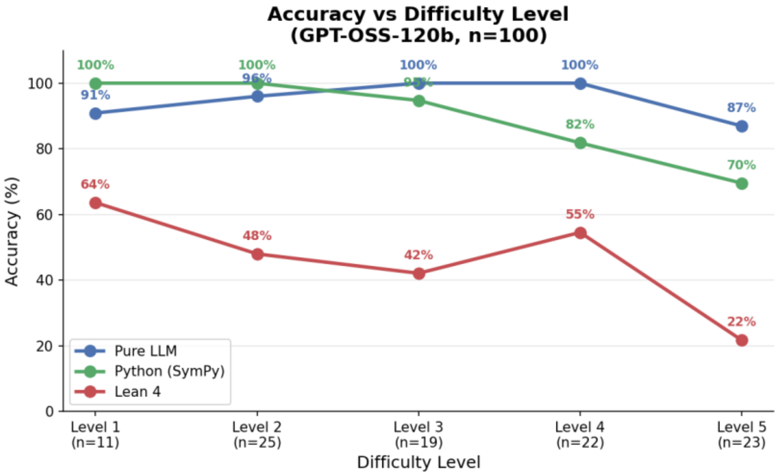
RESULTS

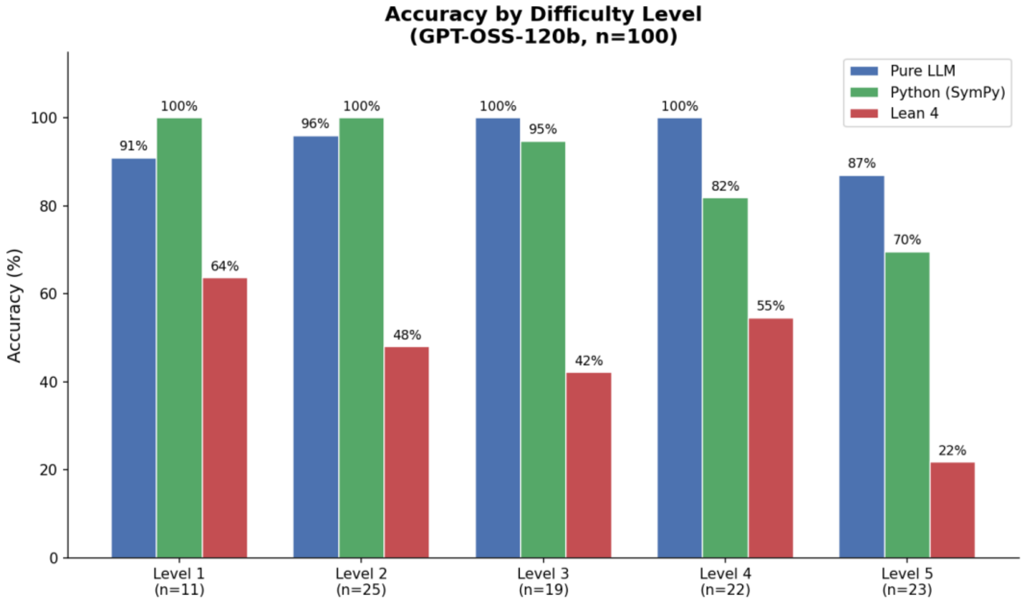
Overall Accuracy:

- Pure LLM 95%
- Python (SymPy) 88%
- Lean 4 44%



RESULTS : ACCURACY BY DIFFICULTY LEVEL

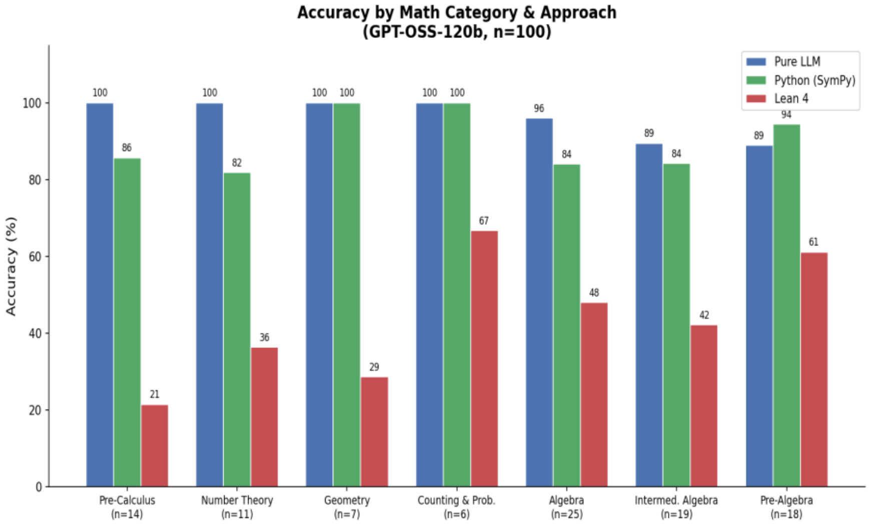




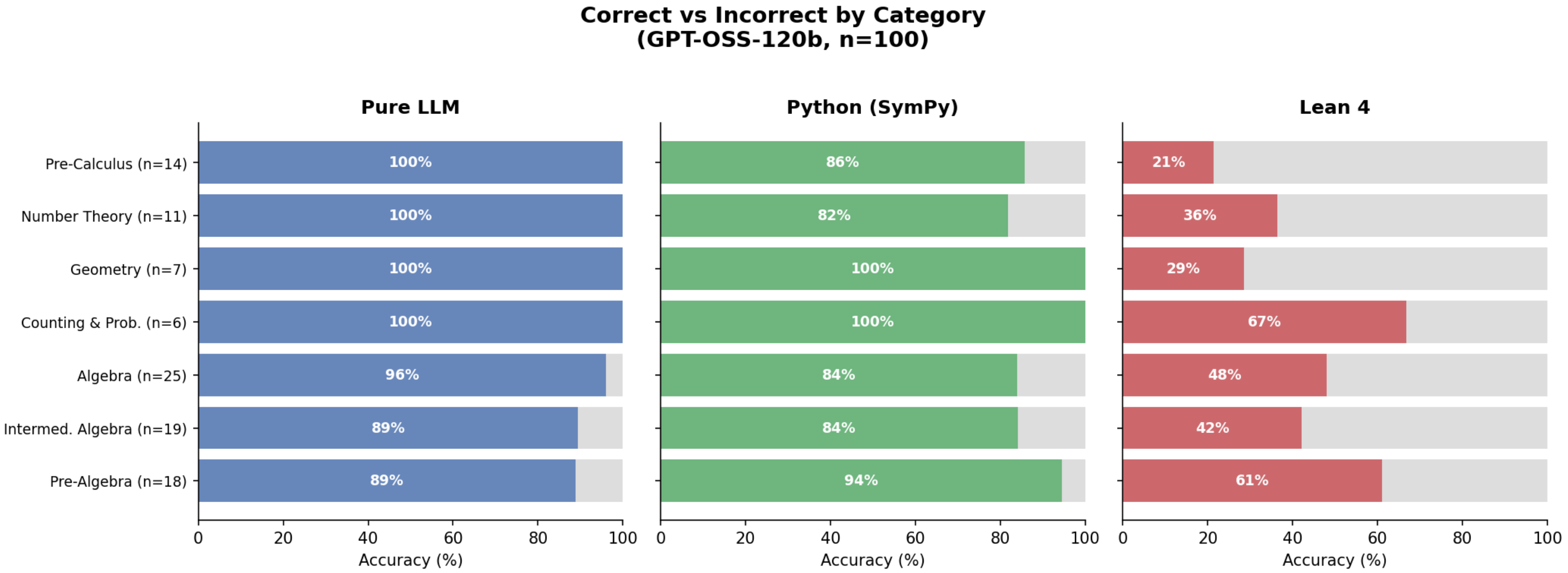
RESULTS: ACCURACY BY MATH CATEGORY & APPROACH

Category Breakdown:

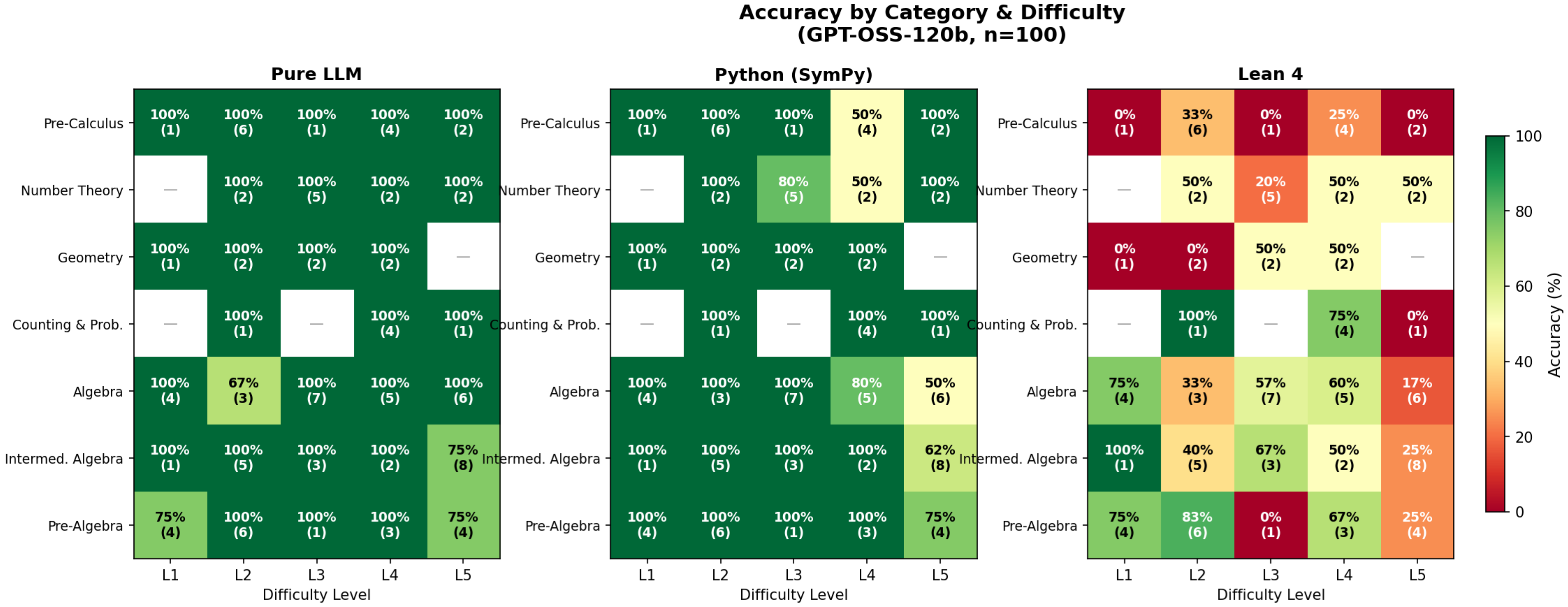
- Pure LLM dominates
- Lean 4 struggles especially in Pre-Calculus (21%) and Geometry (29%)



RESULTS : CORRECT VS INCORRECT BY CATEGORY



RESULTS : ACCURACY BY CATEGORY & DIFFICULTY



CONCLUSION

- Pure LLM: best overall accuracy **(95%)**
- Python (SymPy): competitive but drops at higher difficulty **(88%)**
- Lean 4: significant underperformance across all dimensions **(44%)**
- Difficulty scaling disproportionately hurts Lean 4 **(64% → 22%)**
- Proof generation — not mathematics — is the core bottleneck
- Zero-shot Pure LLM remains the most practical approach today

FUTURE WORK

- Benchmark additional models: GPT-4o, Claude, Gemini
- Test few-shot and chain-of-thought prompting strategies
- Fine-tune models specifically on Lean 4 proof corpora
- Explore hybrid pipelines: Python solving + Lean 4 verification
- Scale dataset beyond 100 problems and more categories
- Analyze Lean 4 failure modes systematically

UNIVERSITÄT LEIPZIG

THANK YOU!

Ralf König
Fritz Körner
Khalid Sabih
Mohamed El Maghari