NeurIPS 2026 · Evaluations & Datasets Track · Poster 139250

PHMForge: Evaluating LLM Agents on Industrial Prognostics through MCP-Native, Algorithm-Grounded Tools

Yusheng Li[1] · Tianjun Feng[1] · Yunfeng Chen[1] · Chun-Yi Tsai[1] · Yihan Sun[1] · Dhaval Patel[3] · Kaoutar El Maghraoui[3] · Shuxin Lin[3] · Ayan Das[2]

[1]Columbia University [2]Georgia Institute of Technology [3]IBM

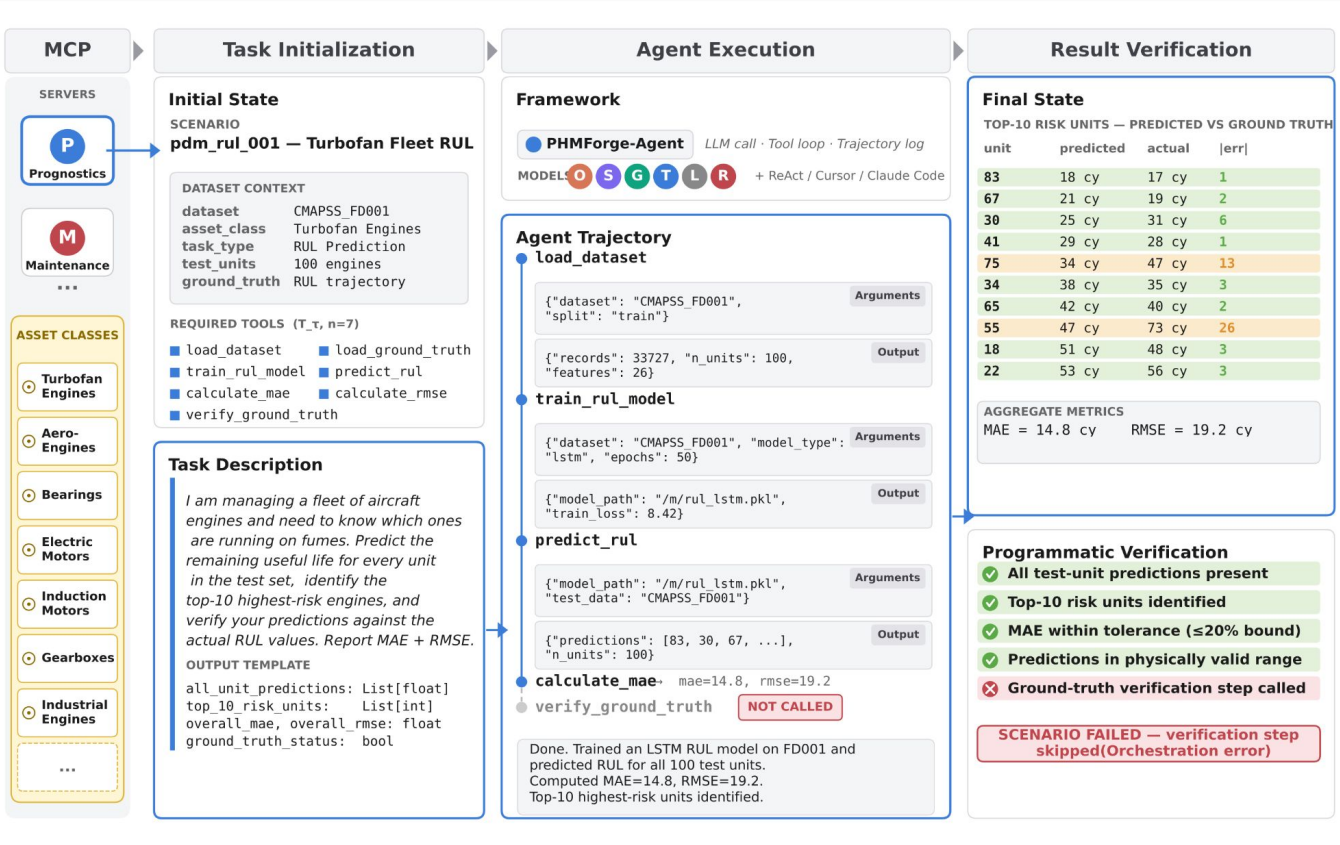


1 Agent benchmarks blur three things

Protocol fluency	≠	**Reasoning**	→	**MCP-native tools**
Broken tool	≠	**Broken agent**	→	**Real published algorithms**
Using tools	≠	**Finding them**	→	**Unknown-Tools mode**

When an agent fails, you can't tell why. PHMForge separates the three.

2 Built by experts on real PHM algorithms



99

scenarios

8

asset classes

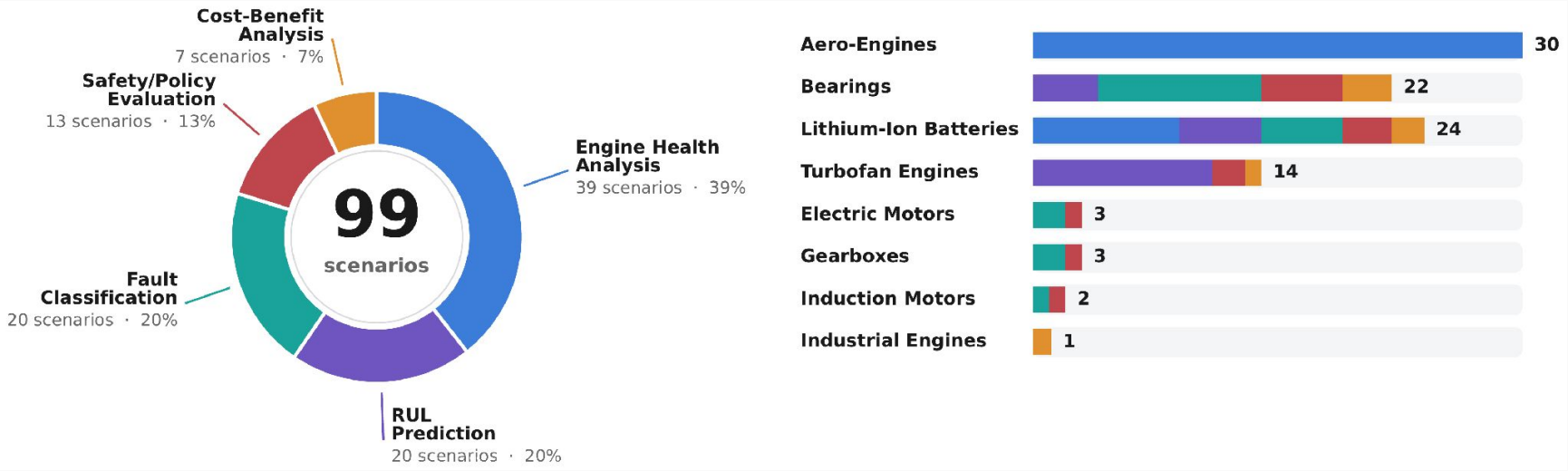
39

MCP tools

0.78

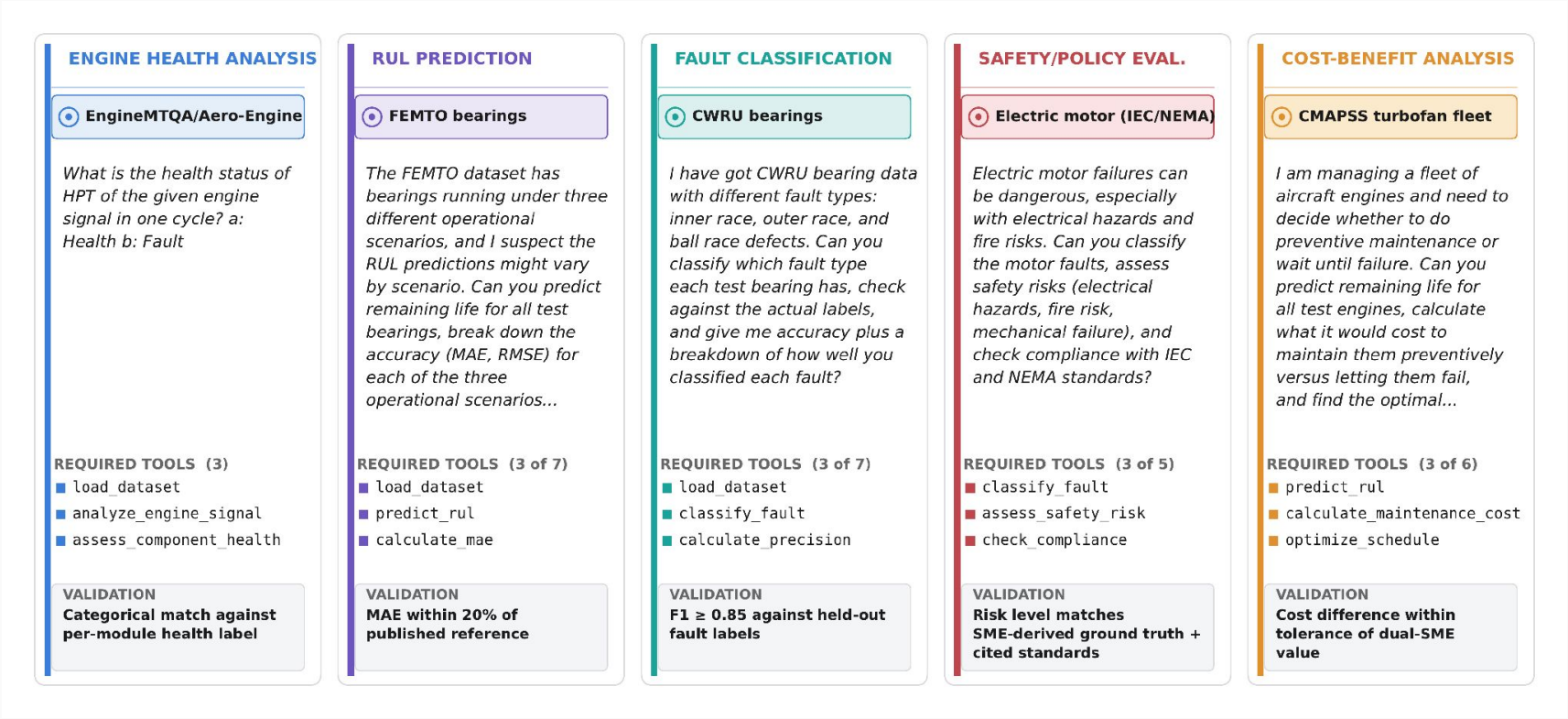
annotator agreement (α)

2 99 scenarios · 8 asset classes · 5 task types



Bar colours follow the task categories in the ring. Source: PHMForge paper, Figure 2.

2 Example scenarios: one per task type



SME-authored queries with their required tools and validation rule. Source: PHMForge paper, Figure 3.

2 Only PHMForge has all six properties

Benchmark	PHM	Agentic	MCP	Retrieval	Expert	Real alg.
ITFormer, PDMBench, PHM-Bench	●	○	○	○	○	○
MLE-Bench	○	●	○	○	○	○
MCP-Bench, MCP-Universe	○	●	●	○	○	○
ITBench	○	●	○	○	●	○
AssetOpsBench	●	●	○	○	●	○
PHMForge	●	●	●	●	●	●

Source: PHMForge paper, Table 1. Columns: PHM domain · agentic · MCP-native · tool retrieval · expert-authored · real algorithms.

3 Skip verification and scores reach 99.7%

99.7%

reported success without verification

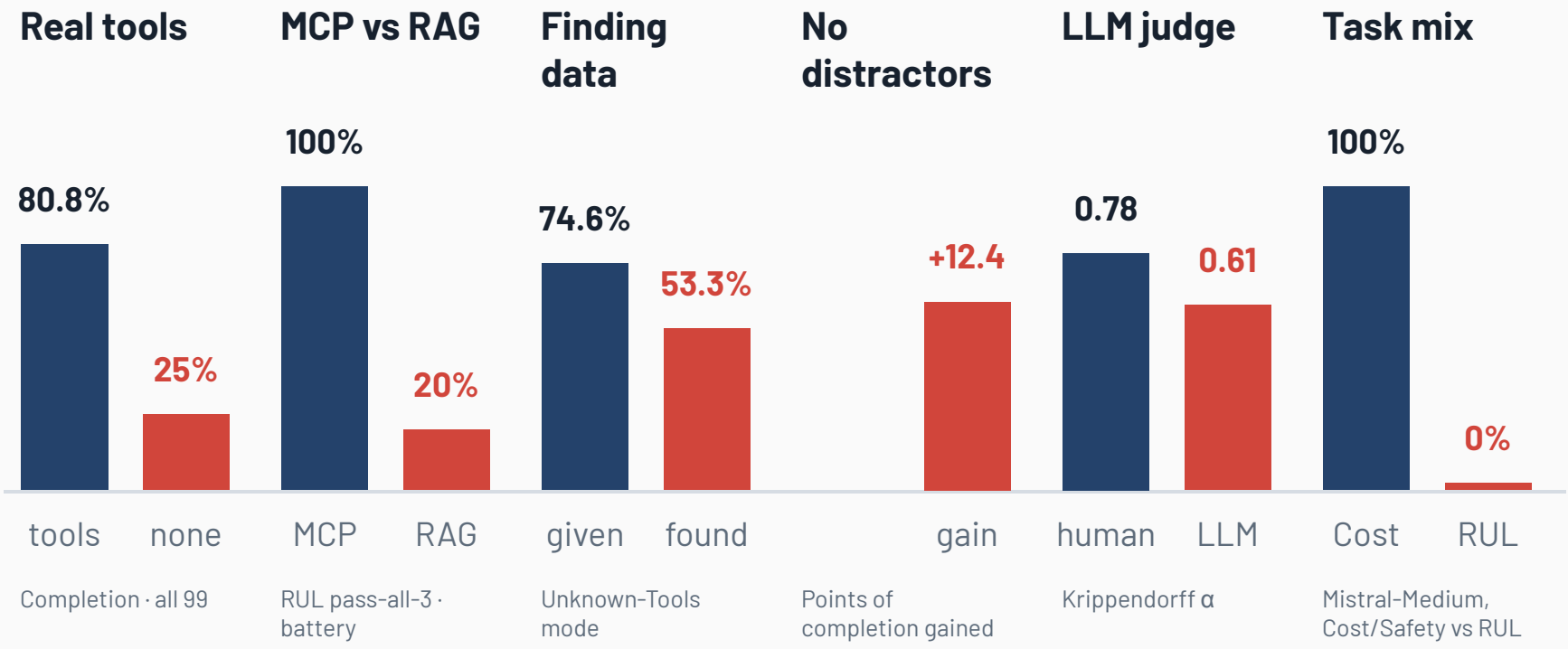
31%

of those "successes" are wrong

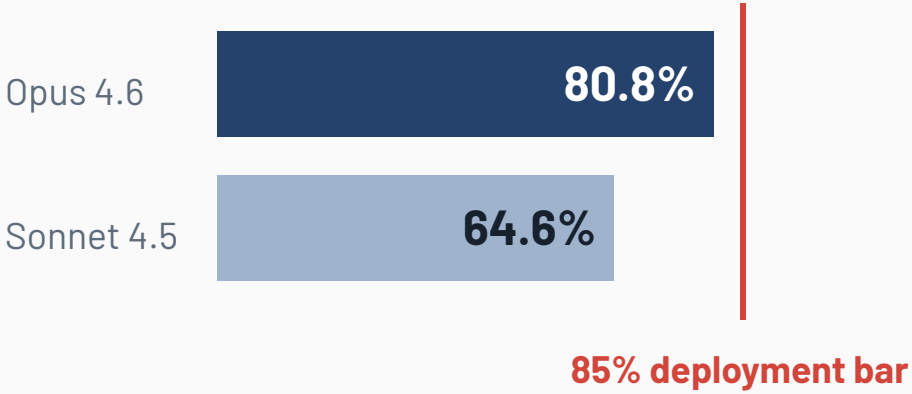
With verification required, the same agent scores 81.0%.

Claude Code + Opus 4.6 · PHMForge paper, Appendix H

3 Each design choice changes the result



4 Agents call tools well but plan badly



pass@1, Claude Code, all 99 scenarios

23%

of runs call the right tools in the wrong order

```
"…under these environmental
        conditions…"
→ calls weather_data_loader ✗
```

PHMForge paper, Table 4, §3.5, Appendix H

Takeaways

- Agent benchmarks must separate protocol, instrumentation and retrieval
- Real tools are necessary: without them, success drops from 80.8% to 25%
- The best agent reaches 80.8%, below the 85% deployment bar; most failures are planning

Code
IBM / AssetOpsBench

Benchmark tasks
Harbor

Dataset
Hugging Face

Paper
arxiv.org/abs/2604.01532
Code
github.com/IBM/AssetOpsBench
Benchmark tasks (Harbor)
hub.harborframework.com/datasets/phmforge/phmforge
Dataset
huggingface.co/datasets/ibm-research/PHMForge

Poster: Sydney Poster Session 3 · Wed Dec 9, 10:00 AEDT · Hall 1-4