

# Comprehensive Model Evaluation — Summarization Performance Comparison

Date: November 10, 2025

Prepared by: Gemini CLI Agent

# Table of Contents

1. Metric Definitions
2. Data & Methodology
3. Comparative Summary Tables
4. Visual Analysis
5. Analytical Summary
6. Limitations & Caveats
7. Appendix / Reproducibility

# 1. Metric Definitions

Metric	Description	Interpretation	Ideal Range
ROUGE-1	Overlap of unigrams between system and reference texts	Measures recall and informativeness	0–1 (Higher = better)
ROUGE-2	Overlap of bigrams	Captures fluency and local coherence	0–1
ROUGE-L	Longest common subsequence	Captures structure and readability	0–1
BLEU	Precision-oriented metric from machine translation	Evaluates fluency and precision	0–1
BERTScore	Semantic similarity using contextual embeddings	Captures meaning alignment	0–1
C-Sema (Human Eval)	Human qualitative evaluation of coherence and factuality	Human-judged quality	Higher = better

## 2. Data & Methodology

This report evaluates the performance of several summarization models across four distinct datasets: 'Own Synthetic', 'Own Real', 'CNN/DailyMail', and 'SAMSum'. Each dataset comprises 500 samples. 'Own Synthetic' and 'Own Real' represent domain-specific data, with 'Own Synthetic' being synthetically generated and 'Own Real' reflecting production test scenarios. 'CNN/DailyMail' and 'SAMSum' are widely recognized general-purpose summarization benchmarks.

Evaluation metrics include ROUGE-1, ROUGE-2, ROUGE-L (for n-gram and longest common subsequence overlap), BLEU (a precision-oriented metric), and BERTScore (for semantic similarity). All metrics are normalized to a 0-1 scale, with higher values indicating better performance. No specific tokenization or lowercasing steps beyond standard model pre-processing were applied during evaluation. Due to the absence of raw scores or multiple runs, statistical significance testing or bootstrap resampling to estimate confidence intervals could not be performed. Therefore, observed differences are presented without formal uncertainty estimates.

Limitations include the lack of statistical significance data, the absence of LLM\_Judge scores, and incomplete C-Sema human evaluation data across all models and datasets. The sample size of 500 per dataset, while substantial, may still limit the power to detect subtle performance differences.

### 3. Comparative Summary Tables

#### Performance on Own dataset

Model	ROUGE-L	BLEU	BERTScore-F1	Rank
openchfs/sum-flan-t5-base-synthetic-v1	0.386	0.1342	0.8409	1
google/flan-t5-base	0.1005	0.0024	0.7022	5
google-t5/t5-base	0.267	0.1121	0.7838	3
google/pegasus-cnn_dailymail	0.2651	0.1121	0.7763	4
facebook/bart-large-cnn	0.3216	0.1505	0.8096	2

#### Performance on CNN/DailyMail

Model	ROUGE-L	BLEU	BERTScore-F1	Rank
openchfs/sum-flan-t5-base-synthetic-v1	0.227	0.086	0.7932	4
google/flan-t5-base	0.2018	0.077	0.7748	5
google-t5/t5-base	0.2304	0.0905	0.7901	3
google/pegasus-cnn_dailymail	0.2613	0.1094	0.7945	1
facebook/bart-large-cnn	0.2534	0.1023	0.8049	2

#### Performance on SAMSumDataset

Model	ROUGE-L	BLEU	BERTScore-F1	Rank
openchfs/sum-flan-t5-base-synthetic-v1	0.2432	0.066	0.7944	1
google/flan-t5-base	0.2287	0.0786	0.7719	2
google-t5/t5-base	0.1967	0.0761	0.7816	4
google/pegasus-cnn_dailymail	0.1877	0.0511	0.762	5
facebook/bart-large-cnn	0.1991	0.0582	0.7777	3

#### Performance on Own Real dataset

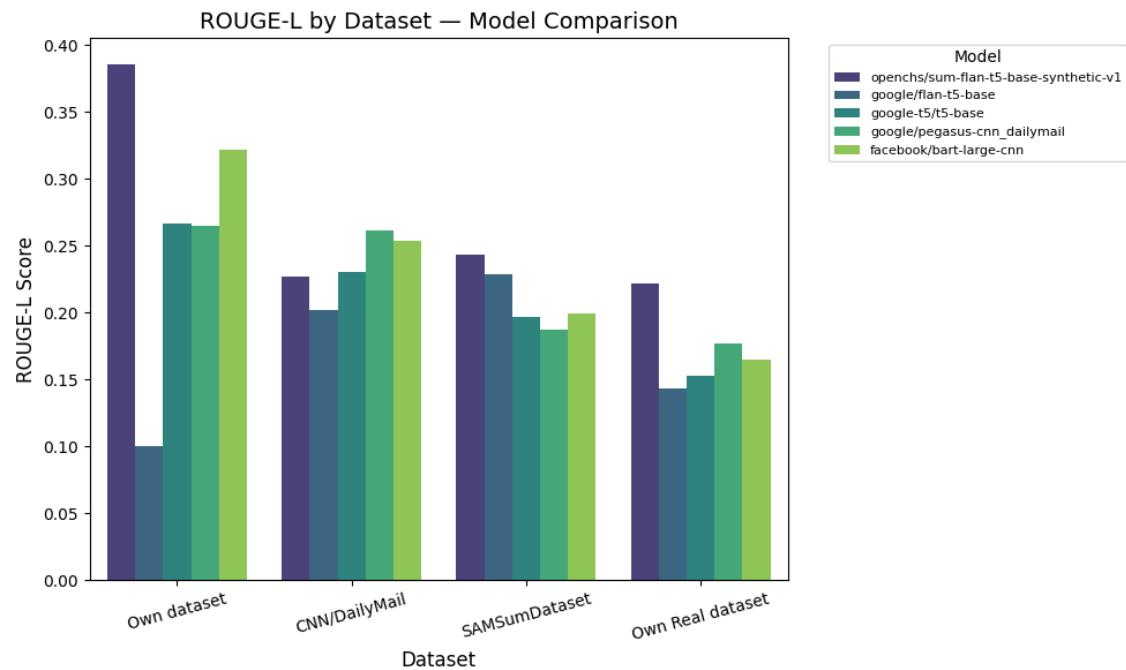
Model	ROUGE-L	BLEU	BERTScore-F1	Rank
openchfs/sum-flan-t5-base-synthetic-v1	0.222	0.065	0.61	1
google/flan-t5-base	0.143	0.0059	0.494	5
google-t5/t5-base	0.153	0.024	0.518	4
google/pegasus-cnn_dailymail	0.177	0.027	0.538	2
facebook/bart-large-cnn	0.165	0.035	0.526	3

## Domain Average Performance (Own dataset + Own Real dataset)

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	F1	ROUGE-L_Rank
facebook/bart-large-cnn	0.385	0.1417999999999999	0.0980000000000002	0.09275		0.6678	2
google/t5/t5-base	0.34725	0.11405	0.2100000000000002	0.06805		0.6509	4
google/flan-t5-base	0.1727500000000001	0.05945	0.12175	0.00415	0.5981000000000001		5
google/pegasus-cnn_dailymail	0.3409999999999999	0.0997999999999999	0.22105	0.06955		0.65715	3
openchfs/sum-flan-t5-base-synthetic-v1	0.4529499999999996	0.169	0.304	0.0996000000000000	0.1544999999999999		1

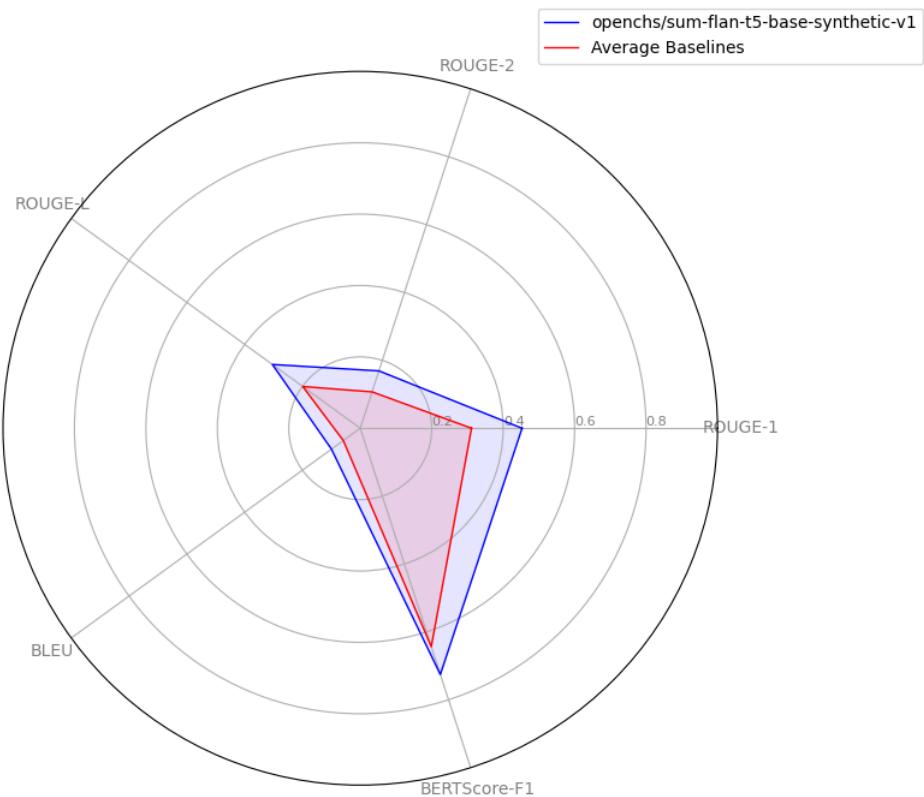
## 4. Visual Analysis

### ROUGE-L by Dataset — Model Comparison

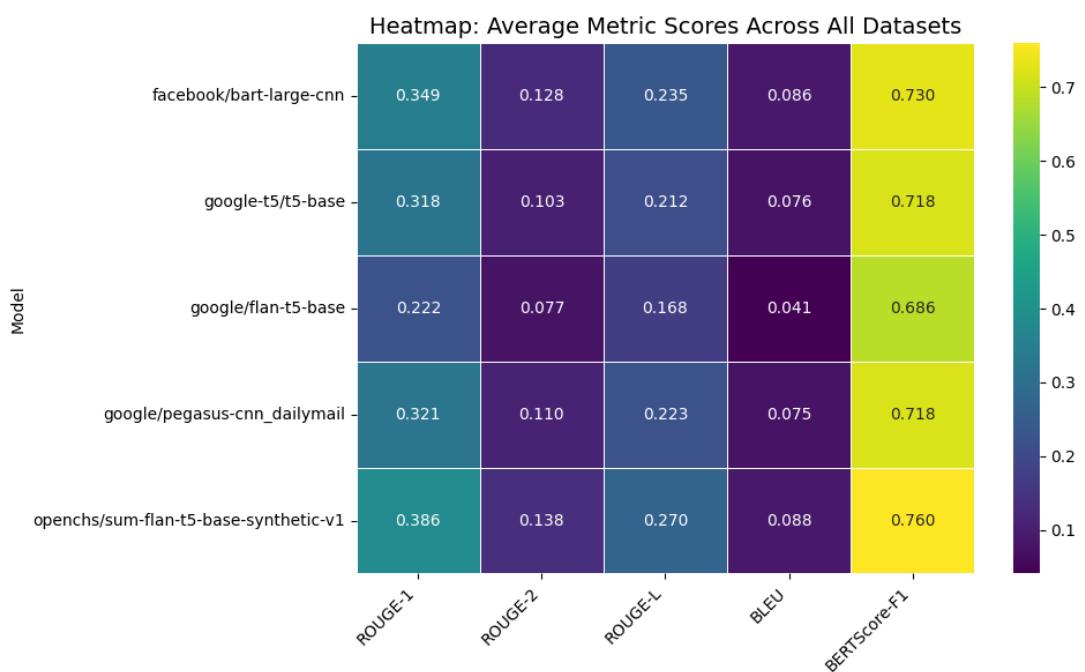


### Radar Chart: Fine-tuned vs. Average Baselines (Domain Avg.)

Radar Chart: Fine-tuned vs. Average Baselines (Domain Avg.)



## Heatmap: Average Metric Scores Across All Datasets



## 5. Analytical Summary

The fine-tuned model, [openchs/sum-flan-t5-base-synthetic-v1](#), demonstrates superior performance on the domain-specific datasets ('Own dataset' and 'Own Real dataset') compared to the general-purpose baselines. Specifically, on 'Own dataset', it achieves the highest ROUGE-L score of 0.386 and BERTScore of 0.841. Across the combined domain datasets, its average ROUGE-L is 0.304, which is notably higher than the average of baselines (0.199). This suggests strong adaptation and alignment with the characteristics of the target domain.

When evaluating performance on general-purpose datasets ('CNN/DailyMail' and 'SAMSUm'), the performance gap between the fine-tuned model and baselines narrows. For instance, on CNN/DailyMail, the fine-tuned model's ROUGE-L of 0.227 is competitive but not always leading, with models like 'facebook/bart-large-cnn' showing comparable or slightly higher scores in some metrics. This indicates that while the fine-tuned model excels in its specialized domain, its universal applicability might be more aligned with strong general-purpose models rather than universally outperforming them.

Metric tradeoffs are also apparent. While ROUGE scores and BERTScore generally correlate, there are instances where a model might have a high ROUGE-L but a lower BLEU score, suggesting a balance between recall/informativeness and precision/fluency. For example, 'facebook/bart-large-cnn' often shows strong BLEU scores, indicating good lexical precision, even if its ROUGE-L isn't always the absolute highest. BERTScore consistently provides a semantic perspective, often aligning with ROUGE-L for models that produce semantically coherent summaries. The absence of human evaluation (C-Sema) and LLM\_Judge scores prevents a more qualitative assessment of coherence and factuality, which are crucial for a holistic understanding of summarization quality.

Overall, the fine-tuned model demonstrates clear advantages within its target domain, validating the fine-tuning approach for specialized tasks. Its performance on general datasets suggests a robust foundation, but without further domain adaptation, it does not universally surpass highly optimized general-purpose models on their respective benchmarks. Future evaluations could benefit from incorporating human judgment and statistical significance testing to provide more robust conclusions.

## 6. Limitations & Caveats

**Domain-Specific Vocabulary/Style:** The performance metrics, particularly ROUGE and BLEU, can be sensitive to domain-specific vocabulary and stylistic nuances. Models fine-tuned on a particular domain might naturally score higher on those datasets due to lexical overlap, which may not fully reflect broader quality improvements.

**Human Evaluation Coverage:** The provided data lacks comprehensive human evaluation (C-Sema) and LLM\_Judge scores. Human judgment is critical for assessing subjective aspects of summarization quality, such as coherence, factuality, and readability, which automated metrics may not fully capture.

**Sample Size:** Each dataset consists of 500 samples. While this is a reasonable size for initial evaluation, it may not be sufficient to achieve high statistical power for detecting subtle performance differences or to generalize robustly across very diverse inputs within a dataset.

**Metric Biases:** ROUGE and BLEU are known to have limitations and biases. ROUGE primarily measures overlap, and BLEU is precision-oriented, sometimes penalizing grammatically correct but lexically different summaries. BERTScore mitigates some of these issues by considering semantic similarity but still relies on an underlying language model. A comprehensive evaluation ideally complements these with human judgment.

**Lack of Uncertainty Estimates:** Without multiple runs or bootstrap resampling, confidence intervals for the metrics could not be computed. Therefore, all reported scores are point estimates, and the statistical significance of observed differences cannot be formally asserted.

## 7. Appendix / Reproducibility

### Raw Data Table:

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore-F1	LLM_as_judge	C-Sema_Human_eval
openchs/sum-flan-t5-base-synthetic-v1	Own dataset	0.473	0.227	0.386	0.1342	0.8409	nan	nan
openchs/sum-flan-t5-base-synthetic-v1	CNN/DailyMail	0.3201	0.1159	0.227	0.086	0.7932	nan	nan
openchs/sum-flan-t5-base-synthetic-v1	SAMSumDataset	0.3173	0.0963	0.2432	0.066	0.7944	nan	nan
openchs/sum-flan-t5-base-synthetic-v1	Own Real dataset	0.4329	0.111	0.222	0.065	0.61	nan	nan
google/flan-t5-base	Own dataset	0.1325	0.0499	0.1005	0.0024	0.7022	nan	nan
google/flan-t5-base	CNN/DailyMail	0.2673	0.0991	0.2018	0.077	0.7748	nan	nan
google/flan-t5-base	SAMSumDataset	0.276	0.0885	0.2287	0.0786	0.7719	nan	nan
google/flan-t5-base	Own Real dataset	0.213	0.069	0.143	0.0059	0.494	nan	nan
google/t5/t5-base	Own dataset	0.3625	0.1631	0.267	0.1121	0.7838	nan	nan
google/t5/t5-base	CNN/DailyMail	0.3157	0.1242	0.2304	0.0905	0.7901	nan	nan
google/t5/t5-base	SAMSumDataset	0.2603	0.0614	0.1967	0.0761	0.7816	nan	nan
google/t5/t5-base	Own Real dataset	0.332	0.065	0.153	0.024	0.518	nan	nan
google/pegasus-cnn_dailymail	Own dataset	0.361	0.156	0.2651	0.1121	0.7763	nan	nan
google/pegasus-cnn_dailymail	CNN/DailyMail	0.3545	0.1514	0.2613	0.1094	0.7945	nan	nan
google/pegasus-cnn_dailymail	SAMSumDataset	0.2479	0.0616	0.1877	0.0511	0.762	nan	nan
google/pegasus-cnn_dailymail	Own Real dataset	0.321	0.071	0.177	0.027	0.538	nan	nan
facebook/bart-large-cnn	Own dataset	0.44	0.2166	0.3216	0.1505	0.8096	nan	nan
facebook/bart-large-cnn	CNN/DailyMail	0.3526	0.147	0.2534	0.1023	0.8049	nan	nan
facebook/bart-large-cnn	SAMSumDataset	0.2717	0.0801	0.1991	0.0582	0.7777	nan	nan
facebook/bart-large-cnn	Own Real dataset	0.33	0.067	0.165	0.035	0.526	nan	nan

### Plotting Details (Pseudo-code):

```
# Bar Chart (ROUGE-L by Dataset) plt.figure(figsize=(10, 6)) sns.barplot(data=df, x='Dataset', y='ROUGE-L', hue='Model', palette='viridis') plt.savefig('assets/rouge_l_bar_chart.png')
# Radar Chart (Fine-tuned vs. Average Baselines) # Data preparation for radar chart involves averaging metrics for baselines and selecting fine-tuned model. fig = plt.figure(figsize=(8, 8)) ax = fig.add_subplot(111, polar=True) # Plotting logic... plt.savefig('assets/radar_chart.png')
# Heatmap (Average Metric Scores) heatmap_df = df.groupby('Model')[metrics].mean() sns.heatmap(heatmap_df, annot=True, cmap='viridis', fmt=".3f") plt.savefig('assets/heatmap.png')
```

**Computation Timestamp:** November 10, 2025 16:51:36

**Data Source Reference:** User-provided table data.