**AI\_Prompt\_Detector**

This is a Python Program created to detect whether a prompt is AI generated.

# Step 1: Media Type Selection

**Chosen Media Type:** Text

**Rationale:** I selected text because of the distinguishable patterns that differentiate AI-generated text from human-written text.

# Step 2: Collected Examples

Human:

1. I think it’s really important to stay healthy. Staying healthy is key to a good life. – me
2. We need to talk about this. We really need to talk about this.
3. Yesterday was so strange. I saw a dog chasing a squirrel, and it just made me laugh

AI:

1. Artificial intelligence optimizes predictive models, enabling enhanced efficiency, automation, scalability, and unparalleled accuracy in diverse applications.
2. By leveraging advanced neural network architectures, generative models produce coherent, contextually accurate outputs for complex natural language processing.
3. Cutting-edge algorithms facilitate real-time data analysis, streamlining operational workflows, reducing costs, and driving innovative technological advancements.

# Step 3: Brainstorm Specific Traits with AI

**Prompt Used:**

What are some common ways to detect if text is AI generated or not?

**AI Response:**

1. **Overuse of Predictable Patterns**
   * Repetitive sentence structures or phrases.
   * Consistently medium-length sentences without much variation.
2. **Lack of Context Awareness**
   * Logical inconsistencies or contradictions.
   * Failure to maintain coherence across long passages.
3. **Overly Formal or Robotic Tone**
   * Text feels overly polished or unnatural for informal contexts.
   * Avoids contractions and lacks a unique personal voice.
4. **Factual Errors or Hallucinations**
   * Provides incorrect or fabricated facts confidently.
   * Lacks supporting evidence for claims.
5. **Generic or Vague Responses**
   * Responses are broad, repetitive, or lack specific details.
   * Avoids addressing nuanced or subjective aspects of a topic.
6. **Difficulty with Complex or Ambiguous Questions**
   * Struggles to handle abstract, contradictory, or open-ended questions.
   * Responses lack originality or creativity.
7. **Overuse of Keywords or Phrases**
   * Repeats keywords excessively, often in SEO-style writing.
   * Forced use of synonyms in unnatural ways.
8. **Lack of Cultural or Emotional Subtlety**
   * Fails to incorporate cultural references, humor, or sarcasm effectively.
   * Lacks nuanced emotional depth or personal anecdotes.
9. **Speed of Response**
   * Instant generation of coherent and lengthy content may indicate AI use.
10. **Poor Handling of Ambiguity**
    * Overgeneralizes or avoids clarification when dealing with vague or ambiguous prompts.
11. **AI Detection Tools**
    * Use tools like GPTZero, Originality.AI, or OpenAI's AI Text Classifier to analyze text for patterns.
12. **Statistical Linguistic Features**
    * **Perplexity:** AI-generated content often has lower randomness.
    * **Burstiness:** Humans write with more varied sentence lengths and complexity.
13. **Overwhelming Formality or "Fake Sophistication"**
    * Text may overuse technical or sophisticated words to appear professional.
    * Feels artificial rather than naturally written.
14. **Testing Edge Cases**
    * Test the text by asking follow-up questions or posing cultural/niche questions.
15. **Grammar and Style Anomalies**
    * Grammatically correct but lacks natural flow or phrasing.
    * Synonyms or paraphrases used inappropriately for context.

**Traits Selected for Analysis:**  
1. Overly Formal/Robotic Tone

2. Generic or vague responses

3. Overwhelming formality or fake sophistication

**Rationale:** These traits were selected based on their commonality in AI-generated text and their alignment with observable limitations in current AI language models. These were selected after identifying that simpler linguistic pattern often failed to distinguish between AI and human-generated text effectively. By focusing on traits like overly formal tones, generic responses, and fake sophistication, this approach targets specific weaknesses of AI systems while reflecting characteristics unlikely to dominate human writing naturally.

# Step 5: Program Testing

|  |  |  |  |
| --- | --- | --- | --- |
| Text | Prediction | Actual | Correct? |
| Text 1 | Human-written | Human-written | Yes |
| Text 2 | Human-written | Human-written | Yes |
| Text 3 | Human-written | Human-written | Yes |
| Text 4 | AI-generated | AI-generated | Yes |
| Text 5 | AI-generated | AI-generated | Yes |
| Text 6 | AI-generated | AI-generated | Yes |

**Result:** The program correctly identified all six text prompts that were used for testing. The detector performed effectively by utilizing features such as rare word ratio, average word length, and formal phrasing to distinguish between AI-generated and human-written text. Adjustments with the AI confidence threshold made a huge difference in improving the accuracy for borderline cases.

# Step 6: Reflection Report

## Program Performance:

* The program successfully distinguished between AI-generated and human-written text.
* Significant trial and error and back and forth with the AI Large Language Model (LLM) were needed to create a successful program up to this point.
* Early iterations of the program struggled significantly with borderline cases, leading to all the prompts being labeled as “AI-generated” or “Human-written”.
* Refining AI confidence thresholds, in addition to analyzing specific text features significantly improved performance.

## Feature Analysis:

* **Rare Word Ratio:** Texts with a higher ratio of rare words tended to be classified as AI-generated due to their overly polished and unnatural vocabulary.
* **Average Word Length:** Human-written texts exhibited more variability in word length, contrasting with AI's tendency toward uniformity.
* **Vague or Overly Sophisticated Phrasing:** AI-generated texts frequently exhibited an exaggerated formal tone, which was a key indicator in classification.

## Limitations and Improvements:

### Limitations

* + The program was tested on a limited dataset of six text prompts, which restricts generalization to other text styles or domains.
  + Longer texts or prompts outside formal or technical contexts may require additional analysis for accurate detection.

### Improvements

* + Expanding the dataset to include diverse text genres could improve the robustness of detection.
  + Incorporating additional linguistic features, such as sentence structure variability or coherence over longer passages, may enhance accuracy for complex texts.
  + Further fine-tuning of the model itself with additional training on AI-generated and human-written text datasets could yield even better results.

# Conclusion

This project provided valuable insights into the differences between human-written and AI-generated text. I learned a lot about how things as little as the average length of a word within a sentence or prompt can make the difference in AI detection for text. By leveraging text features like rare word ratio, average word length, and vague phrasing, the program effectively identified AI-generated text in test cases. Adjusting confidence thresholds and refining feature-based analysis were critical to the program's success. While the program performed well on this dataset, expanding its scope to include diverse text types and larger datasets will be essential for improving accuracy and scalability. Throughout this project, I developed an appreciation for the complexities of text-based AI detection and its reliance on both computational models and text characteristics to achieve impactful results.