Model Performance Comparison

Overview

This document provides a detailed comparison of the performance metrics for three NLP models: **BART**, **PEGASUS**, and **FLAN-T5**. The evaluation is conducted on the test dataset, measuring loss, BLEU score, BERTScore (F1, Precision, Recall), and computational efficiency.

Performance Metrics Comparison

Metric	**BART**	**PEGASUS**	**FLAN-T5**
Test Loss	0.9273	0.7015	0.6794
BLEU Score	0.9636	7.3906	1.0743
BERTScore F1	0.8760	0.8909	0.8776
BERTScore Precision	—	0.9045	0.8974
BERTScore Recall	—	0.8785	0.8593
Test Runtime (s)	88.6975	317.7313	153.7434
Samples per Second	14.1150	3.9400	8.1430
Steps per Second	3.5290	0.4940	2.0360

Table 1: Performance metrics for BART, PEGASUS, and FLAN-T5.

Key Observations

- **Test Loss:** FLAN-T5 has the lowest test loss (0.6794), followed closely by PEGASUS (0.7015). BART has the highest loss (0.9273).
- **BLEU Score:** PEGASUS significantly outperforms the other models with a BLEU score of 7.3906, while FLAN-T5 and BART report much lower scores.
- **BERTScore F1:** PEGASUS achieves the highest F1 score (0.8909), indicating better semantic similarity in generated text compared to FLAN-T5 (0.8776) and BART (0.8760).
- **Precision vs. Recall:**
  - PEGASUS has the highest precision (0.9045).
  - FLAN-T5 has slightly lower precision (0.8974) and recall (0.8593).
- **Efficiency:**

  - BART is the fastest model, processing 14.1150 samples per second with a runtime of 88.6975 seconds.
  - PEGASUS is the slowest, with a runtime of 317.7313 seconds and processing 3.9400 samples per second.
  - FLAN-T5 is intermediate in efficiency.

Conclusion

- **PEGASUS** stands out in text generation quality, achieving the best BLEU and BERTScore metrics, but it comes at the cost of higher computational expense.
- **FLAN-T5** offers a balanced performance in both quality and efficiency.
- **BART** is the most efficient, making it suitable for applications where speed is crucial, though its quality metrics are lower.

Depending on your application—whether quality or efficiency is prioritized—different models may be more suitable. If quality is paramount, **PEGASUS** is recommended. For low-latency needs, **BART** might be the best choice.