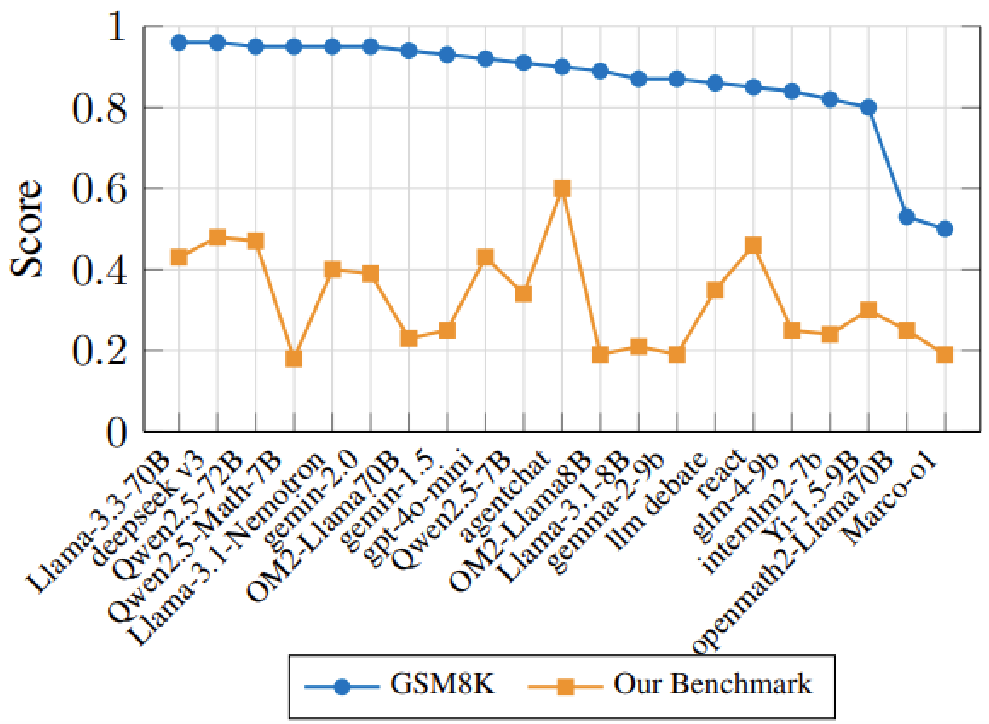
Benchmarking Abstract and Reasoning Abilities Through A Theoretical Perspective

MAC:Media Analytics & Computing Laboratory

Xiamen University

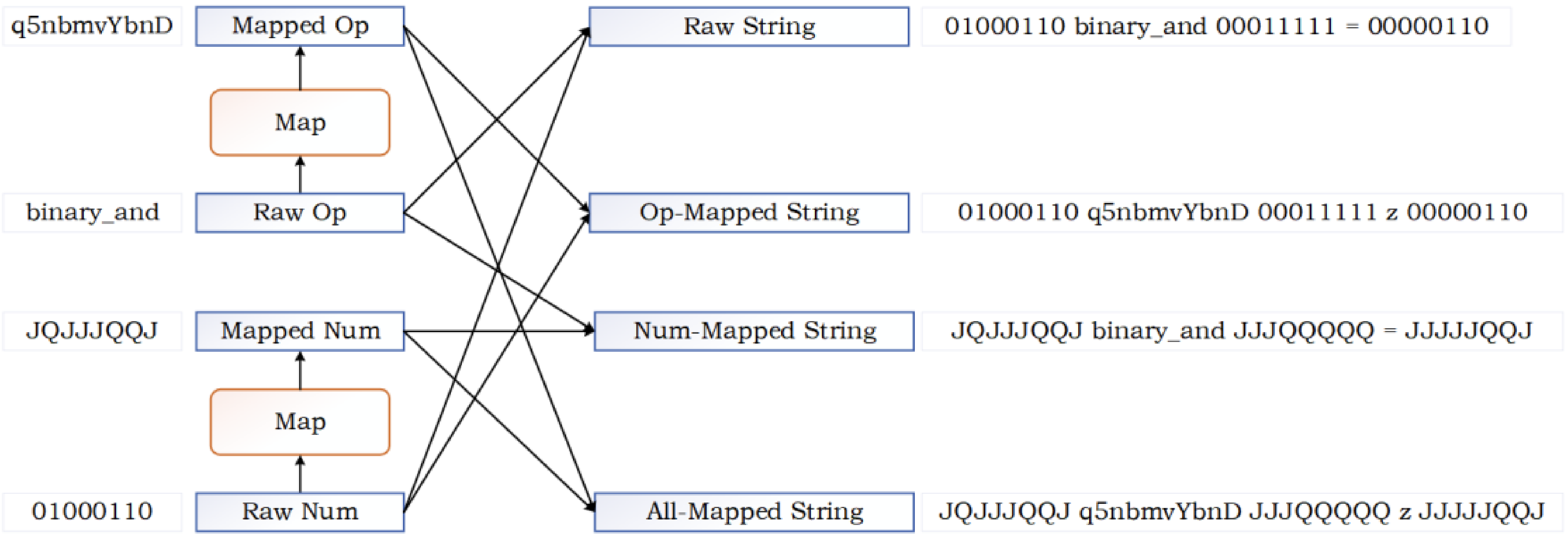
The Problem & Motivation



LLMs: Smart or Memorizing?

High scores (e.g., GSM8K) ≠ True Abstract Reasoning.
Current tests: Surface patterns, not deep understanding.
Goal: Rigorously test **true** LLM abstract reasoning.

Our Approach & Metrics



Theory-Driven Evaluation

Abstract Reasoning: Extract Patterns (**f**) → Apply Rules (**Re**).

Metrics:

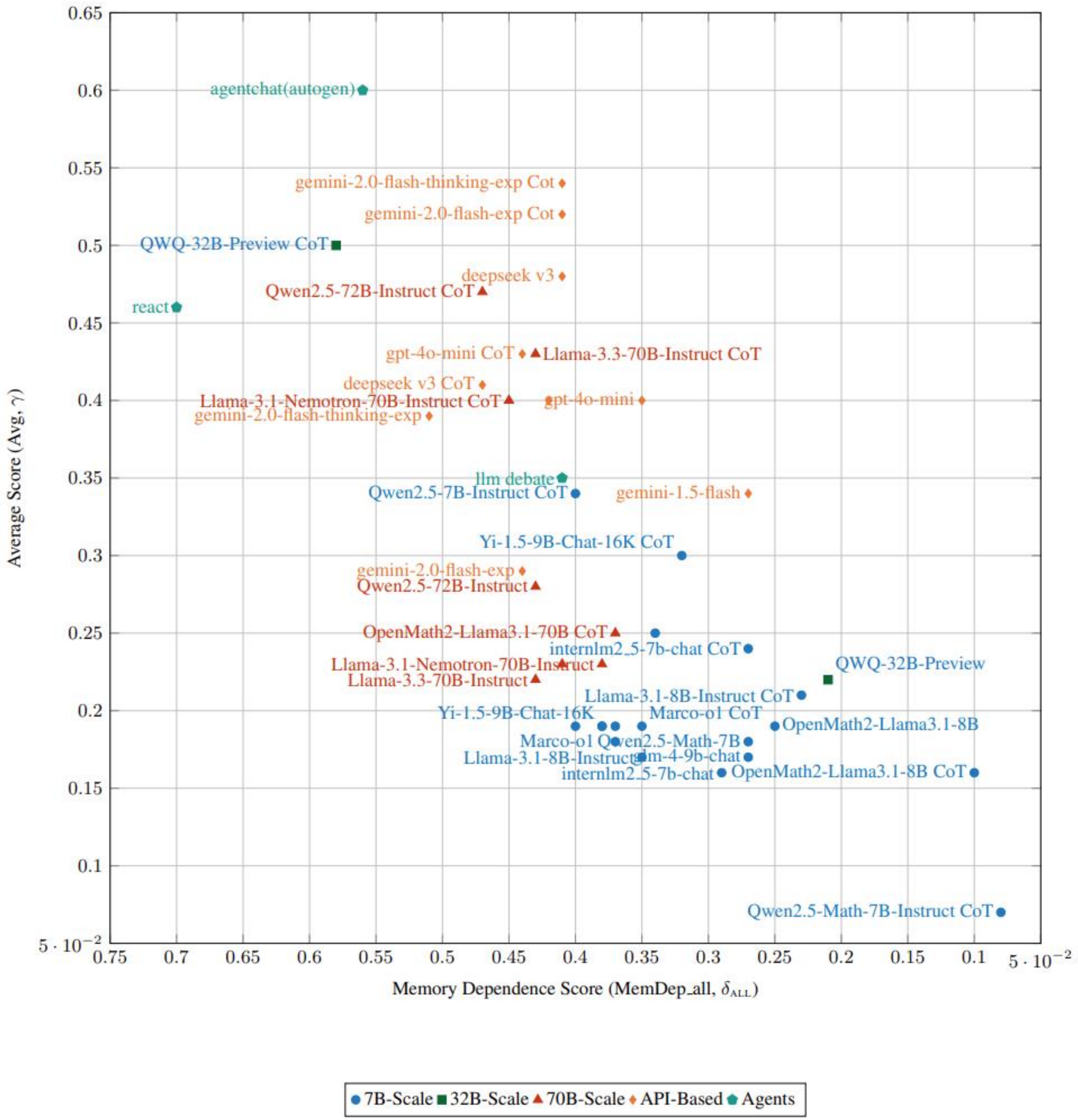
- **Γ**: Base Accuracy.
- **Δ**:Memory Reliance(Γ_original - Γ_remapped).

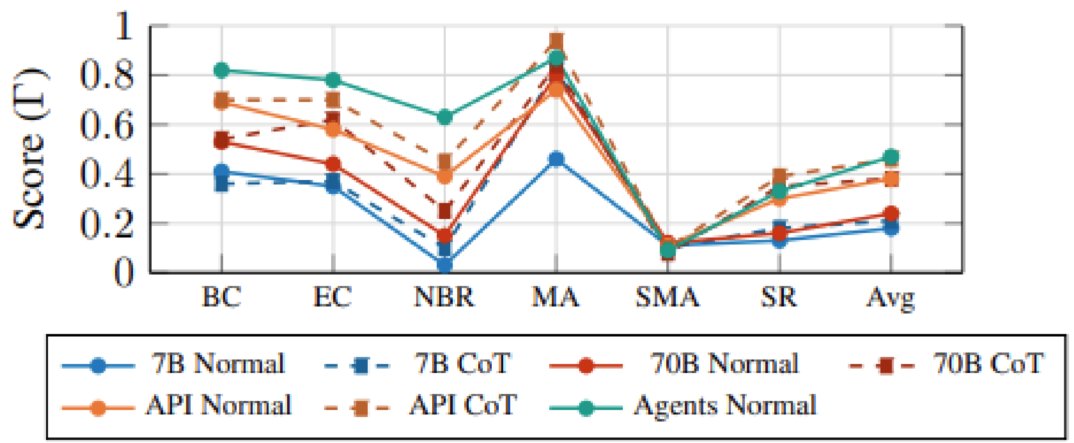
High Δ = Memorization.

Key Design: Symbol Remapping (e.g., ‘1+1=2 ’→‘A op A=B’)

Tests understanding beyond token matching.

Key Findings





LLMs: Memorization Over Abstraction

1. **Failures**: Widespread in non-decimal arithmetic(**NBR**).
2. **High Δ**: Rely on operand symbols (memory), not abstract patterns.
3. **CoT Trade-off**: ↑ Performance often → ↑ Memory Dependence.

Conclusion

Our robust theoretical framework rigorously assessed LLM abstract reasoning. By defining abstract reasoning's interplay, we validated metrics (Γ, Δ) and designed a symbol remapping benchmark for genuine generalization. Evaluations revealed a critical LLM deficit: a profound lack of abstract symbolic reasoning, driven by significant memory dependence and limited generalization, even with advanced techniques.

Impact: Our tools & benchmark guide development of truly intelligent LLMs.
Open Source: github.com/MAC-AutoML/abstract-reason-benchmark

Thank You!