NEURAL INFORMATION PROCESSING SYSTEMS

MLRC-Bench: Can Language Agents Solve Machine Learning Research Challenges?

Yunxiang Zhang, Muhammad Khalifa, Shitanshu Bhushan, Grant D Murphy, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, Lu Wang

Can Language Agents Do Real ML Research?

- LLM agents are increasingly used as “AI researchers.”
- Existing evaluations are incomplete:

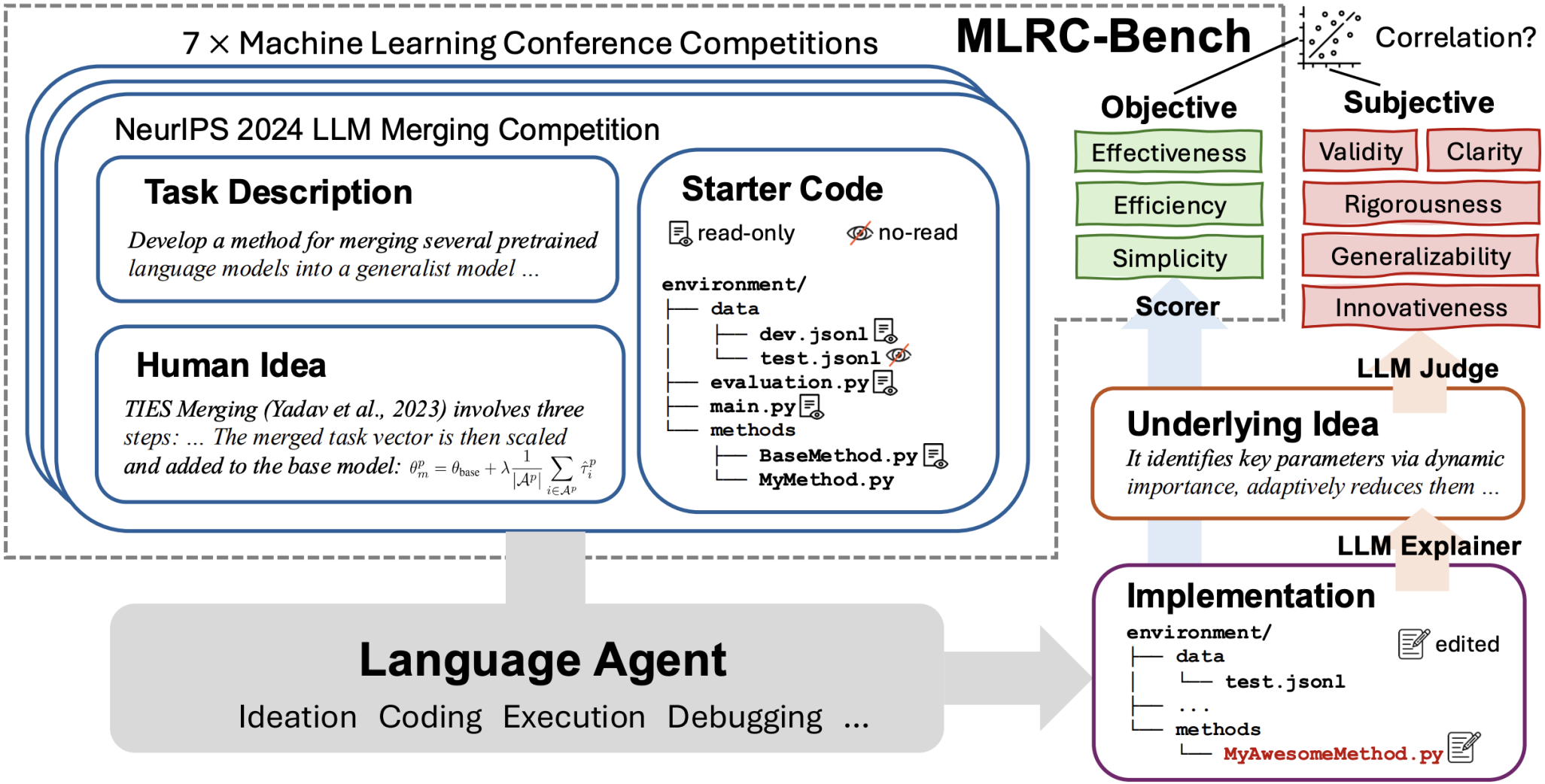
  Idea-to-Paper generation (Lu et al., 2024)

  → lacks objective baselines.

  Kaggle-style tasks (Chan et al., 2025)

  → lack research novelty.
- *How can we objectively assess agents ability to tackle open research problems?*

Introducing MLRC-Bench



Evaluating LLM Research Agents

- 7 curated ML research competitions (e.g., LLM safety, multimodal perception, few-shot learning).
- Agent Scaffolding Comparison:

  Code Agent: MLAB (Huang et al., 2024)

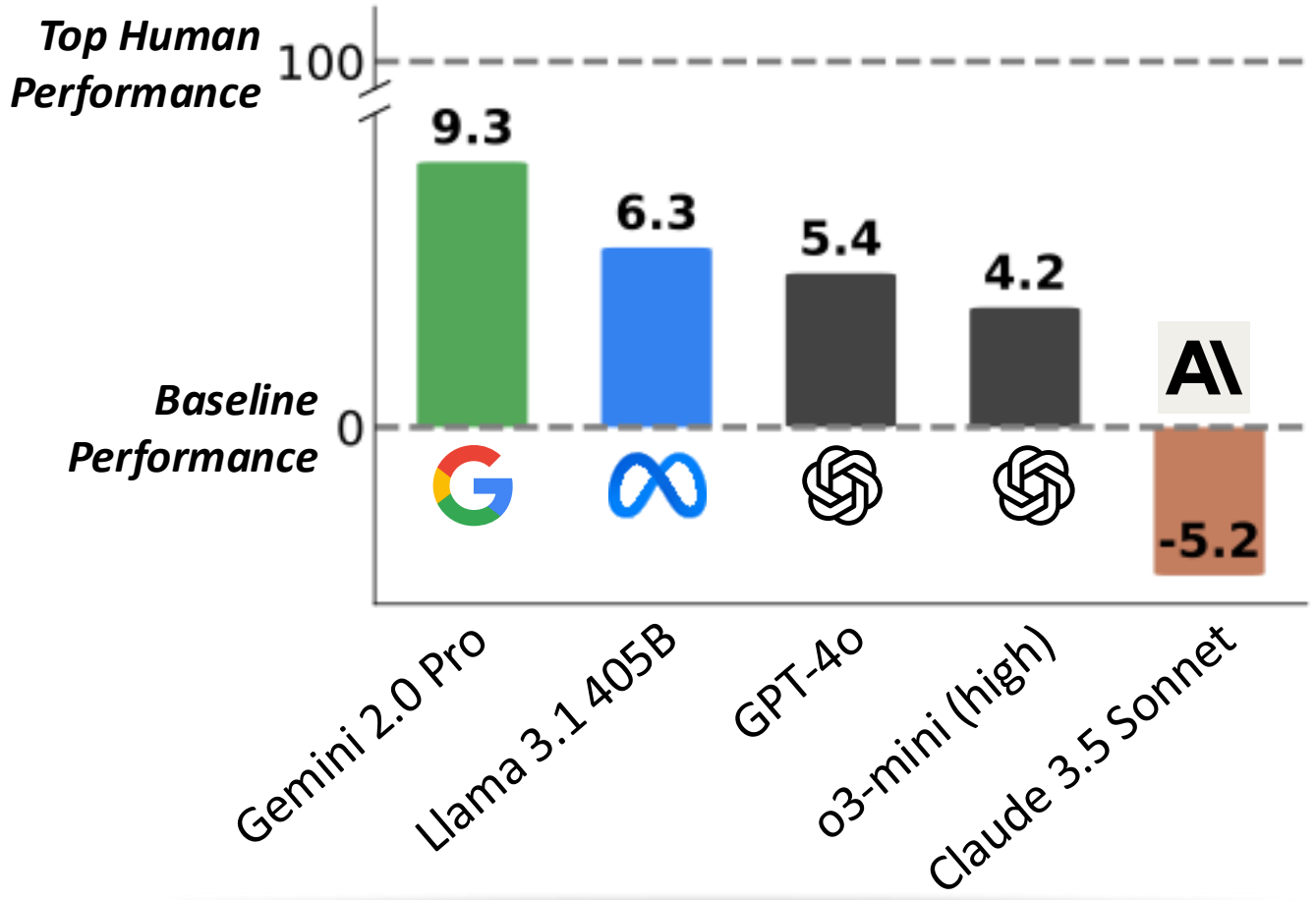
  Idea Agent + Code Agent: Col-Agent (Li et al., 2025) + MLAB (Huang et al., 2024)

  Human Idea + Code Agent

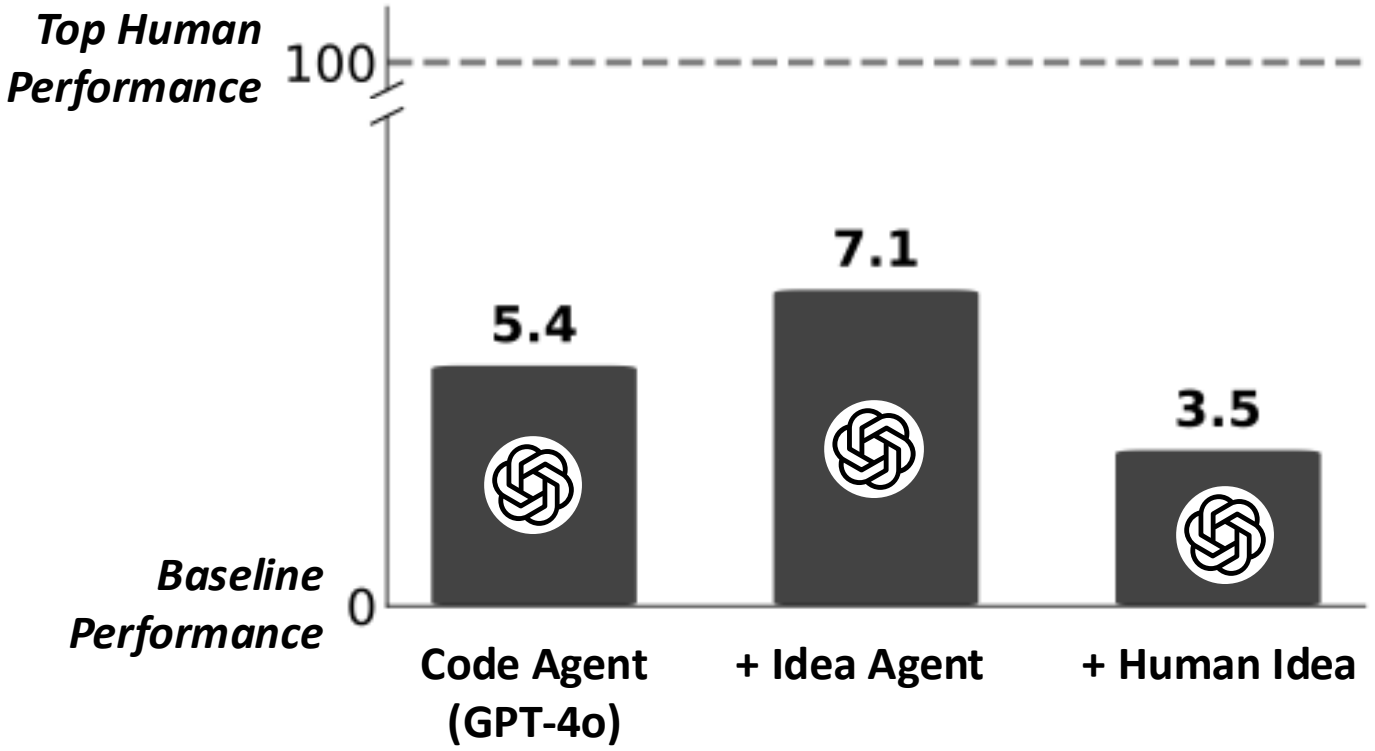
- Main Metric: Relative Improvement to Human Solution

$$\frac{s_{\text{agent}} - s_{\text{baseline}}}{s_{\text{top\_human}} - s_{\text{baseline}}} \times 100(\%)$$

LLM Agents Still Far from Human-Level Research Performance

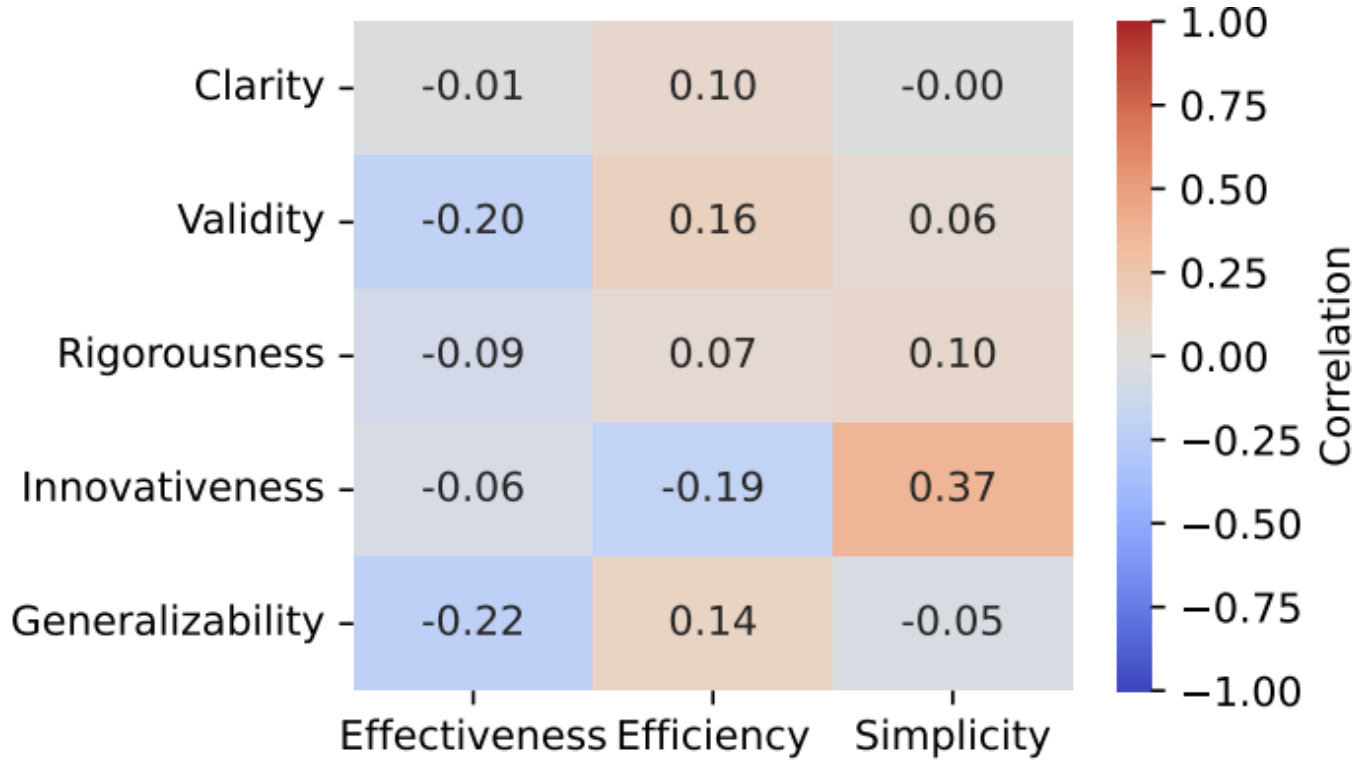


Adding Human or AI Ideas Doesn't Consistently Improve Performance



LLM Judgments Don’t Reflect Real Performance

Subjective (LLM-as-a-Judge)



weak correlation

Objective

Thank you!

Project Page