Evaluating Small Language Models for News Summarization

Implications and Factors Influencing Performance

Aman Agarwal Nakul Siwach Himanshu Shivhare

International Institute of Information Technology, Bangalore

Outline

Motivation

- **Problem**: Small Language Models (SLMs) are gaining importance for resource-constrained environments
- **Challenge**: Which learning approach is best for SLMs in news summarization?
  - Zero-shot?
  - Few-shot?
  - Full fine-tuning?
  - Parameter-efficient fine-tuning (LoRA)?
- **Gap**: Conventional wisdom says "always fine-tune" - but is this true?

Key Question

Does the optimal approach depend on model architecture?

Research Questions

1. **RQ1**: How do different learning approaches affect performance across SLM architectures?
2. **RQ2**: Does fine-tuning improve or degrade instruction-tuned models?
3. **RQ3**: What are the minimum model size requirements for effective LoRA fine-tuning?
4. **RQ4**: How do performance-efficiency trade-offs vary across approaches?

Experimental Setup

Models Evaluated (3)

- FLAN-T5-Small (80M)
- FLAN-T5-Base (250M)
- BART-Base (140M)

Learning Approaches (4)

- Zero-shot
- Few-shot (3 examples)
- Full fine-tuning
- LoRA fine-tuning

Dataset

- CNN/DailyMail
- 1,000 training samples
- 100 test samples

Metrics

- ROUGE-1, ROUGE-2, ROUGE-L
- BERTScore

Total Experiments

3 models $\times$ 4 approaches =
12 evaluations

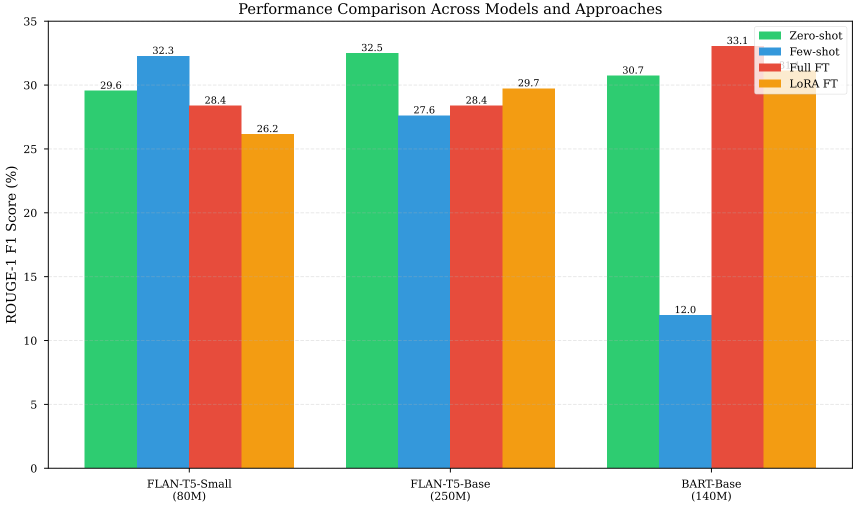
Model Architectures

Model	Parameters
FLAN-T5-Small	80M
FLAN-T5-Base	250M
BART-Base	140M

Key Difference:

- **FLAN-T5**: Pre-trained to follow instructions (zero-shot capable)
- **BART**: Pre-trained for text generation (needs task-specific training)

Overall Performance Comparison



Key Observation

Different models perform best with different approaches!

Top 3 Results

Rank	Model + Approach	ROUGE-1	BERTScore
	BART-Base Full FT	**33.06%**	**87.55%**
	FLAN-T5-Base Zero-shot	**32.50%**	**87.19%**
	FLAN-T5-Small Few-shot	**32.29%**	**87.68%**

Surprising Finding

FLAN-T5-Base zero-shot (no training!) is only 0.56 pp behind the best result!

Results by Model

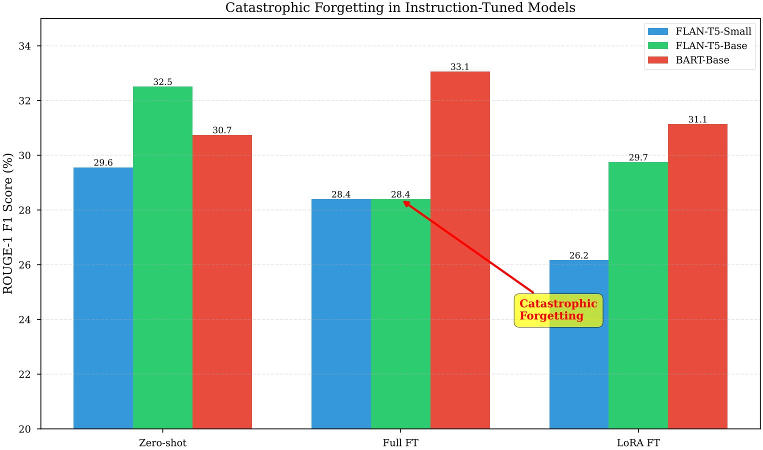
Model	ZS	FS	FFT	LFT
FLAN-T5-Small	29.56	**32.29**	28.39	26.16
FLAN-T5-Base	**32.50**	27.63	28.39	29.75
BART-Base	30.74	11.99	**33.06**	31.13

Table: ROUGE-1 F1 scores (%). ZS: Zero-shot, FS: Few-shot, FFT: Full FT, LFT: LoRA FT

Pattern:

- FLAN-T5: Best with zero-shot/few-shot
- BART: Best with fine-tuning

Finding 1: Catastrophic Forgetting



Critical Discovery

Fine-tuning **degrades** instruction-tuned models (FLAN-T5)!

Finding 1: Catastrophic Forgetting (Details)

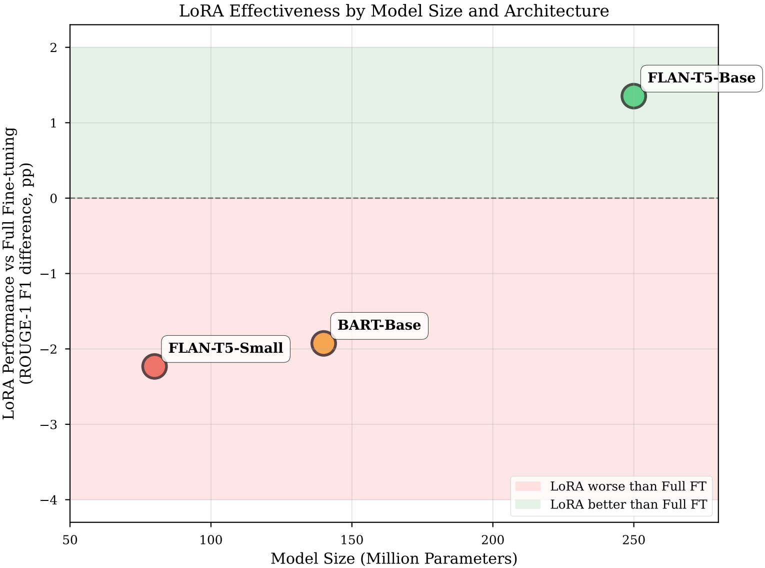
FLAN-T5 Models - Fine-tuning Makes Them WORSE!

Model	**Zero-shot**	**Fine-tuned**	**Change**
FLAN-T5-Small	29.56%	28.39%	-1.17 pp
FLAN-T5-Base	32.50%	28.39%	-4.11 pp
BART-Base	30.74%	33.06%	+2.32 pp

Explanation:

- Instruction-tuned models learn general instruction-following
- Task-specific fine-tuning overwrites this capability
- Result: Performance degrades despite more training!

Finding 2: LoRA Size Thresholds



Threshold Discovery

Finding 2: LoRA Size Thresholds (Details)

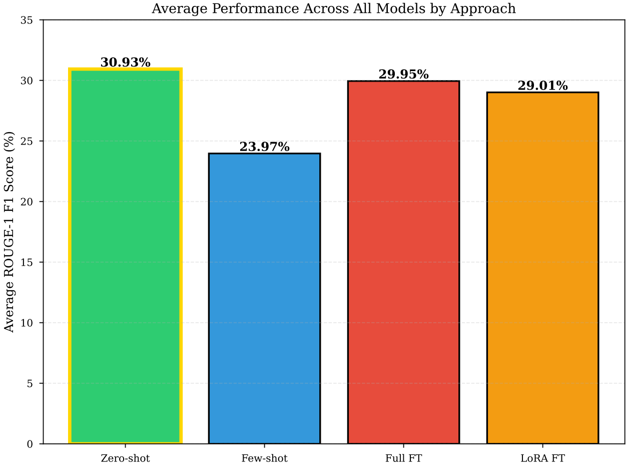
LoRA Performance vs. Full Fine-tuning

Model	Size	LoRA vs FFT
FLAN-T5-Small	80M	-2.23 pp
BART-Base	140M	-1.93 pp
FLAN-T5-Base	250M	+1.36 pp

Key Insights:

- Smaller models (80M): LoRA fails
- Medium models (140M+): LoRA works for generation-focused
- Larger models (250M+): LoRA works well for instruction-tuned
- LoRA can even **outperform** full fine-tuning!

Finding 3: Zero-shot Superiority



Surprising Result

Zero-shot achieves **highest average** performance across all models!

Finding 3: Zero-shot Superiority (Details)

Average Performance Across All Models

Approach	Avg ROUGE-1	Rank
Zero-shot	**30.93%**	**1st**
Full Fine-tuning	29.95%	2nd
LoRA Fine-tuning	29.01%	3rd
Few-shot	23.97%	4th

Implication:

- Training is **not always beneficial**!
- Zero-shot should be the baseline, not an afterthought
- Challenges conventional "always fine-tune" assumption

Finding 4: Few-shot Unreliability

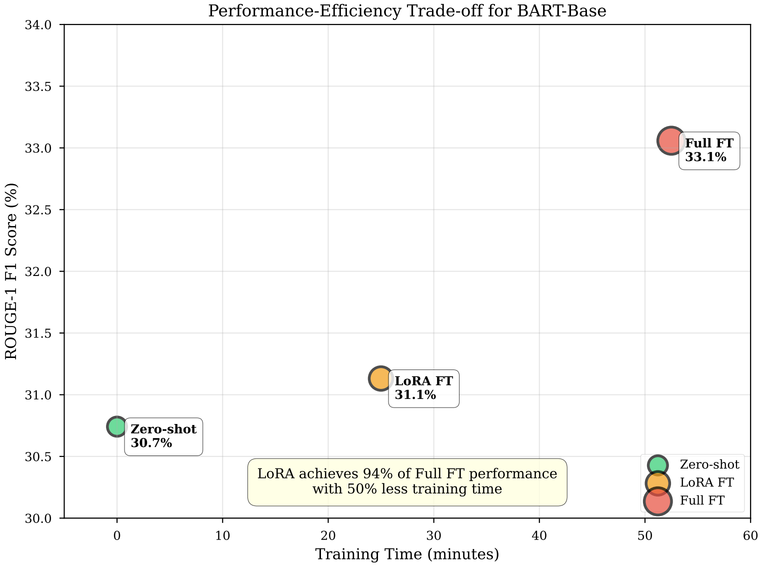
Few-shot Performance is Highly Variable

Model	Few-shot	vs Zero-shot
FLAN-T5-Small	32.29%	+2.73 pp
FLAN-T5-Base	27.63%	-4.87 pp
BART-Base	11.99%	-18.75 pp
Range	**20.3 percentage points!**	

Warning

Few-shot is too unpredictable for production systems!

Finding 5: Efficiency-Performance Trade-off



LoRA Sweet Spot

Finding 5: LoRA Efficiency Benefits

LoRA vs. Full Fine-tuning for BART-Base

Metric	Full FT	LoRA FT	Savings
ROUGE-1	33.06%	31.13%	-1.93 pp (6%)
Training Time	45-60 min	20-30 min	50%
Parameters Trained	140M	0.5M	99.6%
Memory	10-12 GB	6-8 GB	40%

Trade-off: Lose 1.93 pp performance, gain massive efficiency!
Verdict: Excellent for practical deployment!

Architecture-Approach Matching

Instruction-Tuned Models
(FLAN-T5, T0, etc.)

Recommended:

- Zero-shot
- Few-shot (if validated)
- LoRA FT (250M+ only)

Not Recommended:

- Full fine-tuning
- LoRA FT (¡250M)

Generation-Focused Models
(BART, Pegasus, etc.)

Recommended:

- Full fine-tuning
- LoRA fine-tuning
- Zero-shot (no data)

Not Recommended:

- Few-shot learning

Deployment Scenarios

Scenario	Best Choice	ROUGE-1
No training data	FLAN-T5-Base Zero-shot	32.50%
Max performance	BART-Base Full FT	33.06%
Efficiency focus	BART-Base LoRA FT	31.13%
Limited compute	FLAN-T5-Small Few-shot	32.29%
Multiple tasks	FLAN-T5-Base Zero-shot	32.50%

Key Message

Match approach to your constraints, not conventional wisdom!

Research Contributions

1. **Catastrophic Forgetting Discovery**
   - First systematic demonstration in instruction-tuned models
   - Challenges "always fine-tune" assumption
2. **LoRA Size Thresholds**
   - Identified minimum requirements: 250M+ (instruction), 140M+ (generation)
   - Architecture-dependent success criteria
3. **Architecture-Approach Matching Framework**
   - Clear guidelines for approach selection
   - Evidence-based deployment recommendations
4. **Zero-shot Superiority Evidence**
   - Highest average performance (30.93%)
   - Training not always beneficial

Contributions (cont.)

5. **Few-shot Unreliability Documentation**
    - 20.3 pp performance range
    - Not suitable for production
6. **Efficiency-Performance Quantification**
    - LoRA: 94% performance, 50% time
    - Clear trade-off analysis

Impact

Provides evidence-based guidelines that challenge conventional assumptions and enable better deployment decisions

Key Takeaways

1. **Architecture determines optimal approach**
   - Instruction-tuned → zero-shot/few-shot
   - Generation-focused → fine-tuning
2. **Fine-tuning can hurt instruction-tuned models**
   - FLAN-T5-Base: 32.50% (zero) → 28.39% (fine-tuned)
3. **LoRA needs sufficient size**
   - Fails at 80M, works at 140M+
4. **Zero-shot often best**
   - Highest average: 30.93%
5. **Few-shot unreliable**
   - Range: 11.99% - 32.29%

Future Work

- **Expand model coverage**
  - Test decoder-only models (GPT-style, Llama)
  - Evaluate at different size points (100M, 150M, 200M)
- **Mitigate catastrophic forgetting**
  - Develop adaptive fine-tuning methods
  - Test regularization approaches
- **Optimize LoRA configuration**
  - Model-specific hyperparameter tuning
  - Different rank selections
- **Real-world deployment studies**
  - Edge device evaluation
  - User preference studies

Main Message

The optimal learning approach depends on model architecture, not size alone.

Match approach to architecture for best results!

Best Overall:
BART-Base Full FT
33.06% ROUGE-1

Best Without Training:
FLAN-T5-Base Zero-shot
32.50% ROUGE-1

Thank You!

Aman Agarwal, Nakul Siwach, Himanshu Shivhare
International Institute of Information Technology, Bangalore