Fine-Tuning FLAN-T5 for Summarization & Measuring Forgetting

Authors: Shaunak Kapur & Pranav Krishnan
Course: COE 379L

1. Introduction and Project Statement

The goal of this project is to fine-tune a small pretrained language model so that it can generate concise and clear one-sentence summaries of product reviews. Summarization is a well-defined text-to-text task where the model must condense information while preserving meaning, which makes it a useful setting for studying how targeted fine-tuning affects model behavior. In the context of e-commerce and content curation, automated summarization enables users to quickly understand the key points of lengthy reviews without reading the full text, improving information retrieval and decision-making efficiency.

Product reviews are particularly suitable for this task because they follow a consistent pattern, vary in length, and often contain a human-written summary that serves as an ideal reference target. These reviews provide a natural supervised learning setting where we can train models to generate summaries that match human expectations.

The project examines an important side effect of model specialization: forgetting. When a model is fine-tuned on a narrow domain, its performance on tasks outside that domain may degrade. This type of behavior is relevant in many real-world deployments where models are continuously adapted for specific applications. By training the model on summarization and then testing its ability to answer basic factual questions, we can observe whether the specialization comes at the cost of reduced general knowledge.

This project has two connected objectives: first, improving the model's capability on the summarization task through targeted fine-tuning; and second, measuring whether that improvement weakens the model's performance on unrelated prompts. We fine-tuned FLAN-T5-Small on the Amazon Fine Food Reviews dataset, evaluated summarization quality using ROUGE metrics, and assessed potential forgetting through a set of general knowledge questions.

2. Data Sources and Technologies Used

2.1 Data Sources

Amazon Fine Food Reviews Dataset

We used the Amazon Fine Food Reviews dataset, which is publicly available on Hugging Face under the identifier jhan21/amazon-food-reviews-dataset. The original dataset contains approximately 568,000 reviews spanning over 10 years, up to October 2012. Each review includes the full review text and a human-written summary, making it ideal for supervised summarization training.

For computational efficiency, we sampled 20,000 reviews from the full dataset. We filtered out reviews longer than 512 characters and removed rows with missing Text or Summary fields to ensure data quality. The final dataset was split into three subsets: 16,000 samples for training (80%), 2,000 samples for validation (10%), and 2,000 samples for testing (10%).
The dataset is accessible at
https://huggingface.co/datasets/jhan21/amazon-food-reviews-dataset.

Forgetting Evaluation Dataset
To measure potential forgetting of general knowledge, we created a custom test set consisting of 8 general knowledge questions covering multiple categories: geography, basic facts, science, literature, and mathematics. These questions were designed to test the model's retention of fundamental factual knowledge after fine-tuning on the domain-specific summarization task.

2.2 Technologies Used

Model Architecture

We selected google/flan-t5-small as our base model. FLAN-T5 is an encoder-decoder transformer architecture with approximately 60 million parameters. It was pretrained on instruction-following tasks, which makes it well-suited for text-to-text generation tasks like summarization. The relatively small size of this model (compared to larger variants like FLAN-T5-Base or FLAN-T5-Large) makes it computationally feasible to fine-tune on limited hardware while still demonstrating meaningful learning behavior.
The model is available on Hugging Face at https://huggingface.co/google/flan-t5-small.

Software Libraries

We utilized several key libraries from the Python ecosystem:

- transformers (Hugging Face): Model loading, tokenization, and training infrastructure
- datasets (Hugging Face): Data loading and preprocessing utilities
- evaluate: Framework for computing evaluation metrics
- rouge_score: Implementation of ROUGE metrics for summarization evaluation
- pytorch: Deep learning framework providing the underlying computational infrastructure
- accelerate: Distributed training support and optimization utilities

Compute Platform

All training and evaluation was executed on Google Colab using a T4 GPU with 16GB of VRAM. This cloud-based platform provided sufficient computational resources for fine-tuning our model while remaining accessible and cost-effective for academic projects.

3. Methods Employed

3.1 Data Preprocessing

We loaded the Amazon Fine Food Reviews dataset directly from Hugging Face using the datasets library. The preprocessing pipeline involved several steps to prepare the data for training:

First, we filtered the dataset to remove reviews longer than 512 characters to ensure consistent input lengths and avoid memory issues during training. We also removed any rows where either the Text field (review content) or Summary field (target label) was missing.
Next, we sampled 20,000 reviews from the full dataset to balance computational efficiency with sufficient training data. These 20,000 samples were randomly split into training (16,000 samples), validation (2,000 samples), and test (2,000 samples) sets.

For tokenization, we used the T5 tokenizer with a maximum input length of 256 tokens and a maximum target length of 32 tokens. We added the prefix "Summarize this review: " to each input text to provide explicit task instruction to the model, following best practices for instruction-tuned models. Padding and truncation were applied as necessary to handle variable-length inputs.

3.2 Fine-Tuning Strategy

We employed supervised learning with teacher forcing, a standard approach for sequence-to-sequence text generation tasks. The training framework used Hugging Face's Seq2SeqTrainer with Seq2SeqTrainingArguments to manage the training loop and evaluation. All training and evaluation was executed on Google Colab using a T4 GPU (16GB VRAM) for accelerated computation.

Hyperparameters
We trained the model for 2 epochs with a batch size of 4 per device. The batch size was reduced from an initial value of 8 due to GPU memory constraints. We used a learning rate of 2e-4 with the AdamW optimizer (default configuration) and applied weight decay of 0.01 to prevent overfitting.

For mixed precision training, we used BF16 format on modern GPUs with FP32 as a fallback for compatibility. During generation, we set a maximum length of 32 tokens, used beam search with 4 beams, and enabled early stopping to terminate generation when appropriate.

The model was evaluated on the validation set after each epoch, and we saved the best model checkpoint based on ROUGE-1 score, which measures unigram overlap between generated and reference summaries.

3.3 Evaluation Metrics

ROUGE Metrics
We evaluated summarization quality using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, which are standard for assessing automatic summarization systems:

- ROUGE-1: Measures unigram overlap between generated and reference summaries, capturing the percentage of individual words that match
- ROUGE-2: Measures bigram overlap, providing insight into how well the model preserves two-word phrases
- ROUGE-L: Based on the Longest Common Subsequence (LCS), this metric captures sentence-level structure similarity
- ROUGE-Lsum: A sentence-level LCS-based F-score that is particularly suited for multi-sentence summaries

All ROUGE scores are reported as percentages on a 0 to 100 scale. Higher scores indicate better overlap with reference summaries, though it is important to note that ROUGE scores do not capture semantic similarity or factual accuracy, only lexical overlap.
Forgetting Analysis

To measure potential forgetting, we created a custom evaluation protocol using 8 general knowledge questions spanning multiple categories: geography ("What is the capital of France?"), basic facts ("How many days are in a week?"), science ("What gas do plants absorb?"), literature ("Who wrote Romeo and Juliet?"), and mathematics ("What is 2 + 2?"). We evaluated both the base FLAN-T5-Small model and our fine-tuned model on these questions, recording their answers and calculating accuracy as the percentage of correct responses. We used flexible matching criteria that allowed for partial matches and number extraction to account for minor formatting variations in model outputs.

4. Results

4.1 Summarization Performance

ROUGE Scores on Test Set

We evaluated the fine-tuned model on the held-out test set of 2,000 reviews. The quantitative results are presented in the table below:

Metric:	Score:
ROUGE-1	17.24%

ROUGE-2	7.06%
ROUGE-L	16.92%
ROUGE-Lsum	16.87%
Average Generation Length	6.62 tokens
Eval Loss	2.9697

The ROUGE-1 and ROUGE-L scores around 17% indicate reasonable unigram overlap between generated and reference summaries. The ROUGE-2 score of 7.06% shows that the model achieved improved bigram overlap, suggesting it learned to preserve some multi-word phrases from the reviews.

The average generation length of 6.62 tokens aligns well with our target of producing concise summaries. This demonstrates that the model successfully learned to generate shorter, more focused outputs compared to the base model, which had a tendency to copy longer portions of the input text.

4.2 Qualitative Examples

To better understand the model's behavior, we examined several examples comparing the base model output, the fine-tuned model output, and the reference summary:

Example 1

- Review: "This chip has a little tomato taste but the jalepeno seems mild. Good alternative to plain tortilla chips. Not overwhelmed by the taste but a good change...."
- Reference Summary: "Good snack, not very hot"
- Base Model Output: "This chip has a little tomato taste but the jalepeno seems mild. Good alternative to plain tortilla chips. Not overwhelmed by the taste but"
- Fine-Tuned Model Output: "A little tomato taste but a little mild"
- Observation: The fine-tuned model produced a shorter, more concise summary but lost some nuance about the chips being a good alternative. The base model simply copied a longer portion of the input text without true summarization.

Example 2

- Review: "I received my first jar as a gift and promptly fell in love with it! When I found it on this website I was so thrilled. The transaction went very smoothly. Not a problem to be found. I think the pr..."
- Reference Summary: "Best Jam Ever"
- Base Model Output: "Love it!"
- Fine-Tuned Model Output: "Yummy!"

- Observation: Both models generated very short summaries. The fine-tuned model is more concise but neither captures the enthusiastic tone of the reference summary effectively.

Example 3

- Review: "I love it! Great smell and exceptional taste! I've tried lot of different sorts/tastes of coffee from cheap to exclusively expensive, but this one remains one of my favorite for years.<br />Great Coffe..."
- Reference Summary: "Melitta French Vanilla Ground Coffee"
- Base Model Output: "Five Stars"
- Fine-Tuned Model Output: "Love this coffee!"
- Observation: The fine-tuned model generated a more descriptive summary compared to the base model's generic response. However, neither model captured the specific product name mentioned in the reference.

Example 4

- Review: "This product does not taste as sweet as the individual packets made by this same company. Unfortunately they donot sell/package individual packets anymore. So I just add Irish Cream creamer to it to..."
- Reference Summary: "Cappuchino powder"
- Base Model Output: "This product does not taste as sweet as the individual packets made by this same company. Unfortunately they do not sell/package individual packets anymore."
- Fine-Tuned Model Output: "Not as sweet as the individual packets made by this company"
- Observation: The fine-tuned model successfully condensed the base model's longer output, removing redundancy while preserving the key complaint about sweetness.

Example 5

- Review: "Our australian shepard puppies love these! and they are the perfect size for training and to put in their treat balls...."
- Reference Summary: "GREAT"
- Base Model Output: "Great for training!"
- Fine-Tuned Model Output: "Great for training!"
- Observation: Both models produced identical outputs, suggesting the base model already generated an appropriate summary for this particular case.

Overall Qualitative Assessment

Across these examples, we observed that the fine-tuned model consistently showed a tendency toward shorter, more concise summaries compared to the base model, which often copied longer portions of the input text. However, the generated summaries do not always match the reference summaries in style or content, suggesting room for improvement with additional training data or more epochs.

4.3 Forgetting Analysis

Accuracy Results

We evaluated both the base model and the fine-tuned model on our set of 8 general knowledge questions. The results are summarized below:

Model:	Accuracy:	Correct Answers:
Base Model	25.00%	2/8
Fine-Tuned Model	25.00%	2/8
Change in Accuracy	0.00%	No change

Both models achieved the same accuracy of 25%, correctly answering 2 out of 8 questions. This indicates that fine-tuning on the summarization task did not degrade the model's general knowledge performance.

Detailed Question-by-Question Results

The table below shows the specific questions, expected answers, and outputs from both models:

Question:	Expected Answer:	Base Model Answer:	Fine-Tuned Model Answer:	Both Correct?:
What is the capital of France?	Paris	london	French capital	No
How many days are in a week?	7	7 days	7 days	Yes
What gas do plants absorb?	carbon dioxide	helium	gas	No
What is the largest planet?	Jupiter	venus	Earth	No
What is H2O?	water	H2O	H2O	No
Who wrote Romeo and Juliet?	Shakespeare	edward wilson	edmund wilson	No
What color is the sky?	blue	blue	blue sky	Yes
What is 2 + 2?	4	2 + 2	2 + 2	No

The two questions that both models answered correctly were:

1. "How many days are in a week?" (both answered "7 days")
2. "What color is the sky?" (base model: "blue", fine-tuned model: "blue sky")

Example of Change in Response
To illustrate how the fine-tuning affected responses, consider the question "What is the capital of France?":

- Base Model Answer: "london" (incorrect, confusing France with England)
- Fine-Tuned Model Answer: "French capital" (more relevant phrasing but still not the exact answer "Paris")

While the fine-tuned model's response is arguably more relevant than the base model's incorrect answer, neither provided the correct factual answer.

Conclusion on Forgetting
We found no evidence of catastrophic forgetting in our experiment. The fine-tuned model maintained the same 25% accuracy on general knowledge questions as the base model, indicating that specialization on the summarization task did not interfere with its ability to respond to factual questions.

However, it is important to interpret these results cautiously. The base model's accuracy of 25% suggests that FLAN-T5-Small may not be well-suited for factual question answering without task-specific fine-tuning. The low baseline performance makes it difficult to detect subtle degradation. Additionally, our test set of only 8 questions is relatively small and may not comprehensively capture all aspects of general knowledge.

Possible explanations for the lack of observed forgetting include: (1) summarization and question answering tasks may use different parts of the model's knowledge and computational pathways, reducing interference; (2) the limited scale of fine-tuning (only 2 epochs on 16,000 samples) may not have been sufficient to significantly alter the model's general capabilities; and (3) the model's general knowledge performance was already limited, leaving little room for measurable degradation.

5. Discussion and Limitations

5.1 Key Findings

Our project yielded three main findings:

1. Summarization Improvement: The fine-tuned model successfully learned to generate shorter, more concise summaries compared to the base model, demonstrating effective task-specific adaptation. While the ROUGE scores are modest, they represent meaningful improvement over a base model that primarily copied input text.

2. Improved ROUGE Scores: We achieved ROUGE scores around 17% for ROUGE-1 and 7% for ROUGE-2, indicating improved performance in capturing unigram and bigram overlap with reference summaries. The average generation length of 6.62 tokens shows the model learned to produce appropriately concise outputs.

3. No Catastrophic Forgetting: The model maintained its general knowledge performance after fine-tuning, with both base and fine-tuned models achieving 25% accuracy on our test questions. This suggests that task-specific fine-tuning on summarization did not significantly degrade other capabilities, at least not in a measurable way given our evaluation setup.

5.2 Limitations

Several limitations should be considered when interpreting our results:

1. Small Training Set: Using only 20,000 samples may have limited the model's ability to fully learn the summarization task. The full dataset contains over 500,000 reviews, and training on a larger subset might yield better performance.

2. Limited Epochs: Training for only 2 epochs may not have been sufficient for model convergence. Additional training could potentially improve summarization quality, though it might also increase the risk of forgetting.

3. Small Forgetting Test Set: Our evaluation of forgetting relied on only 8 questions, which is too small to draw definitive conclusions about the model's retention of general knowledge. A more comprehensive benchmark with hundreds of questions across diverse topics would provide more reliable insights.

4. Model Size: FLAN-T5-Small has only 60 million parameters, which is relatively small by modern standards. This limited capacity may constrain both summarization quality and general knowledge retention. Larger models might show different patterns of learning and forgetting.

5. Evaluation Metrics: While ROUGE scores are standard for summarization, they may not fully capture the quality of very short summaries. ROUGE focuses on lexical overlap rather than semantic similarity or factual accuracy, potentially missing important aspects of summary quality.

5.3 Future Work

Based on our findings and limitations, we propose several directions for future research:

1. Increase Training Data: Train on the full dataset or a larger sample (50,000+ reviews) to provide the model with more diverse examples and potentially improve summarization quality.

2. Extended Training: Train for more epochs (3 to 5) with learning rate scheduling to allow for better convergence. Monitor validation performance carefully to avoid overfitting.

3. Hyperparameter Tuning: Conduct systematic experiments with different learning rates, batch sizes, and generation parameters (such as number of beams and length penalties) to optimize performance.

4. Comprehensive Forgetting Analysis: Use a larger, more diverse set of general knowledge questions or standardized benchmarks (such as MMLU or TriviaQA subsets) to more rigorously evaluate forgetting behavior across multiple knowledge domains.

5. Larger Model Experimentation: Experiment with FLAN-T5-Base or FLAN-T5-Large to investigate whether increased model capacity affects both summarization quality and forgetting behavior. Larger models may have sufficient capacity to learn new tasks without displacing existing knowledge.

6. References

1. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., & Wei, J. (2022). Scaling Instruction-Finetuned Language Models. arXiv preprint https://arxiv.org/abs/2210.11416.

2. Hugging Face. (2023). FLAN-T5-Small Model Card. Retrieved from https://huggingface.co/google/flan-t5-small

3. Hugging Face. (2023). Transformers: State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX. Retrieved from https://github.com/huggingface/transformers

4. Google Colab. (2023). Colaboratory. Retrieved from https://colab.research.google.com/

5. Jhan21. (2023). Amazon Fine Food Reviews Dataset. Retrieved from https://huggingface.co/datasets/jhan21/amazon-food-reviews-dataset

6. Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out (pp. 74-81). Association for Computational Linguistics. https://aclanthology.org/W04-1013/

7. McAuley, J., & Leskovec, J. (2013). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. (pp. 897-908). ACM. https://dl.acm.org/doi/10.1145/2488388.2488466