The Ghost in the Machine

"Le style, c'est l'homme même" (The style is the man himself). — Georges-Louis Leclerc

Task 0: The Library of Babel

You will build a dataset where the primary variable is **authorship**, not topic.

1. **Class 1:** Download novels from two authors in Project Gutenberg (e.g., Dickens, Not Shakespeare, etc.). Make sure to clean up your data before using it anywhere.
2. **Topic Extraction:** Identify 5-10 core topics from the book (e.g., "The Ethics of Science," "The Loneliness of the Arctic").
3. **Class 2:** Use the Gemini API to generate 500 paragraphs (100-200 words each) on those same topics.
4. **Class 3:** Use the Gemini API to generate 500 paragraphs on those same topics, specifically prompted to mimic your chosen author's unique style.

Task 1: The Fingerprint

Before training, you must prove that these classes are mathematically distinct. Perform the following analyses:

1. **Lexical Richness:**
    - **Type-Token Ratio (TTR):** (Unique words / Total words).
    - **Hapax Legomena:** Count the number of words that appear only *once* in a 5,000-word sample. (Higher in humans, usually).
2. **Syntactic Complexity (SpaCy/NLTK):**
    - **POS Distribution:** Calculate the ratio of Adjectives to Nouns. Does the AI "over-describe" compared to the Human?
    - **Dependency Tree Depth:** Use SpaCy to calculate the average depth of the sentence parse trees. Longer, more nested branches indicate higher complexity.
3. **Punctuation Density:** Create a heatmap of punctuation usage (semicolons, em-dashes, exclamation marks).
4. **Readability Indices:** Calculate the Flesch-Kincaid Grade Level.

Task 2: The Multi-Tiered Detective

Build three detectors to separate AI generated text from Human written ones. **Note:** If your models fail to reach high accuracy, *this is still a valid research finding if you are able to do good analysis.* Document why.

- **Tier A (The Statistician):** An XGBoost/Random Forest model using *only* the numerical features from Task 1.
- **Tier B (The Semanticist):** A Feedforward NN using averaged pre-trained embeddings (GloVe/FastText).
- **Tier C (The Transformer):** Fine-tune a `distilbert-base-uncased` or `roberta-base` using LoRA.

Negative Results: If your models cannot distinguish between the Classes, find out why instead. (e.g., 50/50 accuracy).

Task 3: The Smoking Gun

We need to know *why* the model thinks a text is AI generated.

- **Saliency Mapping:** Use a library like `SHAP` or `Captum` to highlight the words in an "Imposter" paragraph that most strongly signaled "AI" to your Tier C model.
- **The Findings:** Does the model pick up on specific "AI-isms" (e.g., words like "tapestry," "delve," "testament") or is it looking at the rhythm of the sentence?
- **Error Analysis:** Find 3 samples where the Human was labeled as AI. Was the author being particularly repetitive? Was the AI being particularly brilliant?

Task 4: The Turing Test

1. **The Super-Imposter:** Can you "evolve" a paragraph that bypasses your best detector? You will implement a **Genetic Algorithm (GA)** to optimize a piece of AI-generated text until your classifier labels it as "Human."

   The GA Workflow:

   1. **Initial Population:** Generate 10 "Imposter" paragraphs using Gemini.
   2. **Fitness Function:** The "Human" probability score from your model.
   3. **Selection:** Keep the top 3 paragraphs that look "most human" to the model.
   4. **Mutation (LLM-as-Mutator):** For the next generation, prompt Gemini to "perturb" the winners:
      - *"Rewrite this paragraph to change the rhythm of the sentences while keeping the vocabulary."*
      - *"Introduce a subtle grammatical inconsistency or a rare archaic word."*
   5. **Iteration:** Run this for 5-10 generations.
   6. **The Goal:** Can you reach a >90% "Human" confidence score for a machine-written paragraph?
2. **The Personal Test:** Take your Statement of Purpose (SOP) or a recent essay you wrote. Run it through your detector.
   - If it says you are AI, why? Try to "humanize" your own writing manually to lower the AI score.

- If it says you are Human, try to rewrite a paragraph *manually* to sound like an LLM (overly helpful, structured, repetitive). Can you fool your own machine?
