CMPSC 497 Final Project – Manay Lodha

Introduction

Automatic generation of concise "approach" descriptions for research problems can greatly accelerate literature reviews and proposal writing. In this project, we fine-tune an instruction-tuned seq2seq model (Flan-T5) to take as input the combined Abstract + Introduction + Conclusion (AIC) text of a scientific paper and output a short, human-readable "approach" summary. Our goal is to assess how well a medium-sized open-source LLM can learn this extreme summarization task from a domain-specific dataset.

Dataset Construction

We leverage the SciTLDR dataset, which contains 5.4 K TLDR summaries across 3.2 K papers, with an author-written and an expert-derived summary for each entry [Hugging Face](). We use the "AIC" configuration (Abstract + Introduction + Conclusion) as our prompt and the expert-derived TLDR as the target.

1. **Loading & filtering**

python

CopyEdit

raw = load_dataset("allenai/scitldr", "AIC", split="train+validation+test")

We concatenate the "source" sentences into one prompt and "target" sentences into one summary, then filter pairs where the prompt has 20–300 words and the target 10–100 words.

2. **Size & split**

    - **Total examples built:** ~5 400
    - **Train/Val/Test split:** 80 % / 10 % / 10 % ⇒ ~4 320 / 540 / 540 examples

All examples are saved as JSONL (data/train.jsonl, etc.) for downstream processing.

Methodology

Tokenization & Data Pipeline

- **Tokenizer:** AutoTokenizer.from_pretrained("google/flan-t5-base")

- **Maximum lengths:** prompt = 128 tokens, target = 512 tokens
- We apply HF's .map(...) to tokenize and produce input_ids and labels.

Model Selection & Training

- **Base model:** google/flan-t5-base (250 M parameters) Hugging Face
- **Training framework:** 🤗 Transformers Trainer on a single GPU
- **Hyperparameters:**

Parameter	Value
Batch size	4
Learning rate	5×10^{-5}
Epochs	3
FP16	True
Eval & save strategy	per-epoch

Evaluation Metrics & Experiments

We evaluate on the held-out test split (~540 examples) with:

1. **ROUGE** (1, 2, L, Lsum) to measure n-gram overlap.
2. **Perplexity (PPL)** computed via cross-entropy on references.
3. **Qualitative samples** to inspect generation behavior.

Results

Quantitative Results

Metric	Score
ROUGE-1	0.1841
ROUGE-2	0.0753

Metric	Score
ROUGE-L	0.1470
ROUGE-Lsum	0.1489
Avg PPL	**1,777,852.93**

These low overlap scores and extremely high perplexity indicate the model predominantly copies or truncates the prompt rather than generating novel, concise summaries.

Qualitative Analysis

Sample prompt → generated summary → reference:

PROMPT:
We introduce a new procedural dynamic system that can generate a variety of shapes that often appear as curves… (truncated)

GENERATED:
We introduce a new procedural dynamic system that can generate a variety of shapes that often appear as curves… We introduce a new procedural dynamic system…

REFERENCE:
A new, very simple dynamic system is introduced that generates pretty patterns; properties are proved and possibilities are explored.

We observe near-verbatim copying of the prompt and omission of concise paraphrasing.

Discussion

- **Copying behavior:** The model often echoes the input, leading to poor abstractive summarization.
- **High PPL:** Indicates the fine-tuned model assigns very low probability to the reference summaries, suggesting over-reliance on the input distribution.
- **Possible causes:**
  - Insufficient training epochs/dataset size for the model to generalize.
  - Learning rate may be too high, causing early convergence to copying.
  - Lack of explicit instruction prompting (simply feeding raw AIC text).

Conclusion & Future Work

We demonstrated an end-to-end fine-tuning pipeline for FLAN-T5 on the SciTLDR AIC summarization task, complete with dataset construction, training, and evaluation. While the current model underperforms (ROUGE < 0.19, PPL $\gg 10^3$), this establishes a baseline.

Next steps:

- **Prompt engineering:** Prepend explicit instructions (e.g. "Summarize the methods in one sentence:") to guide abstraction.
- **Longer training / larger model:** Experiment with google/flan-t5-large or more epochs with a lower learning rate.
- **Regularization:** Apply label smoothing or dropout to mitigate copying.
- **Augmented dataset:** Incorporate additional summarization resources (e.g. abstract-to-TLDR pairs) to enrich training.

This project lays the groundwork for improved automated generation of scientific approach summaries.