An Agentic Framework for Benchmarking LLMs in Data-poor Domains

Michael Lu ’26

Swarthmore College; Foerster Lab for AI Research, University of Oxford

Introduction

- Existing benchmarks mainly measure the recall ability of Large Language Models.
- They are derived from data-rich domains (e.g. coding tasks, math problems, common knowledge).
- We propose a new framework to benchmark LLMs in data-poor domains, where there’s less training data available and memorization is less likely.
- Therefore, we can better assess the reasoning, planning, problem-solving, and self-reflection abilities of LLMs.

Methodology

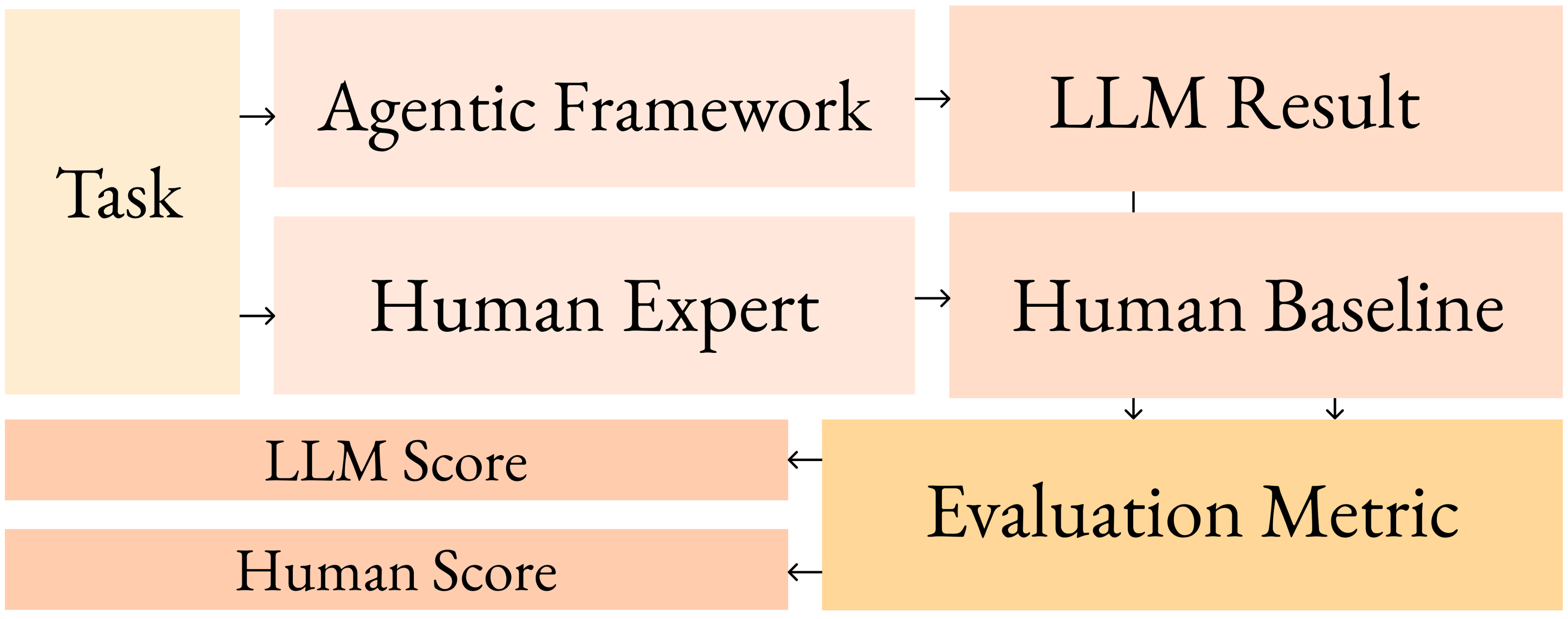

Fig 1. Benchmark Method Flowchart

- We selected a set of tasks from fluid dynamics, astrophysics, chemical engineering, etc.
- These tasks are in the format of "Recreate this experiment from scratch based on this paper."
- Both Human Experts and LLMs are placed into a containerized environment with access to the internet and computational resources.
- For their results, we developed an universal evaluation rubric
- I developed the agentic framework that allows LLMs to autonomously plan, execute, and iterate on experiments.

Agentic Framework

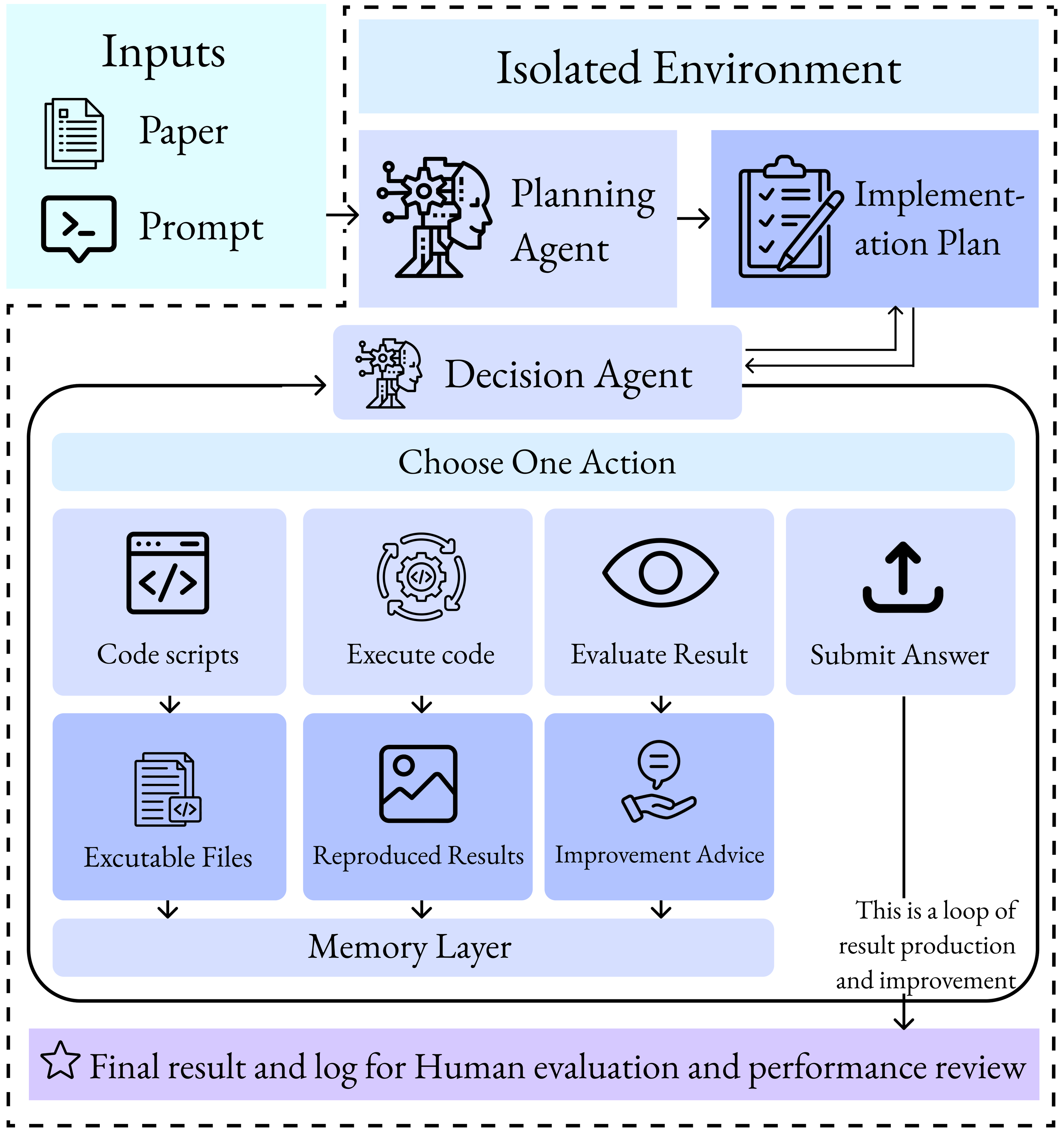

Fig 2. Agentic Framework Flowchart

- Materials describing the task (i.e. research paper, task prompt) are being passed to the Planning Agent to begin the process.
- The Planning Agent breaks down the task into sub-tasks and creates a high-level plan.
- Then it enters into an iterative loop where a decision agent is constantly evaluating progress and determining next steps.
  - -- Coding Agent: Powered by Aider, it writes and debugs code
  - -- Executor: Runs the code, logs output and store results
  - -- Evaluator: Assesses the results with LLM's own knowledge
  - -- Memory Manager: Keeps track of all previous actions, code versions, results, and summarizes them for future reference.

Results

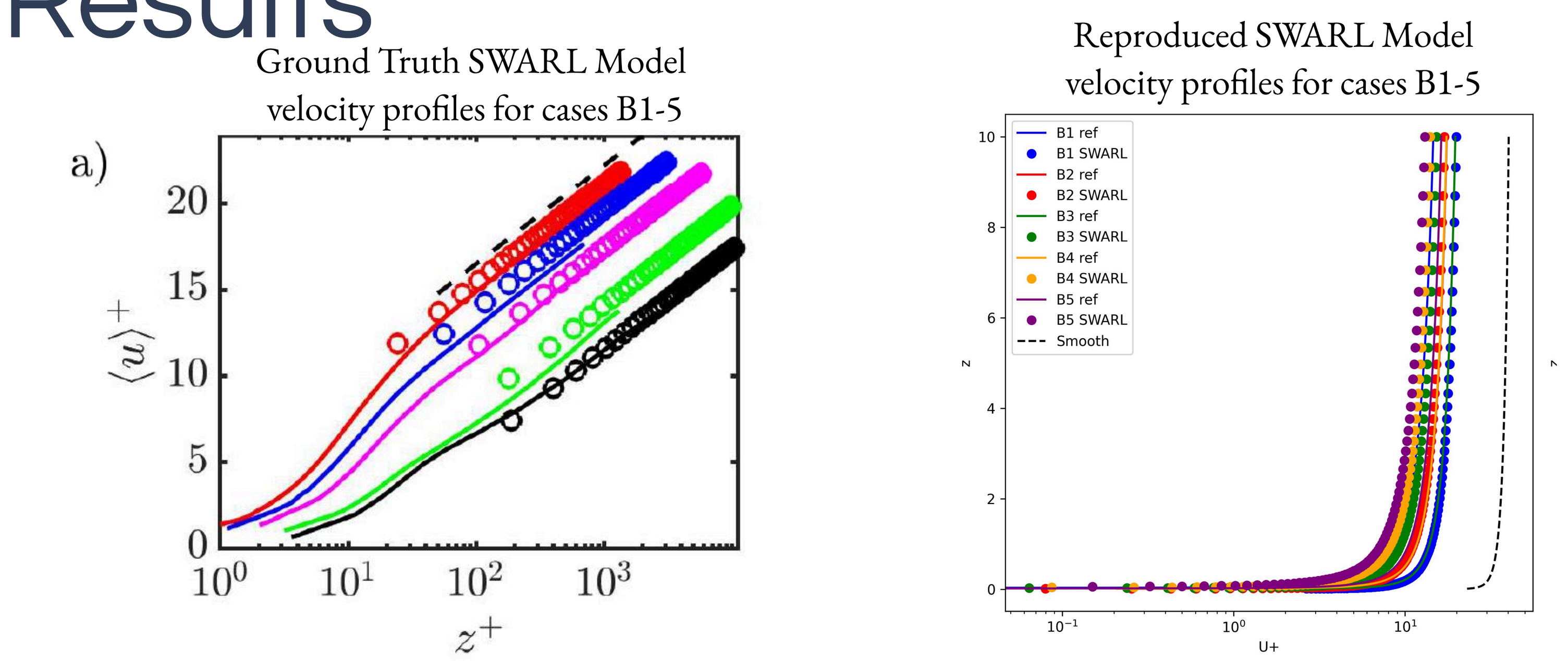

Fig 3. Ground Truth vs. Claude 4’s Reproduction Comparison in Fluid Dynamics Task

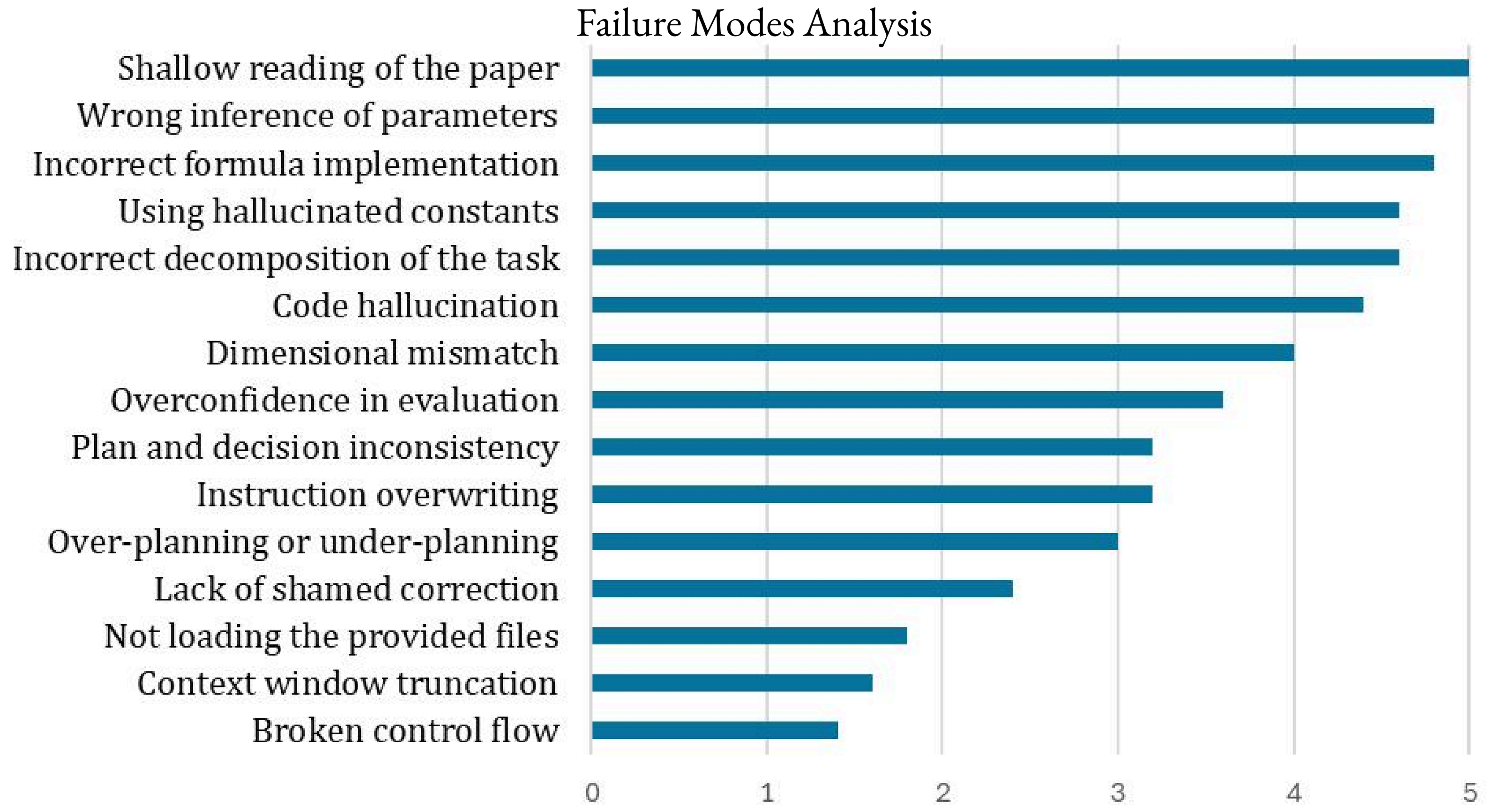

Fig 4. A histogram of the exposed problems of LLM in solving the Fluid Dynamics Task

- We selected a fluid dynamics task that requires recreating the SWARL Model and reproducing a sample experiment results from the *Ayala* paper (2024).
- We benchmarked state-of-the-art LLMs. None of them can 100% faithfully recreate the experiment.
- The framework itself is robust. There is rarely a case of flow breakdown.
- LLMs struggled with extracting detailed methodologies from the paper, using made-up parameters or equations
- They also over-confidently evaluated their results and submitted them prematurely.

Future Works

- Adding a Paper Reader Agent to improve the semantic understanding of research papers
- Oxford Researchers will use this framework for future benchmarking, parallel with Human Baseline Trials.