Automated Atomic Force Microscopy Using Large Language Models

Indrajeet Mandal[1], **Jitendra Soni**[2], **Mohd Zaki**[3], **Morten M Smedskjaer**[4]
Katrin Wondraczek[5], **Lothar Wondraczek**[6],
Nitya Nand Gosvami[2,7*], **N. M. Anoop Krishnan**[3,7*]
[1]School of Interdisciplinary Research, Indian Institute of Technology Delhi, New Delhi, India
[2]Department of Materials Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India
[3]Department of Civil Engineering, Indian Institute of Technology Delhi, New Delhi, India
[4]Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark
[5]Leibniz Institute of Photonic Technology, Jena, Germany
[6]Otto Schott Institute of Materials Research, University of Jena, Germany
[7]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, India
{ngosvami, krishnan}@iitd.ac.in

Abstract

Atomic force microscopy (AFM) is a widely used tool for characterizing material surfaces. Here, we present a framework, namely, artificially intelligent lab assistant (AILA), which enables automation of AFM experiments using large language model-based (LLMs) agents. To evaluate the performance of AILA, we present the first benchmarking dataset, AFMBench, that consists of 45 manually curated tasks corresponding to real-world AFM experiments. These include single-step, multi-step, and mathematical reasoning-based tasks that critically analyze the ability of AILA to perform AFM experiments. Finally, we present two automated experiments using AILA, first, the calibration of the AFM, and second, the imaging of a graphene step. The results presented here highlight the capability of LLMs to guide automated high throughput experiments, accelerating the materials characterizations.

1 Introduction

Large language models (LLMs) have revolutionised several domains due to their ability to process information, which is then used to guide the next set of actions. Recent works have demonstrated the capability of these models to act as a planner, even for large-scale tasks [1]. This capability has been exploited further by the development of automated experiments [2, 3, 4, 5] and agent-based modelling that enables the use of LLMs for experiments [6, 7] and materials discovery [8]. While there has been an increased use of LLMs in chemistry for synthesis and chemical reaction planning [7], there have been only limited efforts towards using them to make characterization techniques more accessible or, for example, for predicting experimental effort and cost [9].

Scanning problem microscopy, including atomic force microscopy (AFM), are widely used to characterize material, surfaces, interfaces, and interactions at the atomic scale [10]. There have been several efforts to automate these techniques [11, 12], as efficient execution of these experiments require years of training. However, most of these techniques are task-specific and restricts an interactive approach to address an out-of-distribution task that may arise during the operation [13].

Here, we address this challenge by developing an LLM-based lab assistant that automates the operation of AFM. The major contributions of the present work are as follows.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- **AI-lab assistant:** We present an LLM-based agent, artificial intelligence lab assistant (AILA), that enables automated operation of AFM.
- **AFMBench:** We present a benchmark dataset comprising question-answer pairs, which evaluates the capabilities of LLMs to perform real-world AFM-based experiments.
- **Automated AFM:** We demonstrate the capability of AILA to capture images in an automated fashion by optimizing the parameters of AFM, thereby, enabling a truly self-driving characterization lab.

2 Methodology

2.1 AILA framework

Figure 1 shows the architecture of AILA. The core of this architecture is the LLM-based planner, namely, AILA. The planner's primary role is to obtain user input and assign tasks between two specialized agents: the AFM Operation Handler Agent and the Data Handler Agent. The planning system utilizes different large language models (LLMs), such as GPT-4 and GPT-3.5 turbo, to enable the delegation of tasks. AILA, the decision-making component, evaluates user queries and assigns them to the suitable agent according to the specifics of the request. Suppose the inquiry refers to AFM instrument operations, it is routed to the AFM Operation Handler Agent. It comes with two essential tools: document retrieval for relevant information extraction from the specific AFM documentation, and a code executor for instrument control. Alternatively, if the query involves image characterization or optimization, it is assigned to the Data Handler Agent. This agent is equipped with an image optimization tool to improve image quality, and an image analyzer tool capable of viewing and extracting crucial information from images. If a query falls outside the scope of both agents, the planner provides the user with suggested actions or alternative solutions.

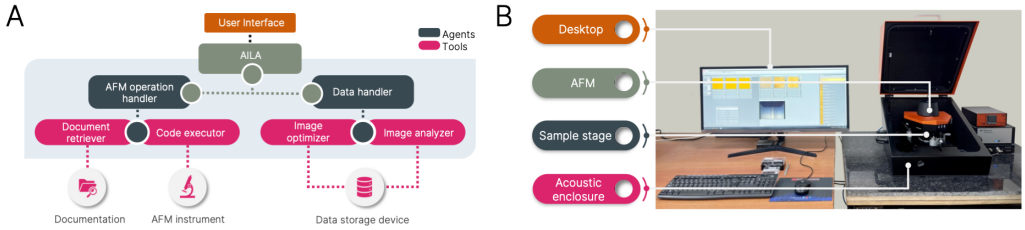

Figure 1: (**A**) Architecture of AILA (Artificially Intelligent Laboratory Assistant). (**B**) Components of the AFM (Atomic Force Microscope) system.

2.2 AFM operation handler agent

Operating an Atomic Force Microscope (AFM) includes numerous complex stages that must be done in a precise order. For example, acquiring a topographic image necessitates the selection of imaging parameters, the suitable AFM tip type, and determining the operational mode, e.g., in sliding or non-contact configuration. The AFM tip must then approach the surface of the specimen and scan the image, which is subsequently stored digitally in the desired format. Deviating from this sequence may lead to inaccurate results, potential harm to the equipment, or loss of data. In our investigation, we utilized the DriveAFM system from Nanosurf and interfaced with the Nanosurf control software via a Python API. Note that this can be integrated into any AFM system with API access. Controlling the AFM with Python requires developing task-specific code in the correct sequence, a process impeded by often unorganised technical documentation. To overcome this challenge, we designed a vector database that organizes the appropriate Python scripts for specific activities. A document retriever tool was built to interface between this database and an AFM operation handler agent, which contains its own language model and memory. This agent dynamically retrieves, updates, and generates Python code in the appropriate order for each task, communicating with the code executor tool that interfaces with the Nanosurf software via API.

The document retrieval tool utilises ChromaDB, an open-source vector database, for storing and retrieving vector embeddings. We utilised the text-embedding-3-lar model to create embeddings by splitting Python codes from technical documentation into segments with a maximum token size of 1000, ensuring no overlaps. These embeddings were thereafter saved together with their

corresponding metadata. For retrieval, we used LangChain's Recursive Similarity Search, providing similarity-based searches based on a minimum similarity threshold. The code execution tool runs Python scripts generated by LLM models using technical documentation. The agent organizes tasks between the document retriever and a code execution tool, assigning actions accordingly. Following the completion of the task, the AFM operation handler can either return the output to the user or transfer it to other agents, depending on the experimental workflow.

2.3 Data handler agent

Attaining precise surface tracking in AFM requires the optimization of scan parameters, namely the Proportional (P), Integral (I), and Derivative (D) gains[14, 15]. Optimally adjusted parameters guarantee that the trace and retrace lines closely track each other, indicating stable scanning conditions. Here we develop a Data Handler agent that can effectively handle user queries relevant to the optimization of AFM scan parameters and image characterization. The agent interfaces with two tools, an image optimizer and an image analyser, which are connected to a storage device for processing saved images. A genetic algorithm was incorporated into the image optimizer to optimize the AFM scan parameters. The method functions by producing and assessing three sets of images per iteration, finally providing the optimized PID gain values to the Data Handler agent for subsequent processing based on query specifications supplied by the user.

An inherent obstacle in contact-mode Atomic Force Microscopy (AFM) is the occurrence of tip wear and fracture, frequently resulting from prolonged scanning, diminishing the tip's longevity[16]. Our method effectively addresses this issue by reducing the number of scans needed for parameter optimization. The Structural Similarity Index (SSIM) was employed as the fitness function in the genetic algorithm to streamline the optimization of scanning settings while minimizing surface contact. SSIM quantifies the degree of similarity between the backward and forward images produced during the trace and retrace procedure. In contrast to the mean square error (MSE), which necessitates extensive surface scans and may fail to provide correct results in the presence of drift, SSIM can capture variations in structural information, brightness, and contrast[17, 18]. This allows the algorithm to optimize parameters even while scanning small regions efficiently.

2.4 AFM Bench

To evaluate the performance of the AILA framework, we created a hand-curated set of question (see appendix 3). These are tasks that are routinely involved in the operation of AFM, such as approaching the cantilever tip, scanning, saving a file, or renaming a file, to name a few. Further, the questions involved single instruction, multi-step instructions, and mathematical reasoning. While single instruction refers to tasks such as saving a file, multi-step instruction refers to tasks such as approaching the tip, scanning a surface, and then saving any associated data as a file. Mathematical reasoning is the most complex category of question, where information provided in the form of data should be analysed to deduce the results and take further action. This involves actions such as scanning an image, analysing its quality, and taking another image with different values of PID gain. Note that these tasks correspond to realistic actions that an experimentalist needs to perform while taking measurements using an AFM. Thus, AFM Bench enables critical evaluation of LLMs for performing real-life AFM experiments.

3 Evaluation of AILA

We start by evaluating the usability of the AILA framework considering two language models namely, GPT-4 and GPT-3.5 turbo. The overall framework is agnostic to the choice of LLMs and hence, any LLM can be used as plug-and-play.

AFMBench: Table 1 shows the performance of AILA on the AFMBench dataset. We observe a notable difference in the performance of GPT-4 and GPT-3.5 turbo in AFMBench. In the case of mathematical reasoning, the LLMs exhibit start differences with GPT-4 giving an accuracy of 80%, while GPT-3.5 turbo gives an accuracy of mere 6.67%. Thus, while GPT-4 is a reasonable candidate as an LLM for AILA, GPT-3.5 turbo is not suitable for real-life AFM experiments within the current framework.

Question type	Total #	Models		Models	
		GPT-4	GPT-3.5 turbo	GPT-4	GPT-3.5 turbo
Single Instruction	15	100%	86.70%	27 s	9 s
Multi-step Instruction	15	93%	20%	75 s	16 s
Mathematical Reasoning	15	80%	6.67%	100s	14s

Table 1: Comparison of accuracy and time efficiency between GPT-4 and GPT-3.5

Calibration of AFM: Imaging using AFM requires careful calibration of PID gain values on a grid sample (or any other standard calibration sample). Due to the continuous values taken by these gains, it requires a trained expert operator to optimize toward high quality images. This limits the equipment's accessibility for a wider base of users. To this end, we optimize the gains using AILA, by minimizing the error on the forward and backward line scan of the AFM. Figure 2 shows several examples of experimental AFM data obtained at variable PID gain values, together with the corresponding line scan. The error between the forward and backward scan is obtained using the SSIM which computes the similarity of the distribution from the two cases. The calibration of AFM is posed as an optimization problem to minimize the SSIM, which is achieved through the genetic algorithm. We observe that already within a few trials (<15), we obtain the optimized PID values for which high-quality images are generated. The quality obtained from these images is comparable or even superior to that obtained by a human expert at a significantly higher experimental cost in terms of operational time.

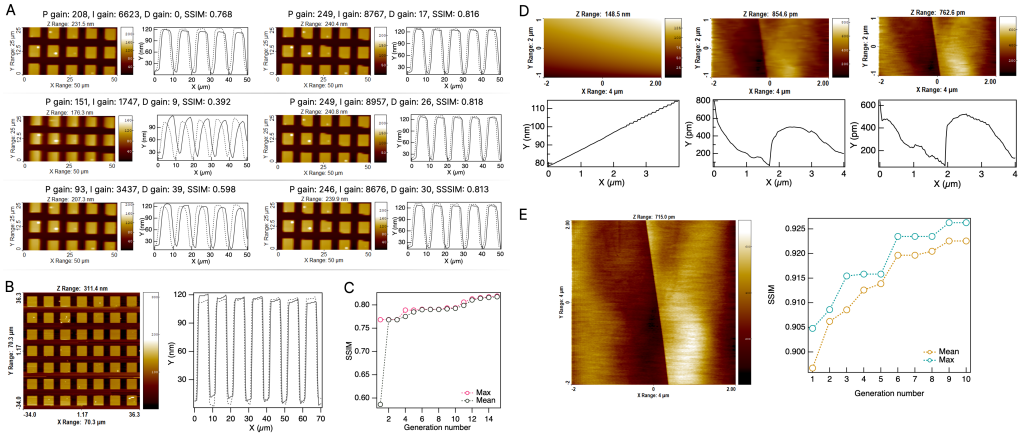

Figure 2: (**A**) The first set of three AFM images (left) with their corresponding horizontal line profile. Dotted lines indicate the backward profile, and the solid line shows the forward profile. The last set of three optimized (right) images after the 15^{th} generation. (**B**) Optimized AFM image with its corresponding horizontal line profile. (**C**) Variation of SSIM across generations in the genetic algorithm. (**D**) Optimizing the AFM to image a step in the graphene. (**E**) Optimized AFM topography of HOPG and variation of SSIM across generations.

Imaging a graphene step: Surface characterization devices, such as atomic force microscopy (AFM), often cause baseline artifacts[19] due to factors such as noise, drift over time[20], mechanical vibrations, temperature changes, and electronic interference. These artifacts might affect actual topographic data, concealing the true surface features of the material. For example, when the surface of a specimen is slightly inclined, it can create a sloping baseline, which can further complicate the identification of surface characteristics. While baseline correction is usually more straightforward for more prominent features, it becomes problematic for weak topographic features, which are more easily masked by baseline noise.

We now utilize a highly ordered pyrolytic graphite (HOPG) sample to collect topographic data containing prominent surface characteristics with minute features, i.e., graphene steps. As illustrated in Figure 2(D), small graphene steps were difficult to detect because of the dominating baseline artifacts cause by the more prominent features. We incorporate a baseline correction technique as part of our image optimization tool to address this issue. By applying a higher-degree polynomial baseline correction, we successfully minimized the baseline noise and enhanced the visibility of small features. Polynomial baseline correction is a generally established method for correcting baseline artifacts in topographic data. Its flexibility allows for mimicking various baseline shapes, from simple linear trends to more complex functions, making it excellent for reducing instrumental noise while keeping the integrity of the surface topography.

4 Conclusions

Altogether, here we proposed a GPT model-based framework, AILA, which takes instructions from users in natural language, and converts them into executable code for AFM imaging. Our results show that LLM-based agents can act as effective lab assistants that enable high-throughput experiments. Our work is currently limited to simple experiments using AFM; doing more extensive evaluation of

other measurement modes, performing open experiments requiring recursive feedback, and extending this approach to other probe measurements are interesting future directions to pursue.

References

[1] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

[3] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[4] Florian Häse, Loïc M Roch, and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1(3):282–291, 2019.

[5] Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 2024.

[6] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[7] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[8] Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029*, 2023.

[9] Santiago Miret and NM Krishnan. Are LLMs Ready for Real-World Materials Discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[10] Ke Bian, Christoph Gerber, Andreas J Heinrich, Daniel J Müller, Simon Scheuring, and Ying Jiang. Scanning probe microscopy. *Nature Reviews Methods Primers*, 1(1):36, 2021.

[11] Sergei V Kalinin, Maxim Ziatdinov, Jacob Hinkle, Stephen Jesse, Ayana Ghosh, Kyle P Kelley, Andrew R Lupini, Bobby G Sumpter, and Rama K Vasudevan. Automated and autonomous experiments in electron and scanning probe microscopy. *ACS nano*, 15(8):12604–12627, 2021.

[12] Mohammad Rashidi and Robert A Wolkow. Autonomous scanning probe microscopy in situ tip conditioning through machine learning. *ACS nano*, 12(6):5185–5189, 2018.

[13] Sergei V Kalinin, Colin Ophus, Paul M Voyles, Rolf Erni, Demie Kepaptsoglou, Vincenzo Grillo, Andrew R Lupini, Mark P Oxley, Eric Schwenker, Maria KY Chan, et al. Machine learning in scanning transmission electron microscopy. *Nature Reviews Methods Primers*, 2(1):11, 2022.

[14] Maja Dukic, Vencislav Todorov, Santiago Andany, Adrian P Nievergelt, Chen Yang, Nahid Hosseini, and Georg E Fantner. Digitally controlled analog proportional-integral-derivative (pid) controller for high-speed scanning probe microscopy. *Review of Scientific Instruments*, 88(12), 2017.

[15] Noriyuki Kodera, Mitsuru Sakashita, and Toshio Ando. Dynamic proportional-integral-differential controller for high-speed atomic force microscopy. *Review of Scientific Instruments*, 77(8), 2006.

[16] Koo-Hyun Chung, Yong-Ha Lee, and Dae-Eun Kim. Characteristics of fracture during the approach process and wear mechanism of a silicon afm tip. *Ultramicroscopy*, 102(2):161–171, 2005.

[17] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[18] Gintautas Palubinskas. Image similarity/distance measures: what is really behind mse and ssim? *International Journal of Image and Data Fusion*, 8(1):32–53, 2017.

[19] Susana Moreno-Flores. Baseline correction of afm force curves in the force–time representation. *Microscopy Research and Technique*, 79(11):1045–1049, 2016.

[20] Francesco Marinello, Paolo Bariani, Leonardo De Chiffre, and Enrico Savio. Fast technique for afm vertical drift compensation. *Measurement Science and Technology*, 18(3):689, 2007.

A Appendix / supplemental material

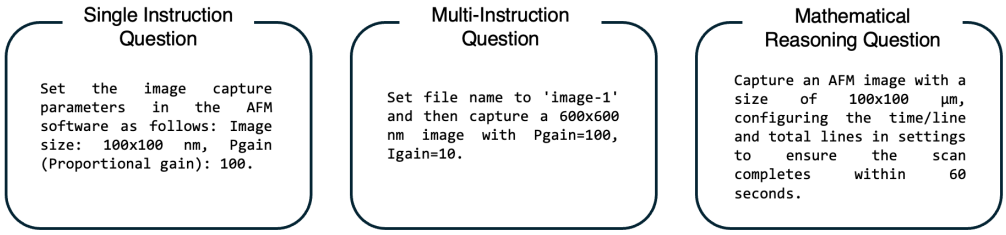

Figure 3: Examples of questions used in AFMBench.