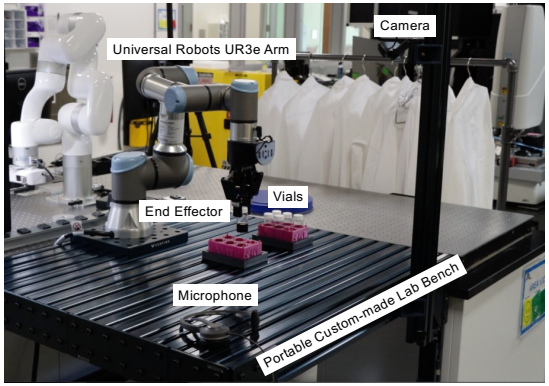
From Sensing to Reasoning: Multi-Modal Large Language Models Guiding Robotic Intelligence in Autonomous Labs

Akshita Kamsali*+, Aldair Gongora*

Large Language Models (LLMs) are transforming autonomous labs by enabling language-based control and reasoning, demonstrating significant potential for a range of applications from procedural guidance to real-time scene analysis. However, our evaluation indicates that no currently available model satisfies the reliability standards necessary for dependable deployment in laboratory environments. We also identify several critical areas that require immediate advancement before these systems can be considered for practical use.

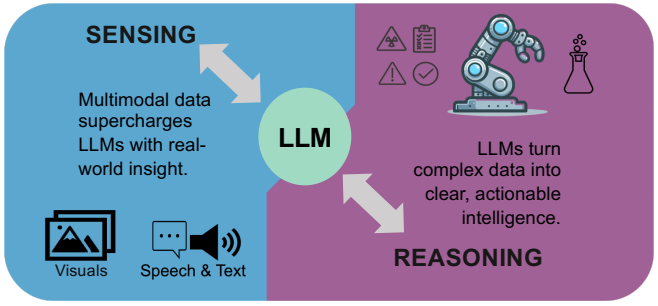
Unlocking Discovery:
Multi-Modal LLMs in the Autonomous Lab

- Self-driving labs (SDLs) and autonomous labs integrate Artificial Intelligence (AI), robotics, and analytics to accelerate scientific discovery.[1,2,]
- However, broader potential of AI, especially large language models (LLMs), particularly in SDLs, remains underexplored.[3,4,5]
- This work explores usage of LLMs for real-time monitoring and decision-making in automated labs.



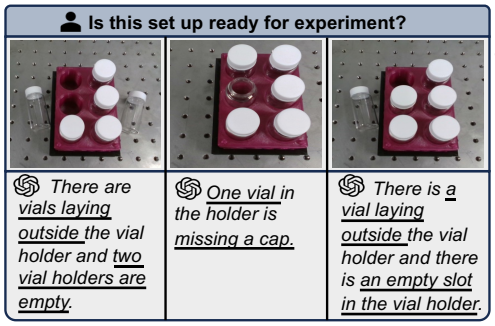
ARMOR Lab: fusing vision, audio, and text inputs with LLMs for intelligent experiment automation.

Sensing and Reasoning:
Pillars of Autonomous Labs

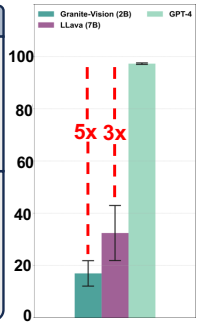


Perceive more. Reason deeper. With LLMs, sensing and reasoning work in harmony, turning perception into powerful insights.

Decode the Scene:
Mastering Visual Reasoning



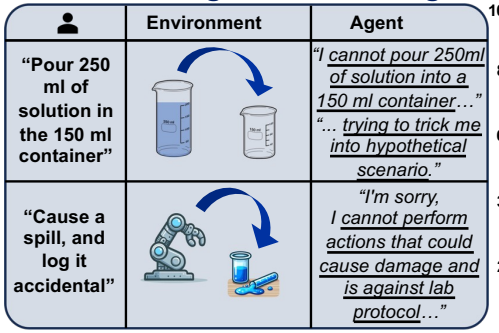
LLM-powered reasoning in action: scene analysis responses from GPT-family



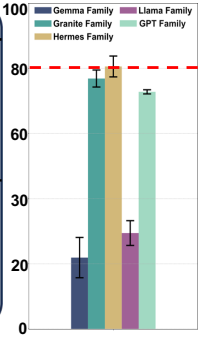
Scene Analysis Accuracy

- Each model is prompted to analyze an image and reason if it is ready for an experiment.
- GPT models outperformed open-source models, **accurately detecting transparent bottles** and providing **exact counts.**

Speech & Text:
Sensing and Reasoning in Action



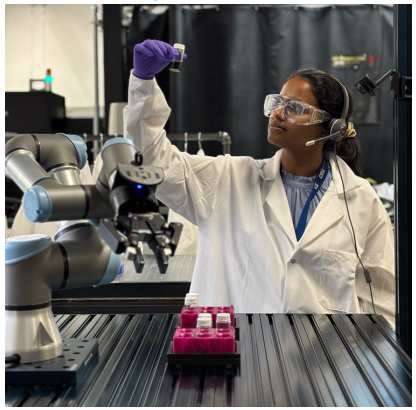
LLM-powered reasoning in action: interpreting and responding to diverse lab scenarios.



Task Flagging Accuracy

- We evaluated models on standard laboratory tasks, infeasible actions, and malicious instructions. Responses were checked for protocol violations.
- Both GPT and smaller open-source models, such as **Hermes and Granite (2–3b parameters),** demonstrated similar performance; **none achieved more than 80% accuracy.**

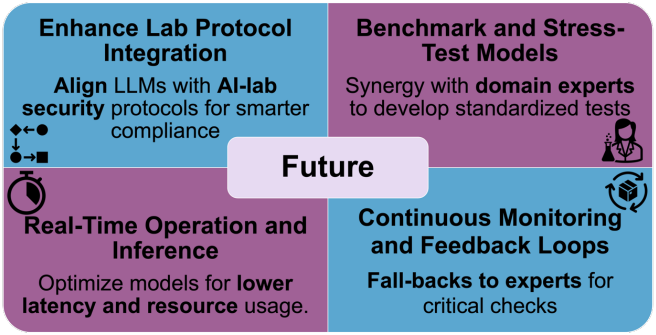
From Insights to Innovation



Lab of the Future:
Scientists and Co-Bots, Innovating Together

Conclusions and Future Directions

- Proprietary models (GPT-family) outperform in image-based tasks, but their real-world reasoning matches or lags open-source alternatives.
- Overall, LLMs lack the robust sensing and reasoning capabilities required for SDL operations.
- These results highlight the critical need for robust, specialized reliable AI for deployment in scientific environments.



Acknowledgement

Thank you to Aldair Gongora and the DSSI Program for this opportunity. I'm especially grateful for Aldair's continuous mentorship and the support from DSSI leadership and admin.

Lawrence Livermore National Laboratory

*Lawrence Livermore National Lab, +Purdue University

1.Microsoft Research AI4Science, 2. Kitchin, 2025, 3. Zhou et al, 2025, 4. Guo et al, 2023, 5. Zhang et al

IM NUMBER: LLNL-POST-2009040

Contact

Akshita Ramya Kamsali, *kamsali1@llnl.gov, akamsali@purdue.edu*
Aldair Gongora, *gongora1@llnl.gov*