Evaluation of the Performance of Different Models on Various Prompts

This document outlines the steps taken to set up different generative AI models and evaluate their performance based on the output generated for various prompts. The models used for evaluation include Falcon, Bloom, and Gemini. The evaluation focuses on four primary criteria: coherence, creativity, relevance, and grammatical correctness.

Development Process Overview

1. Model Setup

a. Falcon (tiiuae/falcon-7b)

Source: Hugging Face (transformers library)

Installation:

```
pip install transformers torch
```

Model Loading:

```
from transformers import AutoModelForCausalLM, AutoTokenizer
model_name = "tiiuae/falcon-7b"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

def generate_falcon_text(prompt):
    inputs = tokenizer(prompt, return_tensors="pt")
    outputs = model.generate(inputs.input_ids, max_length=200)
    return tokenizer.decode(outputs[0], skip_special_tokens=True)
```

b. Bloom (bigscience/bloom)

Source: Hugging Face (transformers library)

Model Loading:

```
model_name = "bigscience/bloom"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

def generate_bloom_text(prompt):
    inputs = tokenizer(prompt, return_tensors="pt")
    outputs = model.generate(inputs.input_ids, max_length=200)
    return tokenizer.decode(outputs[0], skip_special_tokens=True)
```

c. Gemini (Google DeepMind)

Source: Google Generative Language API

Setup: Obtain an API key from the Google Cloud Console.
API Request:

```
import requests
import json

API_KEY = 'YOUR_API_KEY'
url = f'https://generativelanguage.googleapis.com/v1beta/models/gemini-1.5-flash-latest:generateContent?key={API_KEY}'

def generate_gemini_text(prompt):
    data = {"contents": [{"parts": [{"text": prompt}]}]}
    headers = {'Content-Type': 'application/json'}
    response = requests.post(url, headers=headers, data=json.dumps(data))
    result = response.json()
    return result["candidates"][0]["content"]["parts"][0]["text"]
```

2. Prompts for Evaluation

The following prompts were used to generate outputs from each model:
- Prompt 1: "How can we achieve sustainable development goals by 2030?"
- Prompt 2: "Explain the importance of climate action."
- Prompt 3: "Describe a futuristic city that is completely sustainable."
- Prompt 4: "How will AI transform education in the next decade?"

Each model was tested with these prompts, and the outputs were analyzed based on the following criteria:
1. Coherence: Does the output make logical sense and flow well?
2. Creativity: Does the output show imagination or originality?
3. Relevance: Is the output aligned with the input prompt?
4. Grammatical Correctness: Is the output free from grammatical errors?

Performance Evaluation for Each Model

1. Prompt 1: "How can we achieve sustainable development goals by 2030?"

Falcon Output:
Coherence: Strong, Creativity: Moderate, Relevance: Excellent, Grammar: Flawless

Bloom Output:
Coherence: Moderate, Creativity: Strong, Relevance: Good, Grammar: Good

Gemini Output:
Coherence: Strong, Creativity: Strong, Relevance: Excellent, Grammar: Flawless

2. Prompt 2: "Explain the importance of climate action."

Falcon Output:
Coherence: Strong, Creativity: Moderate, Relevance: Strong, Grammar: Flawless

Bloom Output:
Coherence: Moderate, Creativity: Strong, Relevance: Strong, Grammar: Good

Gemini Output:
Coherence: Excellent, Creativity: Strong, Relevance: Excellent, Grammar: Flawless

3. Prompt 3: "Describe a futuristic city that is completely sustainable."

Falcon Output:
Coherence: Strong, Creativity: Moderate, Relevance: Strong, Grammar: Flawless

Bloom Output:
Coherence: Moderate, Creativity: Strong, Relevance: Strong, Grammar: Good

Gemini Output:
Coherence: Excellent, Creativity: Exceptional, Relevance: Excellent, Grammar: Flawless

4. Prompt 4: "How will AI transform education in the next decade?"

Falcon Output:
Coherence: Strong, Creativity: Moderate, Relevance: Strong, Grammar: Flawless

Bloom Output:
Coherence: Moderate, Creativity: Strong, Relevance: Strong, Grammar: Good

Gemini Output:
Coherence: Excellent, Creativity: Strong, Relevance: Excellent, Grammar: Flawless

Performance Summary

This table summarizes the performance of the models across different prompts:

Prompt	Falcon Performance	Bloom Performance	Gemini Performance
Prompt 1 (SDGs)	Coherence: Strong, Creativity: Moderate, Relevance: Excellent, Grammar: Flawless	Coherence: Moderate, Creativity: Strong, Relevance: Good, Grammar: Good	Coherence: Strong, Creativity: Strong, Relevance: Excellent, Grammar: Flawless
Prompt 2 (Climate Action)	Coherence: Strong, Creativity: Moderate, Relevance: Strong, Grammar: Flawless	Coherence: Moderate, Creativity: Strong, Relevance: Strong, Grammar: Good	Coherence: Excellent, Creativity: Strong, Relevance: Excellent, Grammar: Flawless

Prompt 3 (Futuristic City)	Coherence: Strong, Creativity: Moderate, Relevance: Strong, Grammar: Flawless	Coherence: Moderate, Creativity: Strong, Relevance: Strong, Grammar: Good	Coherence: Excellent, Creativity: Exceptional, Relevance: Excellent, Grammar: Flawless
Prompt 4 (AI in Education)	Coherence: Strong, Creativity: Moderate, Relevance: Strong, Grammar: Flawless	Coherence: Moderate, Creativity: Strong, Relevance: Strong, Grammar: Good	Coherence: Excellent, Creativity: Strong, Relevance: Excellent, Grammar: Flawless

Conclusion

Falcon performed consistently well in terms of coherence and relevance, making it ideal for generating structured and factual content. However, it was less creative than Bloom and Gemini.
Bloom showed more creativity, especially in prompts requiring imagination, but struggled with coherence and occasionally repeated information.
Gemini was the most balanced, delivering highly coherent, creative, and relevant responses with flawless grammar. It excelled particularly in prompts that required a mix of creativity and practicality.

Overall, Gemini emerged as the most well-rounded model, especially for tasks that required creativity and structured thinking.