ResiliBench: Evaluating Agentic Workflow Adaptation in Stochastic Environments

Ruicheng Ao*, Zeping Min*, Tingyu Zhu*, Wotao Yin, Xinshang Wang

Massachusetts Institute of Technology, Alibaba Group US, DAMO Academy, UC Berkeley

Introduction

We introduce **ResiliBench**, a benchmark evaluating LLM workflow execution under realistic conditions of instruction quality variability and tool execution uncertainty.

Benchmark Scale:

- 5,040 unique tasks across 5 task types
- 30 canonical software APIs in 6 functional categories
- 4 prompt variants per task (baseline, CoT, MDP-optimal, flawed)
- Base tool success rate: 0.8
- Evaluation: 10 conversational turns per task

Three Main Challenges

1. **Tool Reliability Issues**

   APIs exhibit probabilistic failures including timeouts, validation errors, and resource limitations. We model five primary failure types: `INVALID_INPUT`, `OPERATION_FAILED`, `TIMEOUT`, `CALCULATION_ERROR`, and `OVERFLOW`.

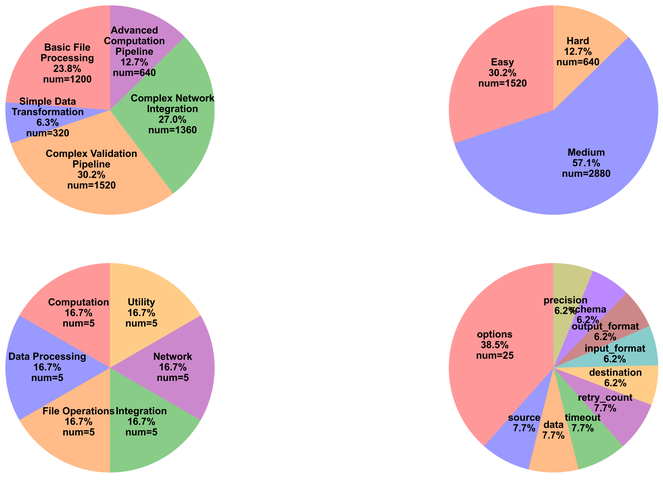
2. **Instruction Quality Variations**

   Users provide instructions that may be incomplete, ambiguous, or contradictory. We systematically generate flawed workflow prompts across seven perturbation categories.

3. **Complexity Challenges**

   Tasks involve intricate tool dependencies and dialogue turn limits that restrict model exploration behavior.

Benchmark Statistics



Benchmark Construction

MDP-Based Workflow Generation:
We develop a Markov Decision Process framework that generates workflows maximizing expected success rates under tool uncertainty.

Reward Function: Two-phase adaptive strategy

- *Phase I* (Coverage): Prioritize tool discovery and exploration
- *Phase II* (Sequence): Optimize execution order and efficiency

Four Prompt Variants:

1. *Baseline*: Essential task information only
2. *Chain-of-Thought*: Enhanced with reasoning instructions
3. *MDP-Optimal*: Detailed execution plan from MDP framework
4. *Flawed*: Systematic perturbations (ordering errors, tool misuse, parameter errors, missing steps, redundancies, logic discontinuity, semantic drift)

Evaluation Methodology

Task Outcome Levels:

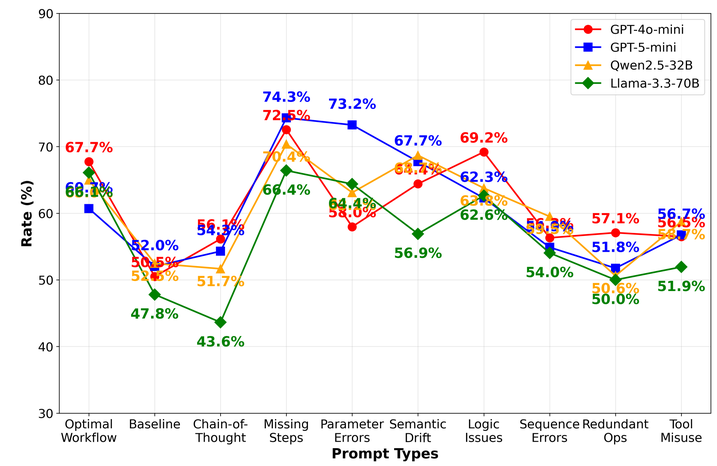
- `full_success`: All tools executed correctly in proper sequence
- `partial_success`: Substantial completion, most requirements satisfied
- `failure`: Insufficient completion due to critical failures

Overall Performance Results

Model	Baseline	CoT	Optimal
GPT-4o-mini	50.5%	56.1%	**67.7%**
O3-0416-Global	52.7%	48.9%	58.5%
Gemini-2.5-Flash	54.3%	51.1%	60.1%
GPT-5-mini	52.0%	54.3%	60.7%
Llama-3.3-70B	47.8%	43.6%	66.1%
Qwen2.5-32B	52.5%	51.7%	65.0%
DeepSeek-V3	50.0%	50.0%	56.8%
Average	51.4%	50.8%	**62.1%**

MDP-optimal workflow prompts achieve significantly higher success rates (62.1% average) compared to baseline (51.4%) and Chain-of-Thought (50.8%) approaches.

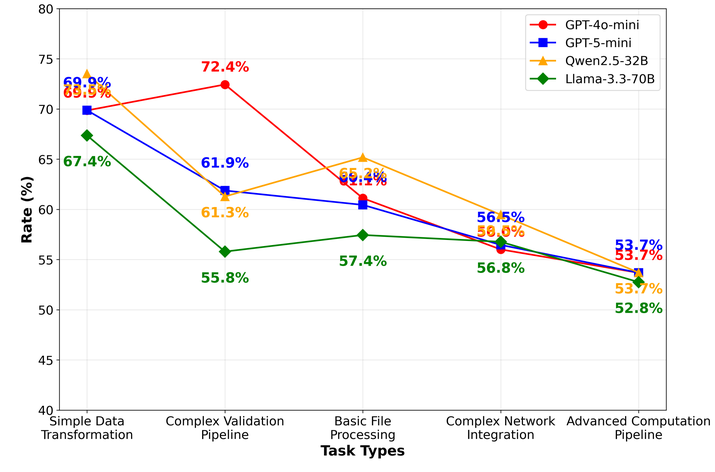
Key Results: Prompt Robustness Analysis



Finding 1: Dramatically Different Robustness Patterns
Models exhibit vastly different responses to flawed instructions:

- **GPT-4o-mini**: Stable performance (optimal: 67.7%, flawed: 62.2%, drop: 5.5 points)
- **Gemini-2.5-Flash**: Substantial degradation (optimal: 60.1%, flawed: 20.0%, drop: 40.1 points)
- Advanced models show resilience to ordering and parameter errors but vulnerability to semantic drift

Key Results: Task Complexity Analysis



Finding 2: Consistent Performance Degradation
All models show performance decline with increasing task complexity.