

Toward an Offline Generative AI Dialogue Model to Support the Automation of a Materials Synthesis Laboratory

Alberto Esteban Reyes Peralta¹, Francisco José López Cortés¹, Rigoberto Cerino Jiménez¹, David Pinto^{1,*}, Enrique Pérez Gutiérrez²

¹ Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

² Instituto de Ciencias, Unidad de Polímeros y Electrónica Orgánica,
Mexico

alberto.reyesp@alumno.buap.mx, franciscojlc21@gmail.com, cerino_rigoberto@hotmail.com,
{david.pinto, enrique.pgutierrez} @correo.buap.mx

Abstract. Materials synthesis laboratories face challenges due to repetitive and error-prone tasks, necessitating innovative solutions. This review explores the potential of large language models (LLMs) to develop an offline dialogue model based on generative artificial intelligence, designed to automate processes in materials synthesis laboratories through conversational interaction with robotic systems and human personnel. Through a systematic literature review from 2019 to 2025, 13 studies are analyzed across three categories: generative AI and language models, automation and robotics in laboratories, and automation and evaluation frameworks. The findings highlight improvements in efficiency (20–90%) and cost reduction (up to 84%) in robotic and scientific tasks but also point out limitations in accuracy (60–80%) and reproducibility (65%). Offline operation requires optimization for local hardware and data security. A hybrid model with human feedback is recommended to maximize impact in materials synthesis, addressing both technical and ethical challenges.

Keywords. Generative AI, large language models, laboratory automation, conversational robotics, materials synthesis, offline operation.

1 Introduction

Materials synthesis laboratories face significant challenges due to the repetitive and demanding nature of their tasks, which are time-consuming and prone to human error. The implementation of advanced technologies, such as automation and

generative artificial intelligence (AI), emerges as a promising solution to overcome these obstacles.

Generative AI enables more natural and efficient interaction between humans and automated systems, optimizing processes and allowing researchers to focus on creative and analytical tasks. Automation increases efficiency by reducing the time spent on repetitive activities and minimizing errors, thereby improving the reproducibility and quality of experimental results. Moreover, the integration of generative AI facilitates seamless communication with automated systems, representing a significant advancement over traditional approaches that rely on complex interfaces.

The development of an offline generative AI-based dialogue model is essential to ensure operational flexibility and protect sensitive data-critical aspects in scientific environments. Operating without an Internet connection eliminates security risks and ensures the confidentiality of information, meeting the specific needs of materials synthesis laboratories. This approach combines technological innovation with practicality and security, aligning with the current demands of scientific research.

The design of an offline dialogue model is grounded in three key areas of research. First, advances in laboratory automation have optimized repetitive tasks, such as sample preparation, enhancing both efficiency and reproducibility.

However, traditional systems lack the flexibility to handle non-predefined tasks, prompting the integration of AI to develop more interactive and autonomous robotic systems. Second, developments in generative AI, particularly large language models (LLMs), have enabled conversational interfaces and precise control of robotic systems through natural language instructions. Yet, their adaptation to technical and offline environments remains a challenge, requiring optimization for local hardware. Finally, AI-based dialogue models have made progress in generating technical responses, making them ideal for assisting in scientific processes. However, operating offline necessitates optimizing these models for limited data and local hardware while ensuring accuracy in specialized domains such as materials synthesis.

This paper aims to support the design and implementation of an offline generative AI-based dialogue model for automating processes in materials synthesis laboratories, enabling conversational interaction with robotic systems and human personnel. Relevant literature in generative AI, robotic automation, and evaluation frameworks is analyzed, regardless of whether the studies are specifically focused on materials synthesis, to inform and enhance the proposed model.

2 Methodology

A systematic literature review was conducted to identify relevant studies for the design of an offline generative AI dialogue model applied to the context of materials synthesis laboratories. The search included terms such as offline generative AI, laboratory automation, conversational robotics, and language models. Articles published between 2019 and 2025 in English were selected, focusing on generative AI models, robotic automation, or evaluation frameworks with potential applicability to the project, even if they were not specifically related to materials synthesis. The selected articles were organized into three categories: Generative Artificial Intelligence and Language Models, Laboratory Automation and Robotics, and Automation and Evaluation Frameworks. The findings are synthesized in tables to compare

applications, performance metrics, and key challenges.

3 Detailed Analysis by Category

3.1 Generative Artificial Intelligence and Language Models

Research in generative artificial intelligence and large language models (LLMs) demonstrates their potential to generate technical responses and assist in scientific processes, with direct applications to the design of an offline dialogue model for materials synthesis laboratories. A study by Negrini and Lippi evaluates generative AI systems in responding to queries about the correlation between troponin and myocardial infarction, reporting a sensitivity of 93% and specificity of 88% in troponin diagnostics.

However, most systems incorrectly describe cardiac troponin as a single protein rather than a protein complex, highlighting limitations in scientific accuracy (Negrini & Lippi, 2023).

On the other hand, Schmidgall et al. present Agent Laboratory, an autonomous framework that automates research processes, from literature review to solving machine learning problems, achieving success rates of 92.8–95.7% in subtasks. The mle-solver component earned four medals in MLE-Bench, outperforming other methods. However, it faces challenges in literature review tasks (60–80% success) and self-evaluation, which limits the quality of generated reports (Schmidgall et al., 2025).

Similarly, Zimmermann et al. analyze a hackathon that showcases the usefulness of LLMs in materials science, covering tasks such as hypothesis generation, data extraction, interface design, and results analysis. Their rapid prototyping capabilities within 24 hours improve research efficiency and creativity; however, the authors note reproducibility challenges due to reliance on proprietary APIs and high computational infrastructure demands (Zimmermann et al., 2025).

Finally, Kahng et al. introduce GAN Lab, an interactive tool based on TensorFlow.js for visualizing Generative Adversarial Networks. It enables users to understand GAN training

Table 1. Generative AI and Language Models

Paper	Application	Contribution	Metrics	Challenges
[9]	Scientific writing	Troponin diagnostics	93% sensitivity	Technical accuracy
[10]	Research assistance	Agents Laboratory	92.8-95.7% success	Self-assessment
[13]	Materials science	Rapid prototyping	Efficiency in 24 hours	Reproducibility
[7]	Visualisation GAN	GAN Lab	Training balance	Technical accessibility

Table 2. Laboratory Automation and Robotic

Paper	Application	Contribution	Metrics	Challenges
[1]	Robotic control	PRM Architecture	66% precision	Adaptability prompts
[3]	Laboratory automation	Robotic arm with LLM	1.9-2.8x reduction time	Self-assessment
[5]	Biological scripts	Scripts for OT-2	90% success	Ambiguous instructions
[11]	Robotics review	LLM applications	20-40% improvement	Generalization
[12]	Robotic manipulation	Generative models	Data efficiency	Technical accessibility

dynamics, such as the balance between discriminator and generator learning rates and explore issues like mode collapse. This tool inspires the development of intuitive interfaces for offline systems in laboratory environments (Kahng et al., 2019).

3.2 Automation and Robotics in Laboratories

The integration of large language models into robotic systems for laboratories presents promising solutions for automating materials synthesis by enabling conversational interactions and precise task control.

Benjdira et al. introduce the Prompting Robotic Modalities (PRM) architecture, which utilizes Modality Language Models (MLMs) to handle specific data types (e.g., image, sound, actuators), coordinated by a central LLM that orchestrates complex tasks through natural language. This architecture is validated by ROSGPT_Vision, which achieves 66% accuracy in monitoring driver attention, and by CarMate, an application that reduces development time and cost by enabling rapid adaptation through prompts (Benjdira et al., 2025).

Similarly, Conrad et al. present a laboratory automation system that integrates a robotic arm with an LLM, achieving a 1.9–2.8x reduction in the processing time of repetitive tasks through meta-tools for dynamic recording and playback of sequences. Their human-in-the-loop collaborative

approach ensures adaptability to changing experimental conditions (Conrad et al., 2025).

Inagaki et al. develop a pipeline that converts natural language descriptions into Python scripts for the OT-2 liquid-handling robot. The system achieves a 95% success rate with step-by-step instructions and 71% with goal-oriented instructions using GPT-4, even when given ambiguous prompts, eliminating the need for programming expertise (Inagaki et al., 2023).

Additionally, Zeng et al. review LLM applications in robotics, highlighting their impact on perception, decision-making, control, and interaction. Models such as PaLM-SayCan and PaLM-E improve efficiency by 20–40%, though they face challenges in training and generalization (Zeng et al., 2023).

Finally, Zhang et al. explore generative models such as GANs and diffusion models in robotic manipulation, organized into Foundation, Intermediate, and Policy layers. They address challenges in data efficiency and long-horizon task planning, with applications in grasp and trajectory generation (Zhang et al., 2025).

3.3 Automation and Evaluation Frameworks

Automation and evaluation frameworks are essential for validating and optimizing LLM-based systems, particularly in offline environments such as materials synthesis laboratories.

Table 3. Automatization and Evaluation frameworks

Paper	Application	Contribution	Metrics	Challenges
[2]	Automated testing	LLMAutoE2E	95% coverage	Security and bias
[4]	Human behavior	Economic markets	70% similarity	Heterogeneity
[8]	Expressive behaviours	GenEM	Competent behaviours	Multi-shift Interactions
[6]	Autonomous AI security	Scorecards	Risk assessment	Standardisation

Chandra introduces LLMAutoE2E, a framework for automated LLM testing that streamlines the generation and execution of test cases, code review, defect detection, and report generation. It achieves 95% test coverage and 80% bias detection, outperforming manual methods in efficiency and reducing the need for human intervention (Chandra, 2025).

Del Río-Chanona et al. explore the use of LLMs to replicate human behavior in economic market experiments, achieving 70% similarity with human behavior by capturing patterns such as positive and negative feedback markets. However, the models exhibited less behavioral heterogeneity compared to humans, highlighting the need to improve conversational diversity in responses (Del Río-Chanona et al., 2025).

Mahadevan et al. present Generative Expressive Motion (GenEM), an approach that uses few-shot prompts to translate natural language instructions into robotic motion code. The system delivers competent and understandable behaviors in user studies, although it is limited to single-turn interactions (Mahadevan et al., 2024).

Finally, Jabbour and Reddi propose a scorecard framework to assess security risks in autonomous systems using generative AI. It classifies applications such as anomaly detection (high risk) and explainability (low risk), offering a standardized method for communicating risks at different levels of the autonomous system (Jabbour & Reddi, 2024).

4 Discussion

The reviewed literature highlights the transformative potential of large language models (LLMs) for automation in materials synthesis laboratories, particularly in the design of an offline dialogue model. In the field of robotics, studies demonstrate that LLMs enable smooth

conversational interactions and precise control of robotic systems, with applications such as script generation for lab robots (95% success rate in Inagaki et al.), accurate manipulation, and architectures like PRM and systems with meta-tools that reduce processing time by up to 2.8 times.

In scientific research, LLMs accelerate processes such as technical writing and experimental design, achieving high success rates (92.8–95.7%) in automated tasks, as evidenced in the work by Schmidgall et al. However, limitations persist in accuracy (e.g., 93% sensitivity but with conceptual errors in troponin as reported by Negrini and Lippi) and reproducibility, particularly due to reliance on proprietary APIs and the high demands of computational infrastructure, as highlighted in Zimmermann et al.

Evaluation frameworks such as Chandra's LLMAutoE2E offer robust strategies for validating offline models, while tools like GAN Lab inspire the development of intuitive interfaces. The evaluation by Del Río-Chanona et al. underscores LLMs' capacity to simulate human behavior in economic contexts, although with reduced heterogeneity, suggesting the need to integrate factors such as demographics or personality traits to improve conversational diversity.

Mahadevan et al.'s GenEM approach demonstrates the feasibility of generating expressive robotic behaviors from natural language instructions, although its limitation to single-turn interactions suggests the need to explore multi-turn dynamics. Jabbour and Reddi's proposal of a scorecard framework for autonomous systems with generative AI emphasizes the importance of standardizing risk assessment, especially in critical applications such as anomaly detection.

For the proposed model, the findings suggest that LLMs can process natural language

instructions to interact with robotic systems and human personnel. However, optimization for local hardware and accuracy in technical domains are crucial. In addition, ensuring security and reliability through rigorous evaluation frameworks is essential to protect sensitive data in offline environments.

Key challenges include accuracy in technical tasks, generalization to dynamic environments, reproducibility, and optimization for local hardware. These challenges may be mitigated through human-in-the-loop collaborative approaches and thorough evaluation strategies.

5. Proposed Methodology for Training a Conversational Model in Laboratory Settings

5.1 Context and Purpose

The increasing complexity of experimental workflows in materials synthesis laboratories highlights the need for intelligent systems capable of assisting in routine tasks, minimizing human error, and improving reproducibility. In future work, we propose to address this challenge by developing and training an offline conversational model designed to facilitate natural and efficient communication between human operators and collaborative robots, such as the ABB Yumi IRB 14000.

The aim of this future effort is to enable the robot to interpret and execute verbal instructions related to essential laboratory tasks, including environmental observation, substance handling, instrument operation, ambiguity resolution, and real-time failure response. Rather than relying on specialized interfaces or manual scripting, the proposed dialogue model would provide an intuitive, speech-driven interaction framework tailored to the needs of scientific personnel.

To ensure security, usability, and adaptability in sensitive research environments, the model will be grounded in domain-specific language and optimized for local deployment. This approach is expected to contribute to the development of secure and autonomous robotic systems that support flexible and reliable automation in laboratory settings.

5.2 Dialogue Corpus Design

As a first step, the construction of a simulated dialogue corpus is planned to replicate common laboratory scenarios. These dialogues would include realistic turns between a human and the robot, manually labeled with elements such as:

- Communicative intent: observation, instruction, inquiry, confirmation.
- Referenced entities: instruments, substances, quantities, units.
- Environmental conditions: states of devices or materials.
- Expected response type: direct action, clarification request, warning.

The corpus would be designed to represent different levels of complexity and serve as a foundation for training and evaluating the model.

Some examples of the expected dialogues follows:

Human Instruction	Yumi Response
Yumi, what objects or instruments do you observe in your environment?	I detect five main instruments on the workbench: 1. A 250 ml beaker 2. A 100 ml Erlenmeyer flask 3. An automatic pipette 4. A bottle labeled 'blue solution' 5. An analytical balance to my left
Can you confirm if the balance is on?	Yes, the analytical balance is on and in standby mode. Current reading: 0.000 g.
Yumi, locate the vial labeled 'blue solution' and check its current volume.	Located. The vial labeled 'blue solution' contains approximately 47 ml according to the level sensor.
Add 10 ml of that solution to the beaker in the center of the table.	Starting transfer. [...5 seconds later...] Transfer completed. 10 ml of blue solution has been added to the beaker.
Do you observe any color change or precipitate?	Yes. After the addition, the solution changed from light blue to purple. No precipitate observed.
Yumi, take a sample from Erlenmeyer flask B and measure its pH.	Taking sample... [...3 seconds later...] Measurement completed. The pH of the solution is 3.8.
Place that flask on heating plate 1 and heat it until it reaches 60 °C.	Placing flask B on heating plate 1. Starting heating. Current temperature: 24.2 °C. [...] Target temperature reached: 60.0 °C.

Maintain that temperature for 5 minutes and monitor any changes in viscosity.	Starting timer. Monitoring viscosity... [...] At minute 5: slight decrease in viscosity observed. Solution is more fluid.
Yumi, add 20 ml of the red solution from the second vial to the 100 ml flask.	Warning: the second vial labeled 'red solution' only contains 12 ml. Do you want to proceed with the available amount?
Yes, add the 12 ml to the flask.	Transfer in progress... Transfer completed. 12 ml of red solution added to the flask.
Note this adjustment as 'Test 4B - limited volume'.	Note recorded: Test 4B - limited volume (12 ml).
Yumi, add acid to the beaker.	Clarification needed: there are three bottles labeled 'acid' in the environment. Should I use 0.1 M hydrochloric acid, 1 M nitric acid, or pure acetic acid?
Use 1 M nitric acid.	Amount to transfer?
5 ml.	Adding 5 ml of 1 M nitric acid to the beaker... Transfer completed. Caution: beaker temperature slightly increased after addition.
Yumi, weigh 0.250 g of sodium carbonate on the analytical balance and transfer it to the beaker.	Attempting to weigh... Error: the analytical balance is unresponsive. Display shows 'ERROR CAL'.
Can you check if it is properly connected?	Yes. The balance is connected to power, but requires recalibration. I recommend maintenance or using the secondary balance.
Use the secondary balance, please.	Weighing with secondary balance... Exact weight: 0.251 g. Would you like to adjust or keep this value?
That's fine, proceed with the transfer.	Transfer to beaker completed.

5.3 Model Architecture

The proposed architecture integrates components from both machine learning and symbolic reasoning to create a robust and context-aware conversational system for laboratory environments. At its core, the system relies on a pretrained language model, such as GPT-4 turbo, which will be fine-tuned through supervised learning using the domain-specific dialogue corpus previously constructed. This adaptation process aims to tailor the model's responses to the specialized vocabulary, typical scenarios, and

procedural logic of a materials synthesis laboratory.

Complementing the language model, a contextual memory module will be implemented. This module is responsible for continuously tracking and updating the state of the laboratory environment, including information about active instruments, substance quantities, and previous dialogue history. By maintaining this dynamic representation of context, the model can generate more accurate, coherent, and situationally appropriate responses.

In addition, a symbolic validation system will serve as a critical safety and consistency layer.

Before any physical action is executed by the robot, this component will validate the proposed command against predefined logical rules and safety constraints. This ensures that instructions issued through natural language are not only syntactically correct but also operationally safe and aligned with laboratory protocols.

Together, these elements form a hybrid architecture that leverages the expressive capabilities of generative language models while grounding them in real-world constraints and procedural logic through symbolic reasoning and environmental awareness.

5.4 Training Stage

The training process for the proposed conversational model would be carried out in several carefully structured stages, each aimed at ensuring linguistic precision, domain relevance, and functional robustness within a laboratory setting.

Preprocessing of the Dialogue Corpus

The initial stage involves preparing the collected dialogue data to ensure consistency and usefulness for training. This includes the normalization of technical terms and measurement units, making sure that variations in terminology do not introduce ambiguity into the learning process.

The data is then subjected to tokenization and manual annotation, where each utterance is labeled with its communicative intent (such as instruction, observation, or clarification) and relevant entities (such as instruments, substances,

and quantities). In addition, the dataset is enriched by the generation of controlled ambiguity and error cases, simulating real-world situations where users may provide vague or incorrect instructions. This enables the model to learn how to identify and handle such scenarios effectively.

Supervised Fine-Tuning

Once the corpus is preprocessed, the language model undergoes supervised fine-tuning. During this phase, the model is trained using backpropagation with a cross-entropy loss function, allowing it to optimize its predictions based on the annotated dialogue turns. The training process includes intermediate validation, using performance metrics such as intent recognition accuracy, BLEU score (to assess the quality of generated responses), and response accuracy. These evaluations help track the model's learning curve and prevent overfitting.

Integration of Environmental Context

To ground the model in real laboratory dynamics, a simulation of the lab environment is developed.

This simulated space includes virtual representations of laboratory instruments, substances, and user actions, allowing for controlled testing of the model's ability to retain and utilize contextual information. Experiments are conducted to evaluate how well the model can recover and apply contextual data, for example, remembering the status of an ongoing procedure or correctly identifying which instrument is in use, across multiple conversational turns.

Evaluation through Cross-Validation

In the final stage, a thorough evaluation of the model's generalizability is carried out through cross-validation. The dialogue corpus is split into distinct training, validation, and test sets, ensuring that the model is assessed on data it has not previously encountered. This phase focuses on measuring the model's performance during unseen interactions, evaluating its ability to correctly interpret new commands, manage unexpected situations, and maintain coherent dialogue with the user.

Together, these stages form a comprehensive training pipeline designed to produce a conversational agent capable of operating safely, efficiently, and intelligently in the context of laboratory automation.

5.5 Interactive Evaluation

Finally, the conversational model is expected to be implemented in either a physical or simulated environment using the YuMi collaborative robot.

This implementation is intended to support an interactive evaluation that mirrors real-world laboratory conditions and allows for a thorough assessment of the model's effectiveness in practical use.

The evaluation will focus on three main aspects. First, it will assess the model's ability to accurately interpret and execute verbal instructions provided by a human operator. These instructions may involve complex laboratory tasks such as manipulating substances, operating instruments, or making observations about the environment. The success rate of these interactions will offer insight into the model's operational reliability and language comprehension.

Second, the system's responsiveness to ambiguous or erroneous input will be examined. In laboratory settings, miscommunication or vague instructions are common, and the ability of the model to identify these issues and respond appropriately by asking for clarification or correcting its course of action is critical. The evaluation will measure how quickly and effectively the system can recover from such situations.

Third, subjective feedback will be collected from human users through usability surveys. These surveys will provide a qualitative perspective on the model's performance, capturing the perceptions of laboratory personnel regarding its ease of use, clarity of responses, and overall utility in supporting their work. This human-centered dimension is essential for ensuring that the system is not only technically capable but also intuitive and beneficial in real-world applications.

This final phase aims to validate both the semantic understanding and practical performance of the conversational model. By embedding it within an experimental environment and engaging

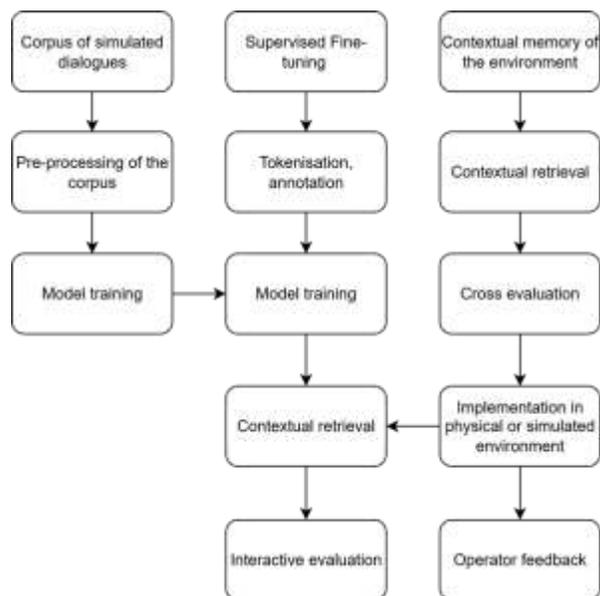


Fig. 1. Proposed Methodology for Training a Conversational Model in Laboratory Environments.

in direct interaction with users and robotic systems, the evaluation will offer a comprehensive view of the model's strengths, limitations, and readiness for deployment in laboratory automation. The full methodology proposed, including this evaluation phase, is illustrated in Figure 1.

6 Conclusions

This review confirms that large language models (LLMs) are versatile tools for automating processes in materials synthesis laboratories, enabling natural conversational interaction and supporting offline operation. Their potential to streamline routine tasks, reduce human error, and enhance reproducibility makes them highly valuable in experimental environments. However, challenges related to domain-specific accuracy, reproducibility of responses, and operational safety must be addressed to enable their broader and more responsible adoption.

To this end, the proposed methodology outlines a structured, multi-stage approach to training and deploying an offline conversational model tailored to laboratory needs. It begins with the design and annotation of a dialogue corpus that captures

realistic human-robot interactions, including varying levels of instruction complexity and ambiguity. This is followed by supervised fine-tuning of a pretrained language model, using the annotated corpus to align the system's responses with the linguistic and procedural demands of laboratory operations. A contextual memory module is integrated to maintain real-time awareness of the laboratory environment, while a symbolic validation layer ensures that generated commands are safe and consistent before execution by the robot.

Further stages include the development of a simulated lab environment to test contextual recall and real-time dialogue coherence, as well as rigorous cross-validation to assess performance on unseen interactions. Finally, the model is deployed in a physical or simulated setup using the ABB YuMi robot, where it is evaluated based on task success rates, ambiguity handling, and human feedback through usability surveys.

Considering these findings and the methodological framework proposed, it is recommended to pursue the development of a hybrid system that combines the generative capabilities of LLMs with human-in-the-loop supervision and experimental validation. Such a system would help bridge the gap between AI-driven language understanding and the practical, safety-critical requirements of scientific laboratories. This approach ensures not only the functional integration of automation and generative AI but also promotes responsible deployment in sensitive, high-stakes research environments like materials synthesis.

References

1. **Benjdira, B., Koubaa, A., Ali, A. (2025).** Prompting Robotic Modalities (PRM): A structured architecture for centralizing language models in complex systems. *Future Generation Computer Systems*. Vol. 166, pp. 107723. doi: <https://doi.org/10.1016/j.future.2025.107723>.
2. **Chandra, R. (2025).** Automation Frameworks for End-to-End Testing of Large Language Models (LLMs). *Journal of Information Systems Engineering and Management*. Vol.

- 10, pp. 464–472. doi: <https://doi.org/10.52783/jisem.v10i43s.8400>.
3. **Conrad, S., Auth, P., Masselter, T., Speck, T. (2025).** Lowering the Entrance Hurdle for Lab Automation: An Artificial Intelligence Supported, Interactive Robotic Arm for Automated, Repeated Testing Procedures. *Advanced Intelligent Systems*. doi: <https://doi.org/10.1002/aisy.202401086>.
 4. **Rio-Chanona, R., Pangallo, M., Hommes, C. (2025).** Can Generative AI agents behave like humans? Evidence from laboratory market experiments. doi: <https://doi.org/10.48550/arXiv.2505.07457>.
 5. **Inagaki, T., Kato, A., Takahashi, K., Ozaki, H., Kanda, G. (2023).** LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation. doi: <https://doi.org/10.48550/arXiv.2304.10267>.
 6. **Jabbour, J., Janapa Reddi, V. (2025).** Generative AI Agents in Autonomous Machines: A Safety Perspective. *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*. doi: <https://doi.org/10.1145/3676536.3698390>.
 7. **Kahng, M., Thorat, N., Chau, D., Viégas, F., Wattenberg, M. (2019).** GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*. 25, 310–320. doi: <https://doi.org/10.1109/TVCG.2018.2864500>.
 8. **Mahadevan, K., Chien, J., Brown, N., Xu, Z., Parada, C., Xia, F., Zeng, A., Takayama, L., Sadigh, D. (2024).** Generative Expressive Robot Behaviors using Large Language Models. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. pp. 482–491. doi: <https://doi.org/10.1145/3610977.3634999>.
 9. **Negrini, D., Lippi, G. (2023).** Generative Artificial Intelligence in (laboratory) medicine: friend or foe? *Biochimica Clinica*. doi: https://doi.org/10.19186/BC_2023.025.
 10. **Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., Barsoum, E. (2025).** Agent Laboratory: Using LLM Agents as Research Assistants. doi: <https://doi.org/10.48550/arXiv.2501.04227>.
 11. **Zeng, F., Gan, W., Wang, Y., Liu, N. & Yu, P. (2023).** Large Language Models for Robotics: A Survey. doi: <https://doi.org/10.48550/arXiv.2311.07226>.
 12. **Zhang, K., Yun, P., Cen, J., Cai, J., Zhu, D., Yuan, H., Zhao, C., Feng, T., Wang, M., Chen, Q., Pan, J., Zhang, W., Yang, B., Chen, H. (2025).** Generative Artificial Intelligence in Robotic Manipulation: A Survey. doi: <https://doi.org/10.48550/arXiv.2503.03464>.
 13. **Zimmermann, Y., Bazgir, A., Al-Feghali, A., Ansari, M., Bocarsly, J., Brinson, L., Chiang, Y., Circi, D., Chiu, M., Daelman, N., Evans, M., Gangan, A., George, J., Harb, H., Khalighinejad, G., Khan, S., Klawohn, S., Lederbauer, M., Mahjoubi, S., Mohr, B., Moosavi, S., Naik, A., Ozhan, A., Plessers, D., Roy, A., Schöppach, F., Schwaller, P., Terboven, C., Ueltzen, K., Wu, Y., Zhu, S., Janssen, J., Li, C., Foster, I., Blaiszik, B. (2025).** 34 Examples of LLM Applications in Materials Science and Chemistry: Towards Automation, Assistants, Agents, and Accelerated Scientific Discovery. doi: <https://doi.org/10.48550/arXiv.2505.03049>.

Article received on 12/06/2025; accepted on 11/10/2025.

*Corresponding author is David Pinto.