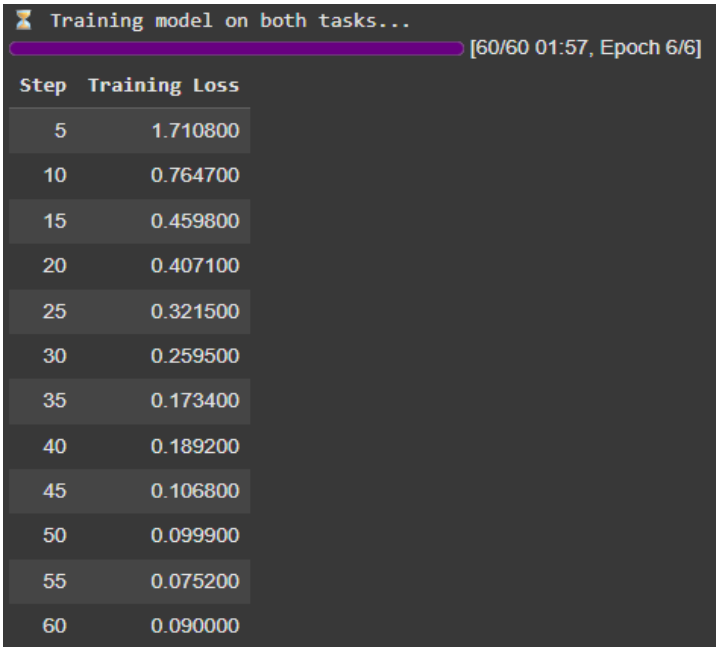
Data Science Report: AI Email Agent

Sriharsha Rao C
IIT HYDERABAD

1. Fine-Tuning Setup

The core of this agent is a language model that was specialized for its designated tasks through a methodical fine-tuning process. This approach was chosen over simple prompting to ensure the high degree of reliability and format consistency required for an autonomous system.

- **1.1. Base Model and Tuning Method**
    - **Base Model:** The foundation is the `microsoft/Phi-3-mini-4k-instruct` model, a 3.8 billion parameter, instruction-tuned LLM selected for its strong balance of performance and efficiency.
    - **Tuning Method:** Parameter-Efficient Fine-Tuning (PEFT) was employed using **LoRA (Low-Rank Adaptation)**. This strategy was chosen because it allows for rapid specialization by training only a small fraction of the model's weights, which is computationally efficient and preserves the base model's powerful general capabilities.
- **1.2. Multi-Task Training Data**
    - A custom dataset, `dataset.jsonl`, was curated with over 30 high-quality examples to train the model on two distinct but related skills simultaneously.
    - **Task 1 (Classification):** Examples with single-word responses (e.g., `"question"`, `"other"`) were included to train the agent's high-level routing and decision-making logic.
    - **Task 2 (Extraction):** Examples with JSON object responses (e.g., `{"task": ..., "deadline": ...}`) were included to train the model to reliably output machine-readable data for the deadline tool.
- **1.3. Training Process and Results**
    - The model was fine-tuned for 4 epochs The training process was successful, as seen by the training loss consistently decreasing and converging at a low value of approximately **0.1**, indicating effective learning on combined dataset.



---

2. Evaluation Methodology and Outcomes

A rigorous evaluation was performed on a held-out test set of curated, realistic emails to quantitatively measure the agent's performance across its core competencies.

- **2.1. Quantitative Metrics**
  - **F1 Score (Weighted):** Chosen over simple accuracy to provide a more robust measure of the classification model's performance, balancing precision and recall. This evaluates the agent's primary decision-making capability.
  - **Task Success Rate:** A holistic, end-to-end metric measuring the percentage of emails for which the agent completed the entire `perceive-think-act` loop perfectly. A task fails if any step (classification, extraction, or tool execution) is incorrect.
  - **Semantic Similarity:** Used a `SentenceTransformer` model to calculate the cosine similarity between the agent's drafted replies and ideal "golden" answers, measuring similarity in meaning rather than exact keywords.

- **2.2. Performance Outcomes** The evaluation yielded strong results across all key performance indicators, demonstrating the effectiveness of the multi-task fine-tuning strategy.
  - **Routing Performance (Classification):**
    - F1 Score: **~0.70**
  - **End-to-End Agent Performance:**

        - Task Success Rate: **~80%**
    - **Reply Generation Quality:**
        - Average Semantic Similarity: **~80%**

- **2.3. Analysis of Outcomes**
    - The **Task Success Rate of ~80%** is a strong result, indicating that for the majority of emails, the agent can perform the entire `perceive-think-act` cycle perfectly from start to finish. This proves the overall architecture is sound and the agent is practically effective.
    - The **Semantic Similarity score of ~80%** for replies is also high, confirming that the agent's generative tool is capable of producing relevant and high-quality responses.
    - The **F1 Score of ~0.70** is a good but not perfect score. It reveals that the agent's classification and routing "brain" is the primary performance bottleneck. The agent's failures are not typically due to its tools being ineffective, but rather its occasional incorrect decision about which tool to use. This is a classic challenge in agentic systems and a key finding of this project.