**Experiment Plan**

**Objective:**  
Develop a robust synthetic email generator that transforms real business emails into anonymized versions. The tool is designed to preserve the original email’s semantic intent, tone, and structure while systematically replacing PII and company-specific data.

**Research:**

* **Literature Review:** Investigated state-of-the-art text anonymization and paraphrasing techniques to ensure that sensitive details are completely removed yet the communication intent remains intact.
* **Techniques & Libraries:**
  + **Regex-based Transformations:** Used extensively for pattern matching and substitution. This includes handling complex patterns like currency with multipliers (e.g., million, billion), date formats (ISO and written), and US phone number formats.
  + **Pandas & Streamlit:** Employed for interactive data manipulation and rapid UI prototyping. These libraries allow for efficient handling of file uploads and direct email input.
  + **OpenAI API Integration:** Utilized to invoke LLM-driven paraphrasing, where custom system prompts and user prompts ensure context-aware synthesis. The API call leverages a model (e.g., “gpt-4o-mini”) to deliver natural and coherent transformations.

**Methodology:**

* **Step 1 – Rule-based Preprocessing (EntityReplacer):**
  + Implements direct string substitutions (e.g., replacing “Enron” with “Agriculture India” and adjusting email domains) using case-insensitive regex matching.
  + Contains specialized functions for:
    - **Currency Conversion:** Converts USD amounts to local currency based on multipliers and formats them in regional styles (e.g., using “crore” and “lakh” for India).
    - **Date Conversion:** Supports multiple date formats, updating years while preserving month and day.
    - **Phone Number Conversion:** Rewrites US phone numbers to target-specific formats.
* **Step 2 – LLM-Driven Paraphrasing (LLMParaphraser):**
  + Constructs a detailed system prompt with explicit transformation guidelines (covering names, dates, currency, and industry-specific terms).
  + Sends the transformed text to the OpenAI API, ensuring that the LLM output refines the initial regex changes into a natural-sounding email.
* **Step 3 – Post-transformation Validation (Validator):**
  + Uses regex to scan for any residual sensitive elements such as unreplaced company names, email domains, currency symbols, or old date formats.
* **Integration & User Interface:**
  + The entire pipeline is built within a Streamlit app that supports both file-based and direct text input modes.
  + A progress bar, side-by-side comparison, and color-coded highlights (via HTML spans) are incorporated to facilitate user review and verification.

**Experiment Results and Findings**

**Data Analysis:**

* **Dataset Complexity:**
  + The Enron dataset presented varied email formats with diverse structures and mixed data types. Detailed analysis revealed inconsistencies such as multiple date formats, varying currency expressions, and different phone number layouts.
* **Processing Insights:**
  + Data cleaning was crucial; missing values were handled using Pandas, and the email identification process dynamically selects the best column based on header names or text length.

**Synthesis Process:**

* **Transformation Pipeline Execution:**
  + **EntityReplacer Module:**
    - Applied regex substitutions for key identifiers, including a two-step process for currency conversion (handling both simple amounts and those with multipliers).
    - Used modular functions to ensure that dates and phone numbers are consistently reformatted.
  + **LLM Paraphrasing:**
    - The paraphraser constructs a comprehensive system prompt detailing the transformation requirements, then refines the regex outputs via an LLM to ensure coherent narrative flow.
    - This step leverages OpenAI’s chat completions and introduces minor adjustments that align with business communication standards.
* **User Interface & Feedback:**
  + The Streamlit app employs interactive tabs for direct input and file uploads, with real-time progress updates and color-coded highlighting for immediate visual feedback.
  + Side-by-side comparisons allow users to directly see the transformation differences, and a built-in export feature supports CSV downloads of the results.

**Challenges:**

* **Regex Complexity:**
  + Devising regex patterns that accurately capture multiple formats (for currencies, dates, phone numbers) without false positives required extensive testing and iterative refinement.
* **LLM API Variability:**
  + Integrating with the OpenAI API demanded careful prompt engineering to balance between preserving semantic content and ensuring complete anonymization. Occasional API response variations necessitated additional error handling.
* **Validation Rigor:**
  + Ensuring the Validator class caught every potential oversight (like hidden references to the original company name) involved extensive trial and error, further fine-tuning regex patterns.