

GenAI Evaluation metrics

1. Summary

This report applies what we studied of evaluation metrics on two leading Generative AI models—**ChatGPT (Model A)** and **Gemini (Model B)**—across both text and image modalities. The evaluation employs classical Natural Language Processing (NLP) metrics for text tasks and state-of-the-art Computer Vision metrics for image generation, providing a quantitative framework for comparing model efficacy, alignment, and realism.

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.

4. Results visualization

(Chatgpt:model A , Gemini:modelB)

