**1. Overview of Metrics**

* **ROUGE Scores (ROUGE-1, ROUGE-2, ROUGE-L):**  
  These scores assess the overlap between generated text and reference text. Higher values indicate a closer match.
* **BLEU:**  
  A metric typically used in machine translation, with higher scores representing better alignment with reference outputs.
* **BERTScore (Precision, Recall, F1):**  
  These scores use BERT embeddings to measure similarity between the generated text and the reference. F1 is a harmonic mean of precision and recall.

**2. Comparison by Model Size**

**7B Models (Smaller Model Size):**

* **100 Data Samples:**
  + ROUGE-1: **0.3924**
  + ROUGE-2: **0.1241**
  + ROUGE-L: **0.2101**
  + BLEU: **0.0676**
  + BERTScore Precision: **0.8611**
  + BERTScore Recall: **0.8802**
  + BERTScore F1: **0.8704**
* **500 Data Samples:**
  + ROUGE-1: **0.3933** (slight increase)
  + ROUGE-2: **0.1274** (slight increase)
  + ROUGE-L: **0.2101** (approximately the same)
  + BLEU: **0.0718** (noticeable increase)
  + BERTScore Precision: **0.8617**
  + BERTScore Recall: **0.8798**
  + BERTScore F1: **0.8706**

*Observation:* For the 7B model, increasing the evaluation data from 100 to 500 results in small improvements, particularly for ROUGE-1, ROUGE-2, and BLEU.

**13B Models (Larger Model Size):**

* **100 Data Samples:**
  + ROUGE-1: **0.3661**
  + ROUGE-2: **0.1070**
  + ROUGE-L: **0.1846**
  + BLEU: **0.0492**
  + BERTScore Precision: **0.8523**
  + BERTScore Recall: **0.8767**
  + BERTScore F1: **0.8642**
* **500 Data Samples:**
  + ROUGE-1: **0.3677** (very slight increase)
  + ROUGE-2: **0.1105** (increase)
  + ROUGE-L: **0.1884** (increase)
  + BLEU: **0.0570** (noticeable increase)
  + BERTScore Precision: **0.8520**
  + BERTScore Recall: **0.8754**
  + BERTScore F1: **0.8635**

*Observation:* Similarly, the 13B model shows marginal improvements when increasing the evaluation set from 100 to 500 data points, with the BLEU score showing the most noticeable enhancement.

**3. Cross-Model Comparison**

* **Performance:**
  + The **7B models** consistently yield higher scores across all metrics compared to the **13B models**.
    - For example, with 100 data points, ROUGE-1 is approximately 0.3924 for 7B versus 0.3661 for 13B.
    - BLEU scores are 0.0676 for 7B compared to 0.0492 for 13B.
* **Data Sensitivity:**
  + Both model sizes benefit from a larger evaluation sample (500 data points), though the improvements are modest.
  + The improvements are seen in slight increases in ROUGE scores and BLEU scores.
* **BERTScore Comparison:**
  + The BERTScore metrics (Precision, Recall, and F1) for the 7B model are marginally higher than for the 13B model in both evaluation setups.

*Overall, while one might expect the larger 13B model to perform better, the evaluations indicate that in these tests the smaller 7B model achieves higher scores. This could be due to various factors such as model calibration on the evaluation dataset, overfitting differences, or the nature of the evaluation task.*

**4. Summary**

* **7B Models:**  
  Outperform the 13B models in this evaluation across ROUGE, BLEU, and BERTScore metrics. They show slight improvements with increased evaluation data (from 100 to 500 samples).
* **13B Models:**  
  Although larger in size, they score lower in all the compared metrics. They too see modest improvements when moving from 100 to 500 evaluation samples, particularly in BLEU and ROUGE-2.
* **Data Impact:**  
  Increasing the evaluation sample size from 100 to 500 benefits both models, with the 7B model maintaining its performance advantage over the 13B model.

This analysis suggests that in this particular evaluation setup, the smaller 7B model not only competes well but actually outperforms the 13B variant across the standard metrics. It’s important to consider that evaluation results might vary with different datasets, tasks, or even with changes in prompting and fine-tuning strategies.

Let me know if you need further details or additional visualizations of the comparison!