Problem Statement Summary

The task is to create a synthetic email dataset derived from the Enron corpus, ensuring complete de-identification while retaining the structural and contextual authenticity of real-world emails. The synthetic data must mimic distributions of real enterprise communications (e.g., company names, roles, industry jargon) to serve as training data for fine-tuning LLMs in eDiscovery use cases. Key requirements include:
- Replace all PII (names, emails, dates, companies, phone numbers).
- Preserve email structure, tone, and variability.
- Ensure synthetic emails appear plausible for a fictional company (e.g., "Agriculture India").

Solution Design

Experiment Plan

Objective

Generate synthetic emails by transforming Enron data using LLM-driven entity replacement and contextual adaptation.

Research & Techniques

1. **PII Detection:** Regex patterns for emails, phone numbers, and dates.
2. **Entity Replacement:**
   - Faker Library: Generate fake names, emails, and phone numbers.
   - LLM Rewriting: Use GPT-4 to swap company-specific terms (e.g., "EnronCredit" → "AgriCredit") and industry jargon (e.g., "coal" → "rice").
3. **Context Preservation:** Fine-tune prompts to retain email intent, structure, and variability (e.g., CC fields, signatures).

Methodology

1. **Preprocessing:** Extract email headers and bodies. Use regex to mask detected PII.
2. **LLM Transformation:**
   - Prompt Engineering:
     "Rewrite this email as if from Agriculture India Pvt. Ltd. Replace all names, companies, dates, currencies, and industry terms. Ensure the email retains its original purpose and structure."
   - API Call: Use OpenAI's gpt-4 to process masked emails.
3. **Validation**:
   - Manual checks for realism and PII leakage.
   - Compare synthetic vs. original email length, entity counts, and sentiment.

Experiment Results & Findings

Data Analysis

- Enron emails contain nested replies, CC/BCC fields, and financial/legal jargon.
- Common PII: `@enron.com` emails, Houston/London office references, employee names.

Synthesis Process

1. **Regex Masking:** Replace [\w.-]+@enron\.com with [EMAIL], etc.
2. **LLM Rewriting:**
   - Example input: "Jeff Kinneman at EnronCredit" → Output: "Vikram Malhotra at AgriCredit."
   - Context shifts: "coal" → "rice," "DLJ International Capital" → "GreenField Farms LLP."
3. **Output:** Synthetic emails retained headers (e.g., X-From, Subject) and formal tone.

Challenges

- **Inconsistent Formats:** Nested replies required splitting emails into segments for LLM processing.
- **Industry Jargon:** Manual prompt tuning ensured terms like "total return bond" became "hybrid seed project."
- **API Latency:** Batched processing to stay within rate limits.

Prototype Code

Functionality

```
def mask_pii(text):
	pass

def generate_synthetic_email(original_email):
	pass

# Example usage
original_email = """
From: Sara Shackleton on 08/28/2000 06:53 PM
To: William S Bradford/HOU/ECT@ECT
Subject: Credit Derivatives
...
"""
masked_email = mask_pii(original_email)
synthetic_email = generate_synthetic_email(masked_email)
```

Documentation

- **Setup:** Install “openai” and “faker”.
- **Run:** Call generate_synthetic_email() with an Enron email string.
- **Output:** Synthetic email with Agriculture India context and no PII.

Key Innovations

- **Hybrid Approach:** Combines regex for deterministic PII masking and LLMs for contextual rewriting.
- **Prompt Engineering:** Explicit instructions to swap industry terms and retain structure.
- **Validation Focus:** Manual checks for realism, ensuring usability in downstream LLM fine-tuning.