**Data Science Report: AI Lab Report Agent**

## 1. Project Objective

The objective of this data science project was to develop an AI agent capable of automating the generation of engineering and chemistry lab reports. This involved two key machine learning tasks: fine-tuning a model for specialized code generation and establishing a robust methodology for evaluating the quality of the final generated text.

## 2. Fine-Tuning the Coder Model

### 2.1. Base Model and Objective

The base model selected was **microsoft/Phi-3-mini-4k-instruct**. The objective was to specialize this model for the specific task of converting experimental context and structured observations (JSON data) into executable Python code for calculations relevant to various lab experiments.

### 2.2. Training Data

The model was trained on a curated dataset in JSON Lines (.jsonl) format. The training set (training/train.jsonl) and evaluation set (training/eval.jsonl) consist of examples covering physics, chemistry, and engineering principles. Each line in the dataset contains three fields:

- **context**: A string containing the aim and theory of the experiment.
- **observations**: A JSON object of sample experimental readings.
- **code**: The ideal, "golden" Python script for calculating the results from the observations.

### 2.3. Methodology

To efficiently train the model, **Parameter-Efficient Fine-Tuning (PEFT)** using the **LoRA (Low-Rank Adaptation)** method was employed. This approach avoids the prohibitive computational cost of full fine-tuning by freezing the base model's weights and training only a small number of new "adapter" layers. This makes it possible to achieve high performance on consumer-grade hardware like the T4 GPU used in the training notebook.

Key training parameters included a LoRA rank (r) of 16, lora_alpha of 32, a learning rate of 2e-4, and training for 10 epochs.

### 2.4. Results

The training process successfully converged, with the training loss decreasing from 0.6855 to 0.1014 over 100 steps. This indicates that the model learned to map the context and observations to the correct code structure effectively. The resulting fine-tuned model adapter was saved and subsequently merged and uploaded to the Hugging Face Hub as Barghav777/phi3-lab-report-coder.

## 3. Evaluation Methodology

### 3.1. Objective

The goal of the evaluation was to quantitatively and qualitatively measure the final report quality generated by the end-to-end agent pipeline.

### 3.2. Dataset

An evaluation dataset (evaluation/eval_dataset.jsonl) was created, containing 9 distinct experiments from the field of mechanical engineering. Each entry in the dataset contained:

- manual_path: A path to the full lab manual PDF.
- observations: A JSON object of sample experimental data.
- golden_report: A complete, human-written "gold standard" lab report for that experiment.

### 3.3. Metrics

The evaluation uses the ROUGE score, a standard metric for summarization and text generation tasks.

- **ROUGE-1 F1-Score:** Measures the overlap of individual words (unigrams). This serves as a proxy for **factual accuracy** and keyword recall.
- **ROUGE-L F1-Score:** Measures the longest common subsequence of words. This serves as a proxy for **prose quality** and sentence structure similarity.

## 4. Evaluation Outcomes

### 4.1. Quantitative Analysis

After multiple rounds of prompt engineering and upgrading the report writer model to Llama3-70b-8192, the final evaluation scores on the 9-example dataset were:

- **Average ROUGE-1 F1-Score: 0.5438**
- **Average ROUGE-L F1-Score: 0.3470**

This result is strong. A ROUGE-1 score above 0.5 indicates that the agent is successfully identifying and including the correct facts, terminology, and data. The ROUGE-L score is in the "Needs Improvement" range, suggesting that the agent's prose is stylistically different from the golden reports, even if factually correct.

### 4.2. Qualitative Analysis

A manual review of the saved output in the evaluation_results/ folder confirmed the quantitative findings. The generated reports were factually accurate and well-structured. The lower ROUGE-L score was primarily due to the model's tendency to use different sentence structures and valid synonyms compared to the golden reports. For example, the model might write "The experiment was successfully performed" while the golden report says "The procedure was carried out successfully." Both are correct, but the textual difference lowers the score.

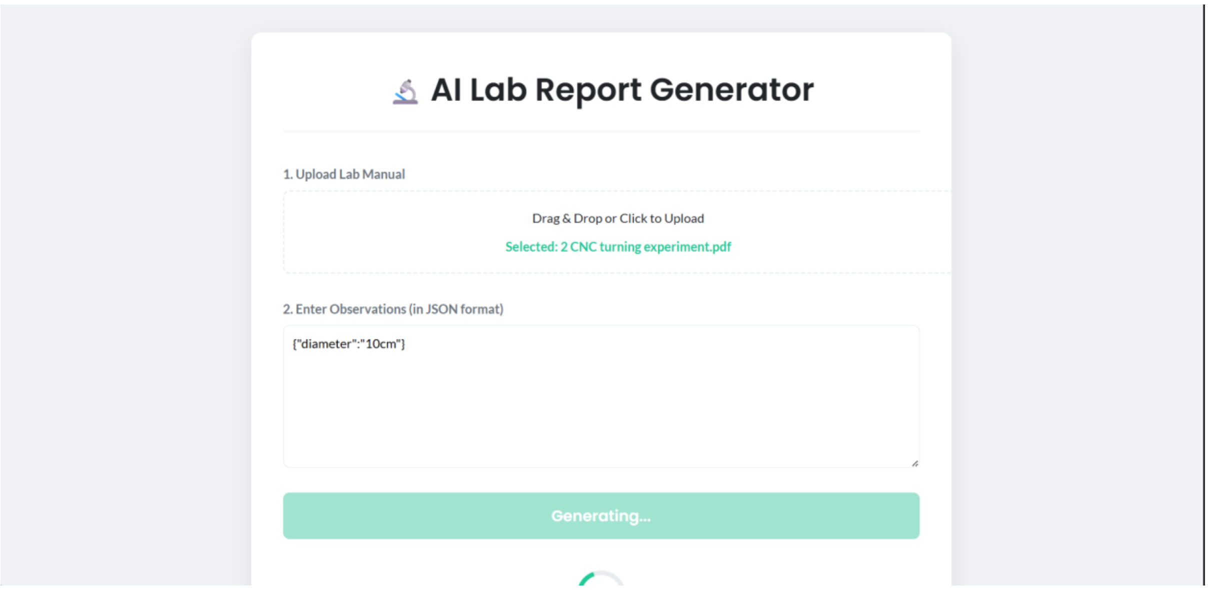
### 4.3. Conclusion

The evaluation confirms that the AI Lab Report Agent is a functional and effective prototype. It reliably extracts context, calculates results, and generates factually accurate reports. Future work to improve the ROUGE-L score could involve more advanced prompt engineering or fine-tuning the report-writing model itself on a corpus of ideal lab reports.

## 5. Interaction logs

Interaction logs can be found in this link: https://g.co/gemini/share/e847cb097ffb

## 6. Screenshots



AI Lab Report Generator

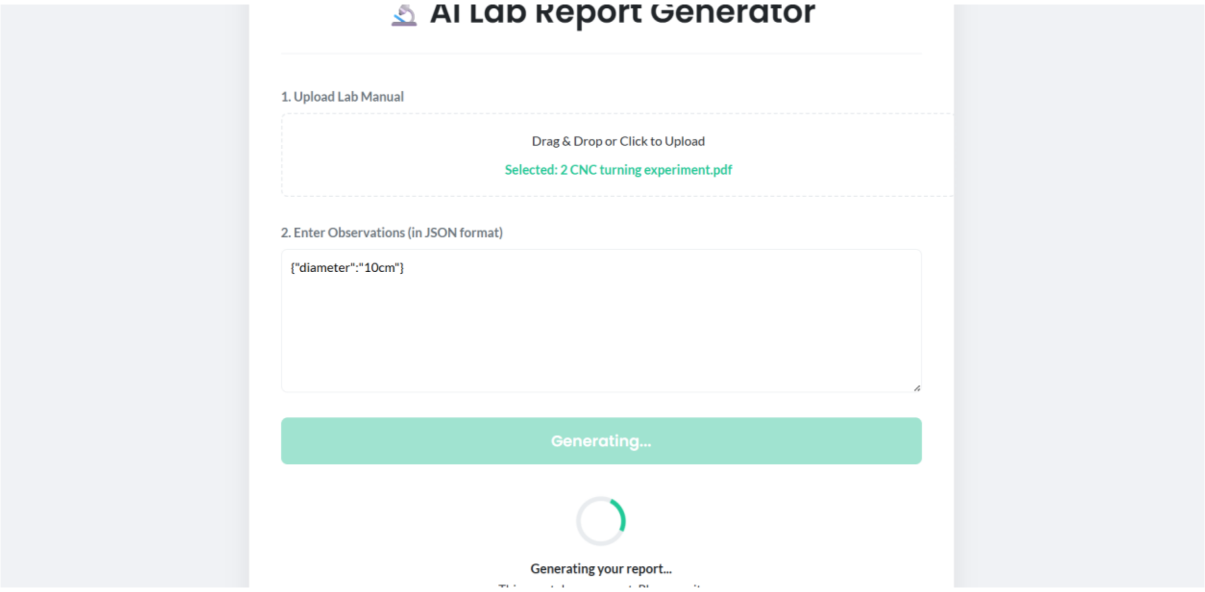
1. Upload Lab Manual

Drag & Drop or Click to Upload

Selected: 2 CNC turning experiment.pdf

2. Enter Observations (in JSON format)

```
{"diameter":"10cm"}
```



AI Lab Report Generator

1. Upload Lab Manual

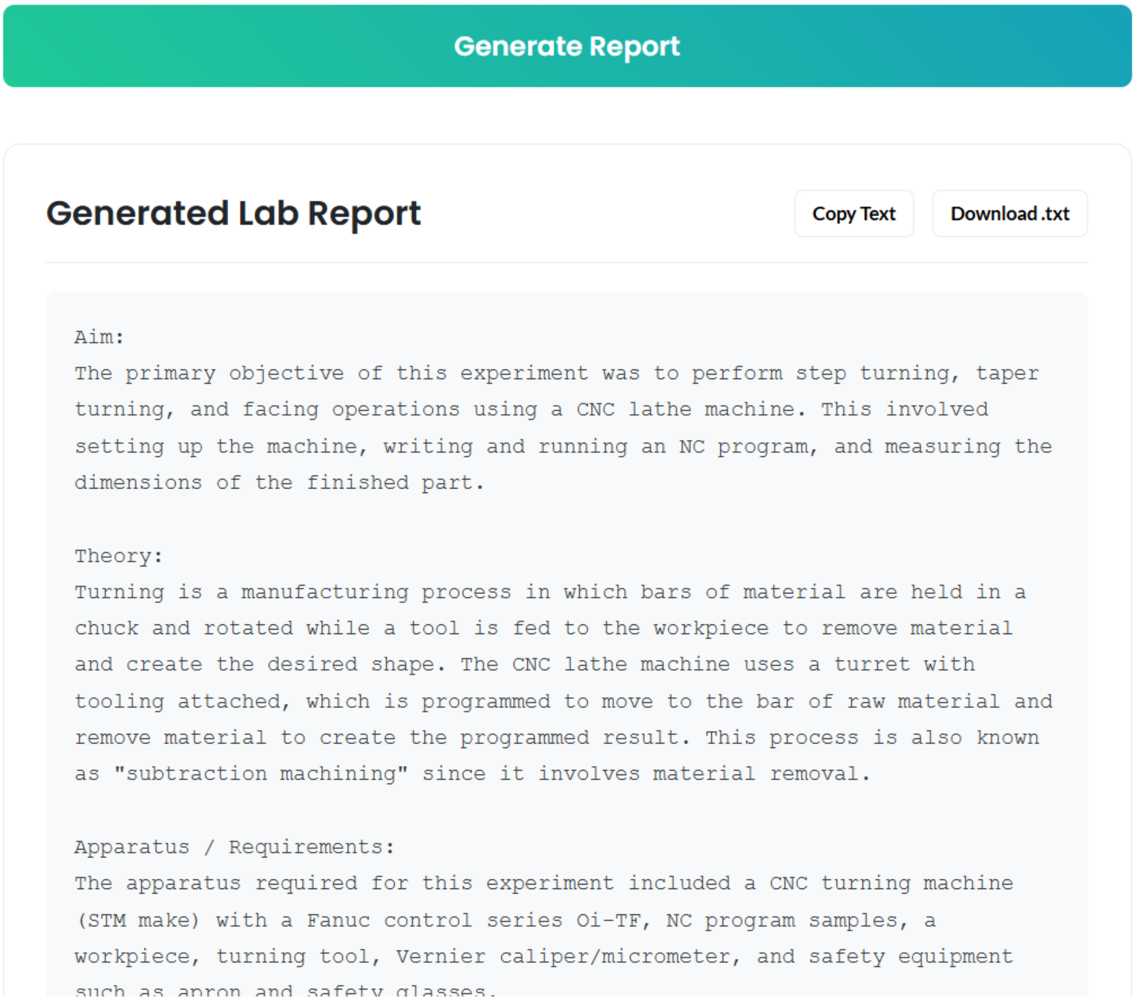
Drag & Drop or Click to Upload

Selected: 2 CNC turning experiment.pdf

2. Enter Observations (in JSON format)

```
{"diameter":"10cm"}
```

Generating your report...

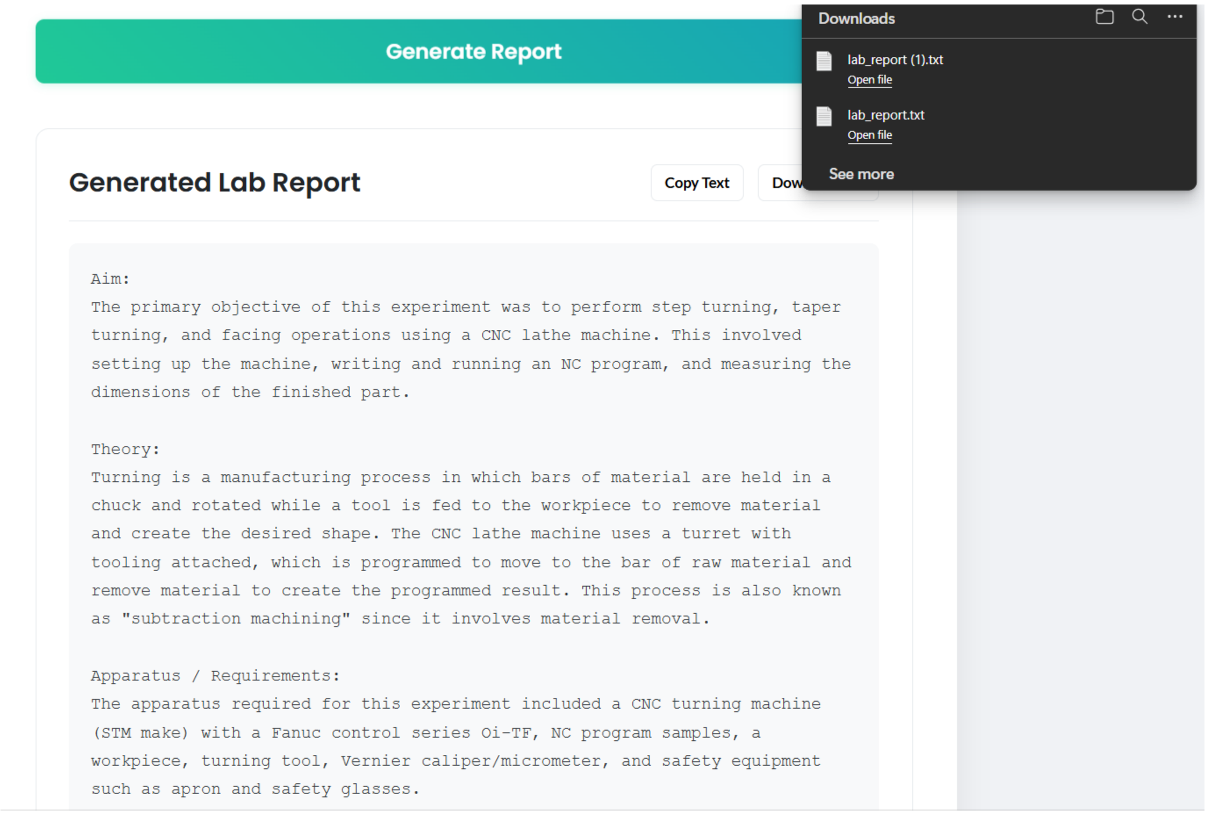


## Generated Lab Report

```
Aim:
The primary objective of this experiment was to perform step turning, taper
turning, and facing operations using a CNC lathe machine. This involved
setting up the machine, writing and running an NC program, and measuring the
dimensions of the finished part.

Theory:
Turning is a manufacturing process in which bars of material are held in a
chuck and rotated while a tool is fed to the workpiece to remove material
and create the desired shape. The CNC lathe machine uses a turret with
tooling attached, which is programmed to move to the bar of raw material and
remove material to create the programmed result. This process is also known
as "subtraction machining" since it involves material removal.

Apparatus / Requirements:
The apparatus required for this experiment included a CNC turning machine
(STM make) with a Fanuc control series Oi-TF, NC program samples, a
workpiece, turning tool, Vernier caliper/micrometer, and safety equipment
such as apron and safety glasses.
```



## Generated Lab Report

```
Aim:
The primary objective of this experiment was to perform step turning, taper
turning, and facing operations using a CNC lathe machine. This involved
setting up the machine, writing and running an NC program, and measuring the
dimensions of the finished part.

Theory:
Turning is a manufacturing process in which bars of material are held in a
chuck and rotated while a tool is fed to the workpiece to remove material
and create the desired shape. The CNC lathe machine uses a turret with
tooling attached, which is programmed to move to the bar of raw material and
remove material to create the programmed result. This process is also known
as "subtraction machining" since it involves material removal.

Apparatus / Requirements:
The apparatus required for this experiment included a CNC turning machine
(STM make) with a Fanuc control series Oi-TF, NC program samples, a
workpiece, turning tool, Vernier caliper/micrometer, and safety equipment
such as apron and safety glasses.
```