AI Email Agent Architecture Document

Ashwin Gaikwad - IIT Goa - IBY DS Internship Assignment

September 16, 2025

Abstract

This document outlines the architecture of the AI Email Agent, a system designed to intelligently automate email management. The architecture is based on a sequential processing pipeline where personally trained models handle classification and extraction before a final orchestration step. This report details the system's components, the interaction flow, the AI models used, and the reasons for their selection.

1 Core Philosophy: A Team of Specialists

The system employs a "team of specialists" approach. This design philosophy uses specific, highly-trained models for initial processing tasks, with a more powerful, general AI handling the final strategic decisions.

- **The Triage Specialist (DistilBERT):** A rapid clerk that instantly sorts all incoming mail into predefined piles.
- **The Data Analyst (Llama 3):** A meticulous analyst that reads specific documents and extracts key information into a structured format.
- **The Project Manager (Gemini 1.5 Pro):** The manager who reviews the sorted piles and the analyst's forms to decide on a final course of action.

This approach is designed to be efficient, accurate, and cost-effective by using specialized local models for 80% of the work and reserving the Gemini API for the final 20% of orchestration tasks.

2 System Components & Architecture

The system is built around a clear, sequential workflow that processes new emails from arrival to action.

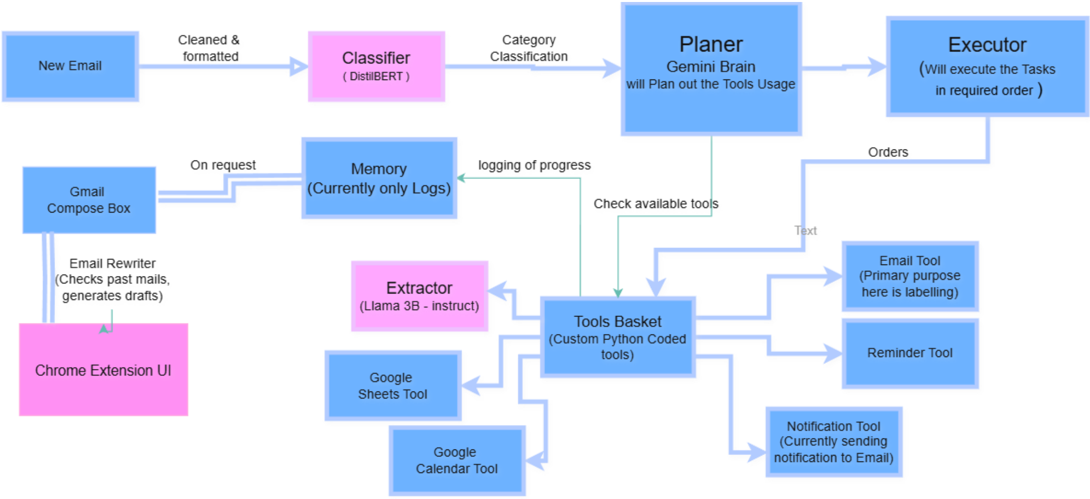

Figure 1: The high-level data flow of the AI Email Agent's sequential processing pipeline.

2.1 Component 1: The Email Monitor

- **Technology:** Gmail API.
- **Role:** This component monitors the user's Gmail inbox for NEW emails only (since the server started). It tracks processed email IDs to prevent duplicate processing.

2.2 Component 2: The Triage Specialist (DistilBERT Classifier)

- **Model:** A fine-tuned distilbert-base-uncased model.
- **Role:** This is the first model in the processing chain. It classifies every new email into one of eight university-specific categories.

2.3 Component 3: The Data Analyst (Llama 3 Extractor)

- **Model:** meta-llama/Meta-Llama-3-8B-Instruct fine-tuned with LoRA.
- **Role:** This model handles all data extraction tasks. If an email requires structured data extraction (e.g., for jobs, events, or deadlines), this model processes it after classification.

2.4 Component 4: The Project Manager (Gemini 1.5 Pro)

- **Model:** Gemini 1.5 Pro.
- **Role:** This model acts as the orchestrator and planner. It receives the pre-classified category from DistilBERT and the extracted data from Llama 3, and then creates an execution plan and selects the appropriate tools. It does not perform classification or extraction itself.

2.5 Component 5: AI-Powered Writing Assistant (Chrome Extension)

- **Technology:** Chrome Extension for UI, with Gemini AI for generation.
- **Role:** Provides a seamless writing assistant within the Gmail interface.
- **Functionality:** It performs style-matched email rewriting by analyzing the user's past emails. It generates context-aware drafts with a tone and style adapted to the specific recipient.

3 Interaction Flow

The following steps describe the journey of an email from its arrival to the completion of an automated action, following the sequential processing flow.

1. **Arrival:** A new email lands in the inbox. The **Email Monitor** detects it.
2. **The Quick Sort:** The email is immediately sent to the **DistilBERT Classifier**. It reads the email and assigns it a category, for instance, Job Recruitment.
3. **The Deep Dive:** If the category requires it, the email is passed to the **Llama 3 Extractor**. The model reads the content and produces a clean, structured JSON output containing the relevant extracted details.
4. **The Master Plan:** The **Gemini Orchestrator** receives the category and the structured data from the previous models. It then formulates an execution plan, such as adding the job details to a tracker and labeling the email.
5. **Execution:** The system's tool execution layer carries out the plan. For a job email, this involves adding the extracted data to a Google Sheet and applying an ”AI-Jobs” label in Gmail.

4 Models Used and Justification

4.1 Email Classifier (Model: Fine-tuned DistilBERT)

- **Model Used**: The base model is distilbert-base-uncased. It was trained on a personal dataset of 5,678 emails to recognize university-specific patterns.
- **Reason for Choice**:
  - **Speed:** It is extremely fast, processing an email in approximately 0.006 seconds.
  - **Accuracy:** It was trained on personal emails, giving it a validation accuracy of about 95.1% on the user's specific email patterns.

4.2 Data Extractor (Model: Fine-tuned Llama 3)

- **Model Used**: meta-llama/Meta-Llama-3-8B-Instruct, modified using Low-Rank Adaptation (LoRA).
- **Reason for Choice**:
  - **Instruction Following:** The "Instruct" version of Llama 3 is exceptional at adhering to complex formatting requests, which is essential for reliably producing structured JSON output. This was its biggest advantage for the task.
  - **Performance:** Upon its release, Llama 3 8B set a new performance standard for models of its size, outperforming previous leaders. Starting with the top performer was a major advantage for a task requiring precision.
  - **Efficiency:** LoRA allows for efficient fine-tuning on a custom dataset to recognize specific patterns in the user's emails.

4.3 Orchestrator & Writing Assistant (Model: Gemini 1.5 Pro)

- **Model Used**: Gemini 1.5 Pro.
- **Reason for Choice**:
  - **Advanced Planning:** For orchestration, Gemini is used because it excels at creating execution plans and coordinating workflows based on the inputs from the specialized models.
  - **Cost Optimization:** Using Gemini only for orchestration, after the local models have done the classification and extraction, significantly reduces API calls and cost.
  - **Writing Assistance:** For the Chrome extension, it performs style-matched email rewriting by analyzing past emails and generating context-aware drafts.