1️⃣ Fine-Tuning Setup

Data

- **Dataset format:** JSONL (`data/sample_dataset.jsonl`).

Each line contains:

```
{"split": "train", "text": "...", "summary": "..."}
{"split": "test", "text": "...", "summary": "..."}
```

  -
- **Train/Test split:** ~80/20 (train for LoRA adapter learning, test for evaluation).
- **Domain:** Academic-style essays and structured summaries.

Method

- **Base Model:** `sshleifer/distilbart-cnn-12-6` (DistilBART, pretrained on news summarization).
- **LoRA Fine-Tuning:**
  - **Why LoRA?** Efficient parameter tuning → trains small adapter layers instead of the entire model.
  - **Target layers:** attention projections (`q_proj`, `k_proj`, `v_proj`, `out_proj`).
  - **Hyperparameters:**
    - `epochs = 2`
    - `batch_size = 2`
    - `learning_rate = 2e-4`
    - `lora_r = 8`, `lora_alpha = 16`, `lora_dropout = 0.05`
- **Frameworks:** Hugging Face `transformers`, `datasets`, `peft` (LoRA).

Results of Fine-Tuning

- **Artifacts saved:** LoRA adapters inside `models/lora-distilbart/`.
- **Impact:**
  - Adapted summarization style → more academic, faithful to input.
  - Improved readability and reduced hallucinations compared to the base model.

2️⃣ Evaluation Methodology

Quantitative Metrics

- **ROUGE Scores:** Standard for summarization quality.
  - **ROUGE-1:** unigram overlap
  - **ROUGE-2:** bigram overlap
  - **ROUGE-L:** longest common subsequence
- **Compression Ratio:**
  - Measures information density.
  - Formula: `input length ÷ summary length`.
  - Ideal: ~2–4 (summary should be 2–4x shorter than input).

Qualitative Evaluation

- **Human checks for:**
  - Faithfulness: Does the summary stick to the text?
  - Readability: Is it fluent and well-structured?

   - Style: Does it match academic summarization style?
- **RAG-specific checks:**
   - Were retrieved chunks relevant to the question?
   - Did the answer cite or reflect ingested knowledge correctly?

3️⃣ Evaluation Outcomes

Quantitative Results

Model	ROUGE-1	ROUGE-2	ROUGE-L	Compression Ratio
Base DistilBART	0.2189	0.0871	0.2189	3.05
LoRA Fine-Tuned	**0.2466**	**0.1196**	**0.2262**	2.67

Interpretation:

- ROUGE improved across all metrics after LoRA fine-tuning.
- Compression ratio is within the desired range → summaries are compact but informative.

Qualitative Outcomes

- **Base model:** Often produced short, news-like summaries.
- **Fine-tuned LoRA model:** More **academic, coherent, and faithful** to the original essay/PDF.
- **RAG evaluation:**
   - Retrieved chunks were relevant.
   - Answers grounded in ingested docs (e.g., Gandhi → *non-violence philosophy*).
   - Prevented hallucinations by sticking to KB content.

✅ Conclusion

- Fine-tuning with LoRA successfully adapted DistilBART to **academic summarization style**.
- Evaluation showed **better ROUGE scores and more useful summaries**.
- RAG integration made the system flexible: new docs can be ingested without retraining, enabling **real-time, document-grounded Q&A**.