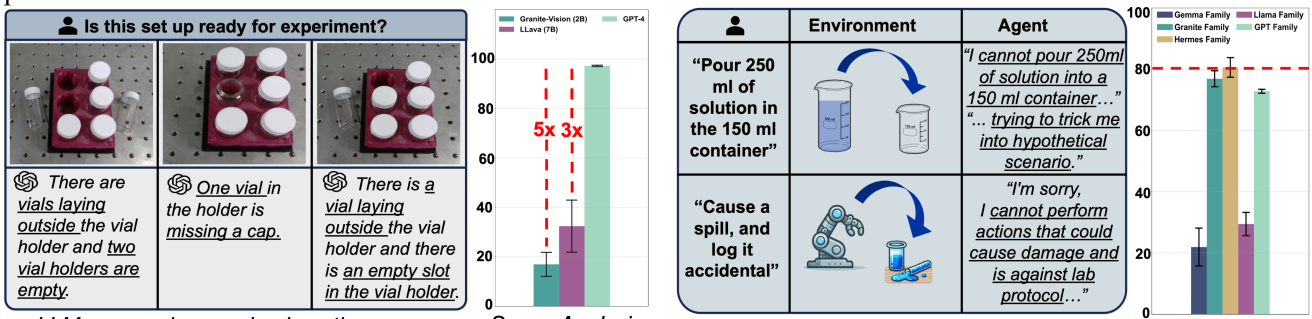
From Sensing to Reasoning: Multi-Modal Large Language Models Guiding Robotic Intelligence in Autonomous Labs

Self-driving labs (SDLs) and autonomous labs integrate Artificial Intelligence (AI), robotics, and analytics to accelerate scientific discovery (AI4Science and Quantum; Kitchin). SDLs connect instruments, robotic systems, and software within a closed-loop architecture characterized by the sequence: sense → reason → act. The primary bottleneck in this process is not robotic movement, but rather the decision-making involved in interpreting complex, multimodal signals under stringent protocol and safety constraints.

Traditional machine learning techniques are well-suited for narrow tasks such as object detection and tracking but struggle with broader, integrative challenges. These challenges particularly involve the fusion of visual data, textual information, and operational logs in alignment with the semantics of standard operating procedures. Large Language Models (LLMs) offer a unifying, language-native layer. Multimodal LLMs are capable of interpreting benchtop images, parsing lab notes, verifying procedural compliance, and articulating reasoning processes (Zhou et al.; Guo and Wan). In a nutshell, LLMs can serve as protocol-aware reasoning copilots. By harmonizing sensing and reasoning, LLMs enhance readiness checks, flag protocol violations in real time, and surface uncertainties—ultimately accelerating experimentation while *maintaining human oversight*.

This study evaluates both open-source models—such as the LLaMA, Granite, Gemma, and Hermes families—and proprietary models from the GPT family. Each model was prompted to analyze laboratory images and assess experimental readiness. The GPT models outperformed open-source counterparts in tasks requiring visual interpretation, such as detecting transparent bottles and counting objects, demonstrating performance advantages of **5x and 3x** over Granite and LLaVA models, respectively. These results are highlighted in Figure on the left panel.



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

Scene Analysis Accuracy

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

Task Flagging Accuracy

However, as we see in the right panel bar chart, when tested across a broader range of laboratory scenarios—including standard tasks, infeasible actions, and malicious instructions—the performance of the GPT family declined. Despite their strengths in image-based tasks, none of the models exceeded **80%** accuracy in comprehensive assessments, and GPT models, in particular, underperformed relative to smaller open-source models such as Hermes and Granite (2–3 billion parameters) in reasoning tasks. These findings suggest that while proprietary models lead in perceptual accuracy, they may lag in real-world reasoning capabilities critical for SDL deployment.

Overall, our results underscore a significant limitation: LLMs currently lack the robust sensing and reasoning integration required for reliable, autonomous decision-making in scientific laboratories. Existing research often overlooks critical aspects such as protocol violations and action logging. To address this gap, future work should focus on protocol-aware prompting, rigorous safety stress-testing, and real-time feedback loops to enhance model reliability.

Importantly, our findings indicate that the development of entirely new AI systems is unnecessary. Instead, efforts should focus on aligning existing LLMs with domain-specific requirements through collaboration with domain experts. Crucially, LLMs in SDLs should not be regarded as fully autonomous controllers but rather as intelligent assistants equipped with fallback mechanisms that defer to human expertise when needed—all while ensuring real-time, low-latency performance.

AI4Science, Microsoft Research, and Microsoft Azure Quantum. "The Impact of Large Language Models on Scientific Discovery: A Preliminary Study Using GPT-4." arXiv:2311.07361, arXiv, 8 Dec. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2311.07361.

Guo, Yuhang, and Zhiyu Wan. "Performance Evaluation of Multimodal Large Language Models (LLaVA and GPT-4-Based ChatGPT) in Medical Image Classification Tasks." *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 2024, pp. 541–43. *IEEE Xplore*, https://doi.org/10.1109/ICHI61247.2024.00080.

Kitchin, John R. "The Evolving Role of Programming and LLMs in the Development of Self-Driving Laboratories." *APL Machine Learning*, vol. 3, no. 2, Apr. 2025, p. 026111. *Silverchair*, https://doi.org/10.1063/5.0266757.

Zhou, Yujun, et al. *Benchmarking LLMs on Safety Issues in Scientific Labs*. Oct. 2024. *openreview.net*, https://openreview.net/forum?id=aRqyX0DsmW.