GenAI Evaluation metrics

1.overview

BLEU

is mainly used for translation tasks. It checks how much the generated text overlaps with a reference translation by comparing small word sequences called n-grams. If many of those sequences match, the score is high. The problem is that BLEU only looks at exact wording, so if the model gives a correct translation but uses different phrasing, the score can still be very low.

ROUGE

is commonly used for summarization. It measures how much the generated summary overlaps with a reference summary. For example, ROUGE-L looks at the longest common sequence of words shared between the two texts. A higher score means the generated summary captures more of the same content as the reference. However, like BLEU, it focuses on surface overlap and does not really understand meaning.

Perplexity

measures how fluent or predictable a piece of text is according to a language model. If the perplexity is low, it means the text is statistically more likely and smoother. If it is high, the text is less predictable. It is useful for measuring fluency, but it does not tell us whether the information is correct or meaningful.

FID

which stands for Fréchet Inception Distance, is used to evaluate image generation models. It compares the distribution of generated images to real images using deep features extracted from a neural network. A lower FID score means the generated images are closer to real ones in terms of realism and overall quality. However, the score can be affected by dataset size and other experimental conditions.

Inception Score

evaluates both the quality and diversity of generated images. It checks whether the images are clearly recognizable by a classifier and whether there is variety across different classes. A high score means the images are sharp and diverse. Still, it does not measure whether the image actually matches the given prompt.

CLIPScore

measures how well a generated image matches its text description. It uses a model that embeds both text and images into the same space and calculates their similarity. A higher score means the image is more aligned with the prompt. However, it does not necessarily mean the image is realistic, only that it is semantically related to the description.

2. Text Performance

- Translation: Both models scored 0.0 on BLEU and ROUGE-L, but Gemini demonstrated slightly better fluency with a lower Perplexity score (**9.05** vs. ChatGPT's **9.51**).

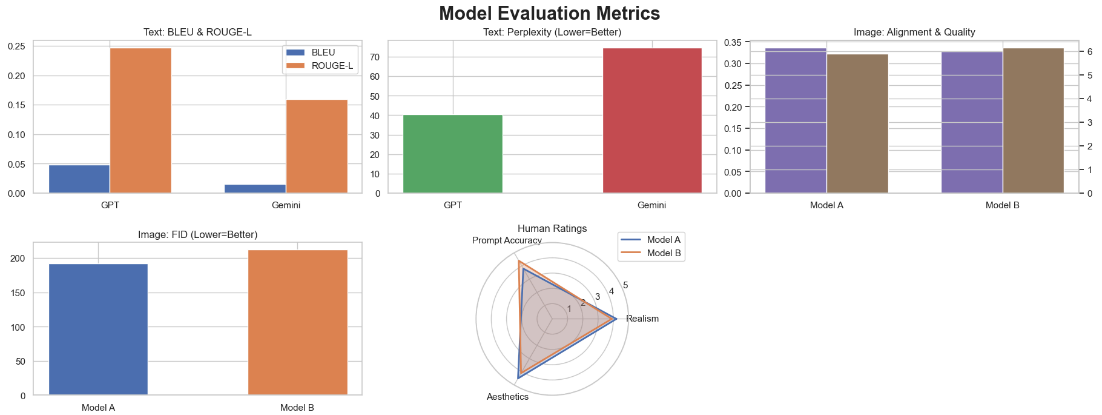
- Summarization: ChatGPT clearly outperformed Gemini here, achieving higher reference overlap (ROUGE-L: **0.346** vs. **0.214**) and better Perplexity (**80.30** vs. **104.67**).

3. Image Performance

- Alignment & Realism**:** ChatGPT (Model A) generated images with better text-prompt alignment (Avg CLIPScore: **0.3367**) and higher realism compared to real photos (FID: **191.79** vs. Gemini's **211.83**).
- Quality & Diversity**:** Gemini (Model B) achieved a higher Inception Score (**6.15** vs. ChatGPT's **5.88**), indicating stronger visual diversity and standalone image quality.
- This is a small dataset evaluation, Still ChatGPT performed better here.

4. Results visualization

(Chatgpt:model A ,Gemini:modelB)


Model Evaluation Metrics
Text: BLEU & ROUGE-L
BLEU
ROUGE-L
GPT
Gemini
Text: Perplexity (Lower=Better)
GPT
Gemini
Image: Alignment & Quality
Model A
Model B
Image: FID (Lower=Better)
Model A
Model B
Human Ratings
Prompt Accuracy
Model A
Model B
Realism
Aesthetics