Project 1 Report - Fine-Tuning LLMs Trustworthy ML (ECE 696B): Spring 2025

Natnael Daba

February 26, 2025

Abstract

This project aimed to fine-tune Large Language Models (LLMs) for text classification and summarization tasks. We fine-tuned GPT-2 Small and GPT-Neo models on the IMDB Large Movie Review and AG's News Topic Classification datasets for sentiment and topic classification, respectively. Additionally, we fine-tuned the Flan-T5-Base model on the SAMSum dataset for dialogue summarization. Our results indicate that fine-tuning GPT-2 Small significantly improved performance in both classification tasks, while GPT-Neo showed moderate gains. The Flan-T5-Base model exhibited modest enhancements in summarization quality post fine-tuning. These findings suggest that model architecture and pre-training influence fine-tuning effectiveness. The code and resources for reproducing these experiments are available at: `https://github.com/nate-daba/llm-sft`.

1 Text Classification

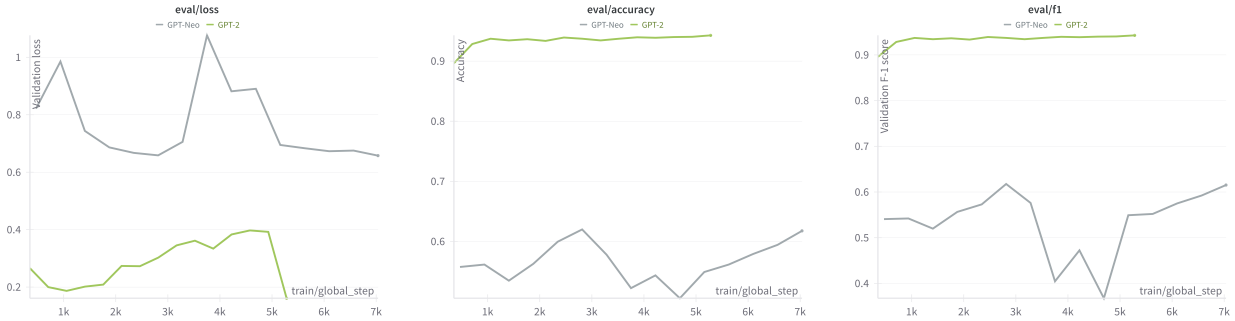
1.1 Sentiment Classification

Given a text input (E.g. tweets, movie reviews, etc), the task of sentiment classification is to classify the input text into one of the binary outputs. I.e. positive or negative sentiment. The dataset used for this task is the **IMDB Large Movie Review Dataset** [11] from [11]. The IMDB dataset contains 50000 movie reviews for binary sentiment classification. The dataset is split into training, validation, and test set resulting in 22500 training, 2500 validation, and 25000 test samples.

Two pre-trained language models are fine-tuned on the IMDB dataset to classify movie reviews as positive or negative: (1) GPT-2 Small (124M parameters) [12], and (2) GPT-Neo (125M parameters) [1]. Both models are fine-tuned for 15 epochs with a learning rate of 2×10^{-5} and weight decay of 0.01. Table 1 shows the evaluation of these models on the test set before and after finetuning. Figure 1 shows the training and validation accuracy and F1-score curves of the training process. Figures 2 and 3 show a sample movie review from the IMDB Large Movie Review Dataset and the corresponding model predictions with ground truth, respectively.

Table 1: Sentiment Classification Results

Model	Accuracy	F1 Score
Baseline GPT-2	0.5001	0.3344
Baseline GPT-Neo	0.4788	0.4500
Fine-tuned GPT-2	**0.9446**	**0.9446**
Fine-tuned GPT-Neo	0.6176	0.6152



(a) Validation loss (b) Validation accuracy (c) Validation F-1 score

Figure 1: Validation loss, accuracy and F-1 score of the fine-tuning process of GPT-2 and GPT-Neo for sentiment classification on the IMDB Large Movie Review Dataset [11]

Sample Movie Review

Isaac Florentine has made some of the best western Martial Arts action movies ever produced. In particular, *US Seals 2*, *Cold Harvest*, *Special Forces*, and *Undisputed 2* are all action classics. You can tell Isaac has a real passion for the genre, and his films are always eventful, creative, and sharp affairs, with some of the best fight sequences an action fan could hope for. In particular, he has found a muse with Scott Adkins, as talented an actor and action performer as you could hope for. This is borne out with *Special Forces* and *Undisputed 2*, but unfortunately, *The Shepherd* just doesn't live up to their abilities.

There is no doubt that JCVD looks better here fight-wise than he has done in years, especially in the fight he has (for pretty much no reason) in a prison cell, and in the final showdown with Scott. But look in his eyes. JCVD seems to be dead inside. There's nothing in his eyes at all. It's like he just doesn't care about anything throughout the whole film. And this is the leading man.

There are other dodgy aspects to the film, script-wise and visually, but the main problem is that you are utterly unable to empathize with the hero of the film. A genuine shame, as I know we all wanted this film to be as special as it genuinely could have been. There are some good bits, mostly the action scenes themselves. This film had a terrific director and action choreographer, and an awesome opponent for JCVD to face down. This could have been the one to bring the veteran action star back up to scratch in the balls-out action movie stakes.

Sincerely a shame that this didn't happen.

Figure 2: Sample movie review from the IMDB Large Movie Review Dataset.

Model Predictions

Baseline GPT-2 Prediction: *POSITIVE* (Score: 0.99998) ✗
Fine-tuned GPT-2 Prediction: *NEGATIVE* (Score: 0.99998) ✓
Baseline GPT-Neo Prediction: *POSITIVE* (Score: 0.69614) ✗
Fine-tuned GPT-Neo Prediction: *POSITIVE* (Score: 0.59411) ✗
Ground Truth: *NEGATIVE*

Figure 3: Model predictions and ground truth of sample movie review shown in Figure 2

Discussion Fine-tuning GPT-2 Small on the IMDB dataset resulted in a significant performance improvement, with accuracy and F1 scores increasing from approximately 0.50 and 0.33 to 0.94, respectively. In contrast, GPT-Neo's performance improved modestly, achieving an accuracy of 0.62 and an F1 score of 0.62 after fine-tuning. The superior performance of GPT-2 Small may be attributed to its training objectives and architecture, which are better suited for sentiment classification tasks compared to GPT-Neo. This observation aligns with findings that GPT-2 delivers higher overall accuracy in emotional detection tasks [14]. The disparity in performance suggests that model selection should consider the specific characteristics and requirements of the task at hand.

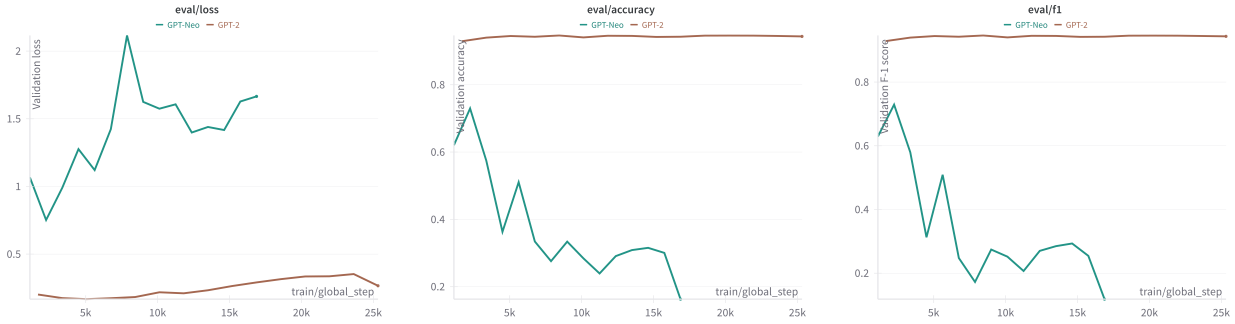
1.2 Topic Classification

Topic classification involves assigning documents to predefined categories, such as sports, business, politics, etc. For this task, we utilized the **AG's News Topic Classification Dataset** [15], a subset of AG's corpus comprising titles and descriptions from articles across four primary classes: World, Sports, Business, and Sci/Tech. The dataset includes 30,000 training samples and 1,900 test samples per class.

Building upon the methodology outlined in Section 1.1, we fine-tuned GPT-2 Small (124M parameters) and GPT-Neo (125M parameters) models for this classification task. GPT-2 was trained for 15 epochs, while GPT-Neo underwent 50 epochs of training; other hyperparameters remained consistent with those previously described. Table 2 presents the evaluation metrics—accuracy and F1 score—of these models on the test set, both before and after fine-tuning. Figure 4 illustrates the training and validation accuracy and F1-score trajectories throughout the training process. Figures 5 and 6 present a sample news article from the AG's News Topic Classification Dataset and the corresponding model predictions alongside the ground truth, respectively.

Table 2: Topic Classification Results on AG News Dataset

Model	Accuracy	F1 Score
Baseline GPT-2	0.2501	0.1003
Baseline GPT-Neo	0.2616	0.1936
Fine-tuned GPT-2	**0.9442**	**0.9442**
Fine-tuned GPT-Neo	0.7247	0.7250



(a) Validation Loss (b) Validation Accuracy (c) Validation F1 Score

Figure 4: Validation loss, accuracy, and F1-score during fine-tuning of GPT-2 and GPT-Neo models on the AG News Topic Classification Dataset.

Sample News Article

Losing the War on Terrorism

"Sven Jaschan, self-confessed author of the Netsky and Sasser viruses, is responsible for 70 percent of virus infections in 2004, according to a six-month virus roundup published Wednesday by antivirus company Sophos."

"The 18-year-old Jaschan was taken into custody in Germany in May by police who said he had admitted programming both the Netsky and Sasser worms, something experts at Microsoft confirmed. (A Microsoft antivirus reward program led to the teenager's arrest.) During the five months preceding Jaschan's capture, there were at least 25 variants of Netsky and one of the port-scanning network worm Sasser."

"Graham Cluley, senior technology consultant at Sophos, said it was staggering..."

Figure 5: Sample news article from SAMSum dataset [6].

Model Predictions

Baseline GPT-2 Prediction: *Business* (Score: 0.4514) ✗
Fine-tuned GPT-2 Prediction: *Sci/Tech* (Score: 0.99999) ✓
Baseline GPT-Neo Prediction: *World* (Score: 0.6823) ✗
Fine-tuned GPT-Neo Prediction: *Business* (Score: 0.3984) ✗
Ground Truth: *Sci/Tech*

Figure 6: Model predictions and ground truth of the sample article shown in Figure 5.

Fine-tuning significantly enhanced both models' performance on the AG News dataset. GPT-2's accuracy improved from 25.01% to 94.42%, while GPT-Neo's accuracy increased from 26.16% to 72.47%. The substantial improvement in GPT-2's performance underscores its capacity to adapt effectively to topic classification tasks. Notably, despite GPT-Neo's comparable parameter count, its post fine-tuning performance was inferior to that of GPT-2. This discrepancy may stem from architectural differences between the models or variations in their pre-training corpora, which could influence their ability to generalize across different tasks.

2 Text Summarization

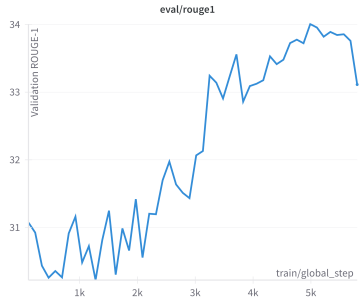
Given a lengthy text input (e.g., a paragraph or larger body of text), the task of text summarization is to generate a shorter summary of the input. The dataset used for this task is the **SAMSum dataset** [6], obtained from [3]. The SAMSum dataset contains approximately 16,000 messenger-like conversations with summaries. The dataset is split into training, validation, and test sets, comprising 14,731 training, 818 validation, and 819 test samples.

The model employed for this task is Flan-T5-Base, introduced by Google. It is a 250M parameter model that is a fine-tuned version of T5, instruction-tuned on a variety of tasks to enhance zero-shot and few-shot learning capabilities [2].

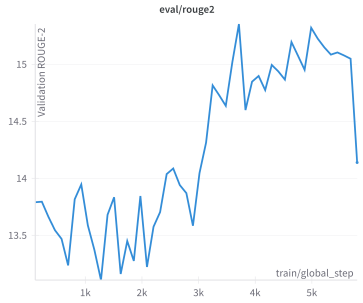
Flan-T5-Base was fine-tuned for 50 epochs with a learning rate of 2×10^{-5} on the SAMSum dataset. The ROUGE metric [10] was used to evaluate the quality of the generated summaries. ROUGE-1 measures the overlap of unigrams (single words) between the generated and reference summaries. ROUGE-2 measures the overlap of bigrams (two consecutive words). ROUGE-L focuses on the longest common subsequence, capturing sentence-level structure similarity. ROUGE-Lsum is similar to ROUGE-L but operates at the summary level, considering sentence boundaries. These ROUGE scores provide insights into how well the model's summaries align with human-written references in terms of content and structure. Table 3 below shows the performance of the Flan-T5-Base before and after fine-tuning. Figure 7 illustrates the performance of the Flan-T5-Base on the validation set during the fine-tuning process. Figures 8 and 9 present a sample dialogue from the SAMSum dataset and the corresponding model-generated summaries alongside the reference summary, respectively.

Table 3: Summarization Performance of Fine-tuned T5-Flan-Base on SAMSum Dataset

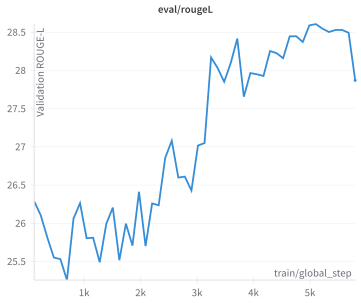
Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Baseline Flan-T5-Base	0.3158	0.1287	0.2627	0.2881
Fine-tuned Flan-T5-Base	**0.3522**	**0.1539**	**0.2948**	**0.3223**



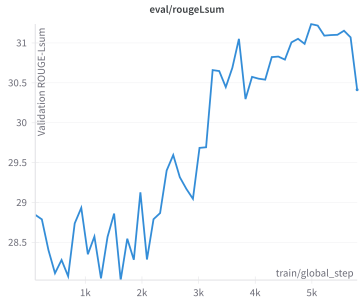
(a) ROUGE-1 Score



(b) ROUGE-2 Score



(c) ROUGE-L Score



(d) ROUGE-Lsum Score

Figure 7: Validation ROUGE scores across min-batch training steps.

Sample Dialogue

Richie: Pogba
Clay: Pogboom
Richie: What a strike, yoh!
Clay: Was off the seat the moment he chopped the ball back to his right foot.
Richie: Me too, dude.
Clay: Hope his form lasts.
Richie: This season he's more mature.
Clay: Yeah, Jose has his trust in him.
Richie: Everyone does.
Clay: Yeah, he really deserved to score after his first 60 minutes.
Richie: Reward.
Clay: Yeah, man.
Richie: Cool then.
Clay: Cool.

Figure 8: Sample Dialogue between Richie and Clay from the SAMSum dataset [3].

Model Summaries

- **Flan-T5-Base Summary:** Pogba scored a goal after his first 60 minutes. Richie and Clay hope his form lasts this season and he's more mature.
- **Flan-T5-fine-tuned Summary:** Pogba scored a strike after his first 60 minutes. Richie and Clay hope his form lasts this season and he's more mature.
- **Reference Summary:** Richie and Clay saw a very good football game, with one football player chopping the ball back to his foot, which was particularly exciting. Jose has trust in that player.

Figure 9: Model summaries and reference summaries of the sample dialogue shown in Figure 8.

Discussion After fine-tuning the Flan-T5-Base model on the SAMSum dataset for 50 epochs, the ROUGE scores showed modest improvements: ROUGE-1 increased from 0.3158 to 0.3522, ROUGE-2 from 0.1287 to 0.1539, ROUGE-L from 0.2627 to 0.2948, and ROUGE-Lsum from 0.2881 to 0.3223. This indicates that while the model's performance enhanced slightly, the gains were not substantial.

Several factors may contribute to this outcome. Firstly, the SAMSum dataset, comprising approximately 16,000 dialogue-summary pairs, might be insufficient for significant performance leaps, especially for a model with 250 million parameters like Flan-T5-Base. Secondly, the model's pre-training on a diverse range of tasks could mean it already possesses a strong capability for summarization, leaving limited room for improvement through fine-tuning on a single dataset. Additionally, the quality and consistency of the SAMSum dataset's summaries play a crucial role; any noise or variability in the data can hinder the fine-tuning process.

Regarding the similarity between the summaries generated by the fine-tuned and base models, this suggests that the pre-trained model was already proficient in handling dialogue summarization tasks. The fine-tuning process, therefore, resulted in only marginal refinements to its outputs.

In summary, while fine-tuning the Flan-T5-Base model on the SAMSum dataset yields slight improvements in summarization performance, the extent of enhancement is constrained by factors such as dataset size, pre-existing model capabilities, and data quality.

3 Resources

1. Used [9] and [7] for implementing the training and evaluation routines of the text classification task.

2. Used [8] and [13] for implementing the training and evaluation routines for the text summarization task.

3. Downloaded the SAMSum dataset [6] from [3].

4. Downloaded AG News topic classification dataset [15] from [5].

5. GPT-Neo 125M model card [1]

6. Used [4] for launching distributed training accross multiple GPUs.

References

[1] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. https://doi.org/10.5281/zenodo.5297715, March 2021. If you use this software, please cite it using these metadata. doi:10.5281/zenodo.5297715.

[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL: https://arxiv.org/abs/2210.11416, doi:10.48550/ARXIV.2210.11416.

[3] Hugging Face. knkarthick/samsum dataset. https://huggingface.co/datasets/knkarthick/samsum, 2025. Accessed: 2025-02-20.

[4] Hugging Face. Transformers: Distributed training and mixed precision. https://github.com/huggingface/transformers/tree/main/examples/pytorch#distributed-training-and-mixed-precision, 2025. Accessed: 2025-02-20.

[5] fancyzhx. Ag news dataset, 2025. Accessed: 2025-02-20. URL: https://huggingface.co/datasets/fancyzhx/ag_news.

[6] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. URL: https://aclanthology.org/D19-5409/, doi:10.18653/v1/D19-5409.

[7] Hugging Face. *Evaluate: TextClassificationEvaluator*, 2025. Accessed: 2025-02-19. URL: https://huggingface.co/docs/evaluate/v0.4.0/en/package_reference/evaluator_classes#evaluate.TextClassificationEvaluator.

[8] Hugging Face. *Summarization*, 2025. Accessed: 2025-02-20. URL: https://huggingface.co/docs/transformers/en/tasks/summarization.

[9] Hugging Face. *Text Classification*, 2025. Accessed: 2025-02-19. URL: https://huggingface.co/docs/transformers/v4.49.0/en/tasks/sequence_classification.

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL: https://aclanthology.org/W04-1013/.

[11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/P11-1015.

[12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[13] Phil Schmid. Fine-tuning flan-t5 for summarization on samsum dataset, 2025. Accessed: 2025-02-20. URL: https://github.com/philschmid/deep-learning-pytorch-huggingface/blob/main/training/flan-t5-samsum-summarization.ipynb.

[14] Armand Stricker and Patrick Paroubek. A unified approach to emotion detection and task-oriented dialogue modeling. *arXiv preprint arXiv:2401.13789*, 2024. URL: https://arxiv.org/abs/2401.13789.

[15] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.