RE-IMAGINE: Symbolic Benchmark Synthesis for Reasoning Evaluation

Xinnuo Xu[1] * Rachel Lawrence[1] * Kshitij Dubey[2] * Atharva Pandey[2] * Fabian Falck[1] Risa Ueno[1] Aditya V. Nori[1] Rahul Sharma[2] Amit Sharma[2] Javier Gonzalez[2]

[1]Microsoft Research Cambridge [2]Microsoft Research India

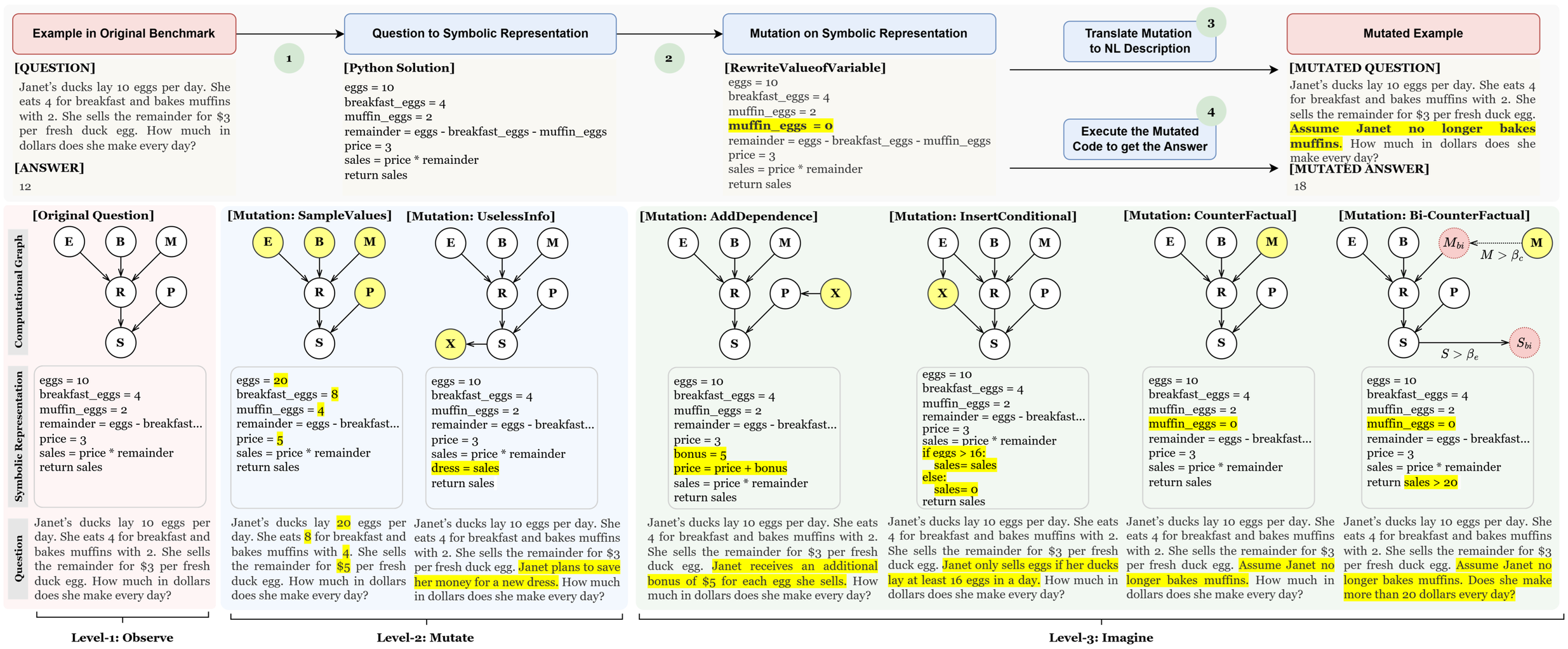

Figure 1. Benchmark Transformation Pipeline

The Ladder of Reasoning

Inspired by the **Ladder of Causation (Pearl, 2009)**, we define and label a three-layer hierarchy — *observe*, *mutate*, and *imagine* — that characterizes different levels of reasoning abilities in LLMs, similar to the cognitive skills captured by Pearl's ladder of causation. **"Only machines that can correctly perform correlations, interventions and counterfactuals will have reasoning abilities comparable to humans."**

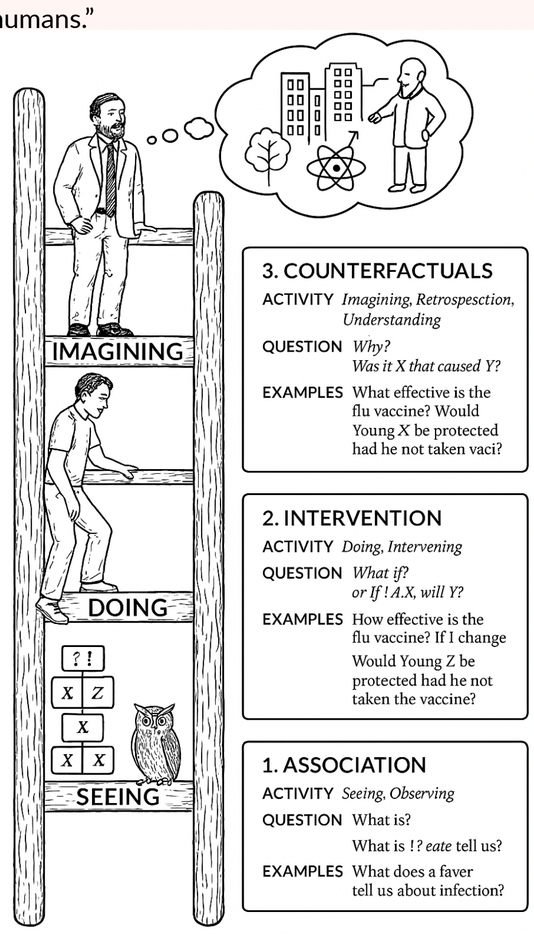

Figure 2. Pearl's Ladder of Causality

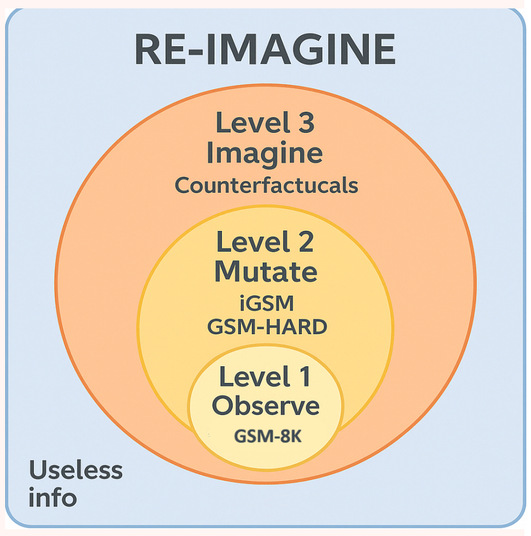

Figure 3. Re-Imagine Coverage Of Benchmarks

Results

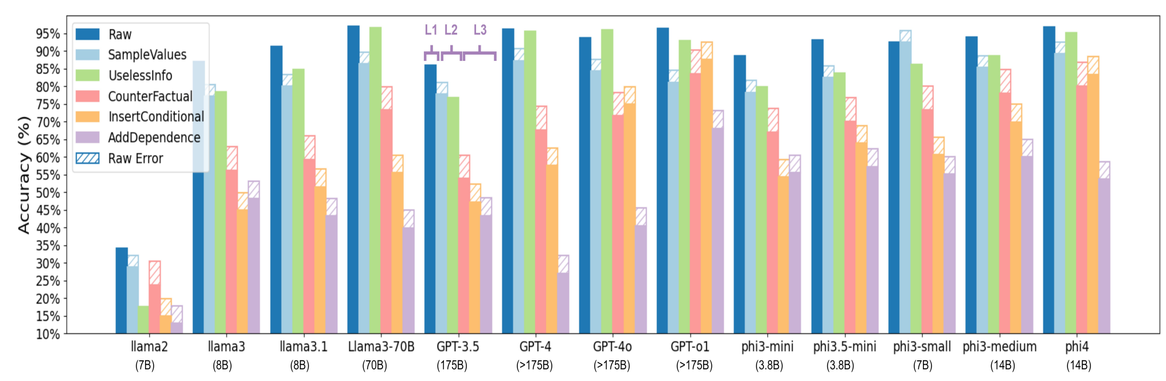

Figure 4. GSM8K Results

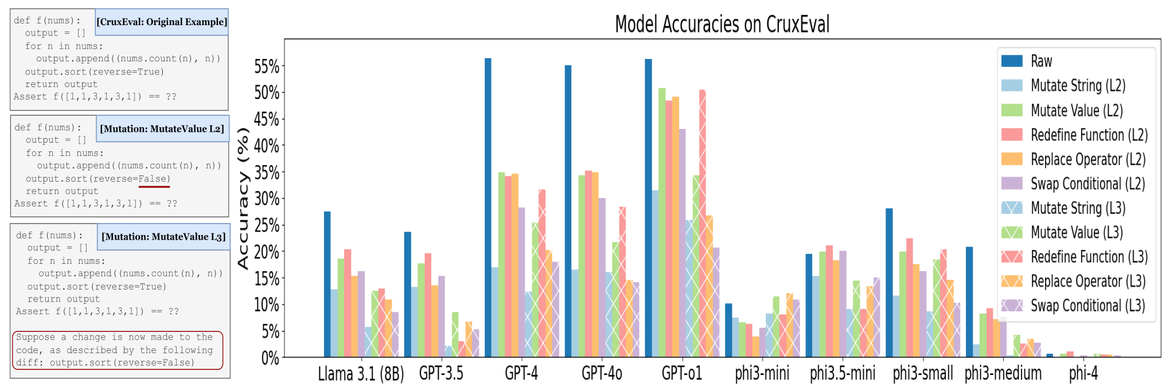

Figure 5. CruxEval Example Mutations and Results

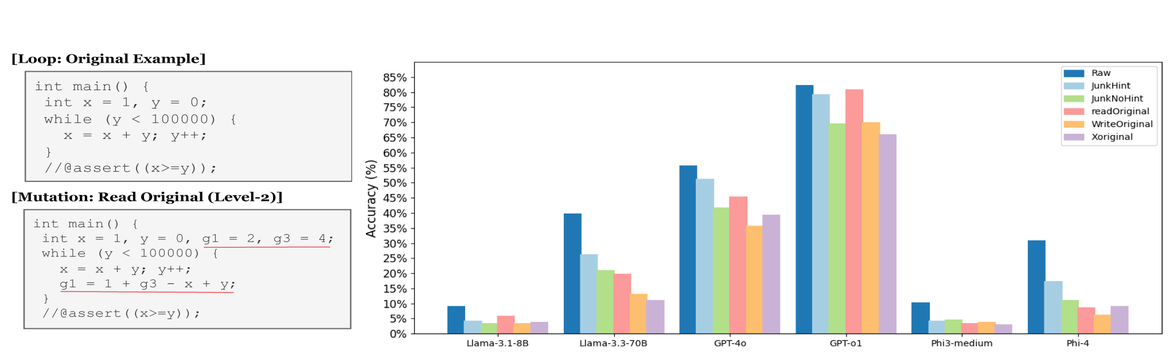

Figure 6. Loop Example Mutations and Results

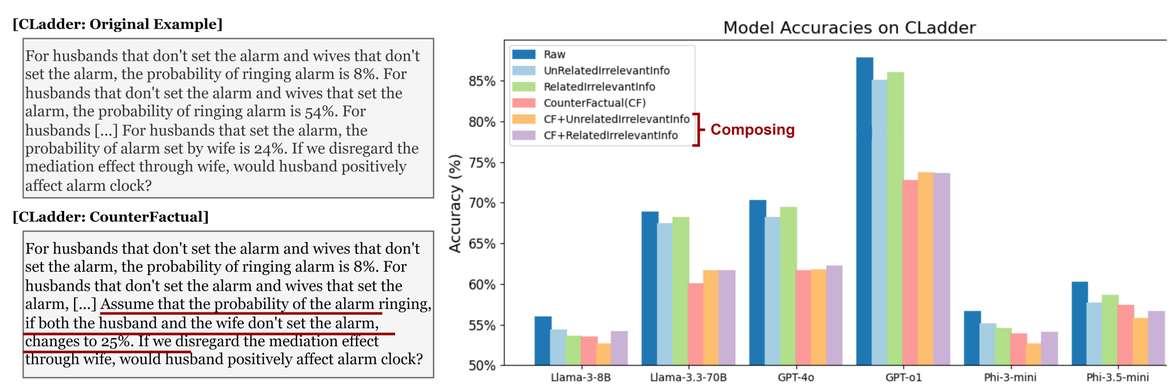

Figure 7. Cladder Example Mutations and Results

Novelty and the Influence of the Question Difficulty

- We define a hierarchy of reasoning difficulty that unifies prior mutation strategies with new ones introduced in our work.
- We present the first **scalable mutation generation pipeline**, applicable across benchmarks and tasks, enabling arbitrary mutations at each reasoning level.
- Prior work is mostly limited to **Level-2**; existing **Level-3** efforts rely on manual rules and lack scalability.
- Our framework overcomes these gaps by supporting multi-level, cross-benchmark problem variation.
- **More results and ablations are available in the paper.**

How do the three levels relate to the causal ladder?

- Our framework connects to Pearl's ladder through the problem's *computation graph*, viewed as a causal model.
- Each benchmark problem represents a specific realization of this graph.
- Perturbations correspond to operations in Pearl's ladder.
- **However**, not all mutations have causal counterparts (e.g., adding irrelevant info or modifying operations).