FINE-TUNING NEWS SUMMARIZATION MODEL WITH HUMAN PREFERENCES

Rimika Dhara, Nikhitha Gollamudi, Akhil Mavallapalli & Xifan Li
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
{dhara015, golla063, maval003, li003646} @umn.edu

1 INTRODUCTION AND PROBLEM STATEMENT

Traditional news summarization models lack optimization with respect to human preference. Currently, transformer-based summarization models are facing issues including inaccurate information and verbose summaries that are not aligned with how humans would normally prefer their news to be summarized. Given these issues, the problem we aim to tackle is improving how accurately, precisely, and concisely we can tune our model to align more closely with human preferences. We also want to understand how much difference this tuning would make to the established pre-trained models.

Our approach to this problem is to implement reinforcement learning (RL) by training a reward model on human-labeled preferences. More specifically, we employ two RL algorithms, Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO), in hopes of increase the performance of the summarization model. This will help directly tackle current issues by optimizing the model to create summaries that humans prefer. Building on the work of Ziegler et al. (2019) and Stiennon et al. (2020), we apply RL from human feedback (RLHF) specifically to the news domain, with particular attention to factual accuracy and conciseness.

2 SIGNIFICANCE OF THE PROBLEM

Automated summarization plays a vital role in modern information consumption, enabling quick understanding of large texts by highlighting key ideas. This is especially impactful in news summarization, where journalism can benefit from faster, more digestible content. However, current systems face major issues—most notably, factual inaccuracy and verbosity—which limit their effectiveness. Inaccurate or overly long summaries can lead to misinformation, reduce readability, and erode trust in AI-generated content, ultimately defeating the purpose of aiding information processing. Beyond journalism, accurate and concise summarization has critical applications in domains like healthcare, legal analysis, and finance, where mistakes can be costly and time is a premium resource. Although transformer-based models have advanced the field, existing limitations show the need for better alignment with human preferences. Incorporating human feedback into training—through methods like RL—can improve both reliability and utility. This work has the potential to reshape how AI contributes to information consumption, setting new standards for trustworthy, human-aligned summarization systems.

3 PREVIOUS AND RELATED WORK

The field of language model fine-tuning from human preferences has seen significant developments in recent years. Notably, Ziegler et al. (2019) demonstrated that RL from human feedback (RLHF) can successfully fine-tune language models for tasks including summarization. Their work on the TL;DR and CNN/Daily Mail datasets showed that models trained with 60,000 human comparisons outperformed supervised baselines according to human evaluators, although they observed that the resulting models often acted as ”smart copiers” that selected relevant sentences from the input. Building on this foundation, Stiennon et al. (2020) further refined the approach for summarization tasks, showing that human feedback can be more effective than traditional supervised learning methods. Their work emphasized the importance of high-quality human feedback and introduced iterative

data collection procedures to improve model performance. In parallel, recent work by Böhm et al. (2019) also explored using human evaluations to learn reward functions for summarization, providing detailed investigations of learned policies on the CNN/Daily Mail dataset. Meanwhile, advances in large language models like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) have established new state-of-the-art results on summarization benchmarks using traditional supervised approaches. More recently, researchers have explored various techniques to improve RLHF, including Constitutional AI (Bai et al., 2022), which combines rule-based and preference-based learning, and Direct Preference Optimization (Rafailov et al., 2023), which reformulates the preference learning problem to avoid the instabilities associated with traditional RL approaches. These advancements provide significant evidence that this is a promising direction.

4 Goals and Progress So Far

The overall workflow pipeline for our project is illustrated in Figure 1. We begin with two key datasets: the CNN/DailyMail summarization dataset, which is used to train and benchmark the base summarization model (T5 encoder-decoder model), and the OpenAI TL;DR preference dataset, which is used to train a reward model from human-labeled comparisons that is used to understand human preferences. The base summarization models are evaluated to set benchmarks that will later be used to evaluate our reward models. The reward model will be used to guide fine-tuning via reinforcement learning methods—specifically, PPO and DPO—producing a model optimized for accuracy, conciseness, and human preference.

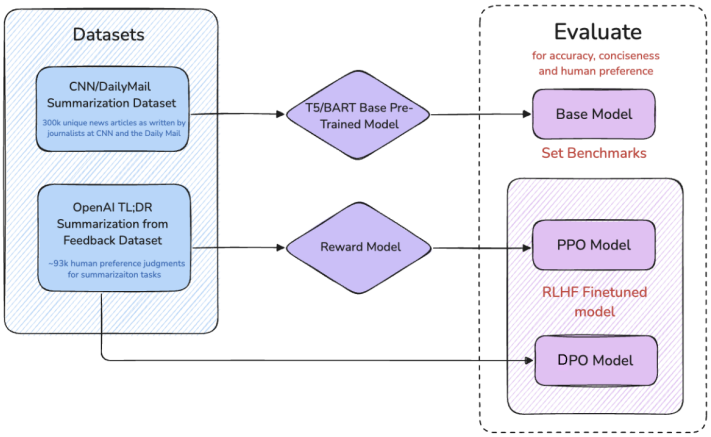

Figure 1: Project Workflow Overview

On setting up the benchmark statistics side, we implemented two levels of benchmarking using the CNN/DailyMail dataset and two T5 configurations: `t5-base` trained on 2,000 samples, and `t5-large` trained on 1,000 samples due to compute constraints. Both setups followed the same training pipeline, evaluation metrics, and preprocessing logic, differing only in model size and dataset scale. The smaller-scale setup was used for rapid prototyping and early comparison, while the `t5-large` model was explored to assess potential improvements from model scaling.

While we initially planned to include `facebook/bart-large-cnn` in our benchmarks, we encountered persistent training instabilities and evaluation inconsistencies across runs. Based on our observations, the issues likely stemmed from a combination of factors: a high learning rate, insufficient input length due to a low `max_input_length`, and hardware limitations when increasing dataset or batch sizes. These constraints led the model to overfit on shorter reference summaries and crash under larger configurations. As a result, we decided to focus our benchmarking efforts on the more stable and interpretable T5-based models.

Our pipeline includes:

- Loading and trimming the CNN/DailyMail dataset to target subset sizes

- Applying T5-style summarization prompts (e.g., `summarize: <article>`)
- Preprocessing using HuggingFace tokenizers with max-length padding and truncation
- Training using `Seq2SeqTrainer` with ROUGE and METEOR evaluations
- Saving model checkpoints and generating sample summaries for both qualitative and quantitative analysis

The results from these experiments, shown in Table 1, serve as our baseline for comparison against the RLHF fine-tuned models.

Model	Dataset Size	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
`t5-base`	2K	0.3302	0.1345	0.2523	0.1495
`t5-large`	1K	0.2468	0.0890	0.1998	0.1495

Table 1: Baseline performance comparison across T5 configurations.

As shown in Table 1, the `t5-base` model trained on 2,000 samples outperformed the `t5-large` model trained on 1,000 samples across all ROUGE metrics, despite the latter's larger model size. This suggests that dataset size had a stronger impact than model capacity under our current setup. Interestingly, both models produced identical METEOR scores, indicating similar word-level alignment despite structural differences in the summaries. These benchmarks serve as a strong reference for evaluating our RLHF fine-tuned models. We plan to revisit BART once the training stability issues are resolved.

On the human preferences side, we focused on preparing and modeling the human feedback data to build a reward model that would eventually guide reinforcement learning for summarization through PPO and DPO.

We used the OpenAI's `Summarize from Feedback` dataset, specifically the `comparisons` split, which contains approximately 93k human preference judgments. This dataset is well-suited for training a pairwise reward model due to its format: each example includes a Reddit post, two candidate summaries (`summary_0` and `summary_1`), and a `choice` field indicating which summary was preferred by a human annotator. Since high-quality human preference datasets are expensive and thus relatively scarce, especially in the context of preference tuning, we found this dataset to be the most relevant to our work. While it doesn't explicitly capture the nuances between general and news-specific summaries—each of which emphasizes different kinds of details—it remains the most practical and well-established option for our objectives for this project.

We completed the following steps:

- Verified dataset structure and number of examples
- Analyzed the preference distribution to ensure balance between choices
- Built a custom PyTorch dataset class that returns chosen and rejected summary pairs
- Designed a reward model using `distilbert-base-uncased` and a regression head
- Trained the reward model for 5 epochs using margin ranking loss
- Tracked and visualized training loss over epochs
- Evaluated on the validation set and reported average pairwise validation loss
- Saved the model as `reward_model.pt` for use in PPO

Initial training showed expected behavior with high loss in early epochs (1.02), decreasing as the model learned to rank preferred summaries higher. This behavior is normal in early training and indicates that the model has room to learn. Our trained reward model demonstrated consistent improvement over the first three epochs, as shown in Figure 2. The validation loss reached its lowest point at epoch 3 (0.5053), after which it began to plateau and increase slightly, indicating potential overfitting. Based on this trend, the model checkpoint from epoch 3 was selected as the final version for downstream use. The average pairwise validation loss across all epochs was approximately 0.5324. As next steps, we plan to continue training the reward model for up to 10 epochs to observe whether the overfitting trend persists beyond the initial five. 10 epochs seems reasonable for

Figure 2: Training and validation loss per epoch for the reward model

preference-based reward models trained on datasets of this size, providing enough runway for learning while allowing early stopping if validation loss begins to rise. We will also explore strategies to reduce overfitting, such as adding dropout, using early stopping more aggressively, or applying stronger regularization. While KL divergence is not used in reward model training directly, it will play an important role in the subsequent reinforcement learning phase as a regularization technique during policy optimization, where we expect to see further improvements.

5 Updated Plan and Milestones

We have successfully completed the tasks planned for Week 1 and Week 2. Below is an updated milestone list, with adjustments reflecting current progress and challenges:

- **Week 1 (4/6 – 4/12): Dataset collection and preprocessing.**
  - ✓Preprocessed and tokenized the CNN/DailyMail dataset.
  - ✓Completed literature review on transformers and RL for summarization.
- **Week 2 (4/13 – 4/19): Reward model training.**
  - ✓Trained a reward model using the OpenAI `Summarize from Feedback` dataset.
  - ✓Validated model and saved best-performing checkpoint.
  - ✓Built PyTorch dataset and implemented margin ranking loss.
- **Week 3 (4/20 – 4/26): Fine-tuning with RLHF (PPO/DPO).**
  - Begin PPO fine-tuning of the T5 model with the reward model.
  - Explore DPO as an alternative method.
  - Apply KL divergence for regularization during PPO.
- **Week 4 (4/27 – 5/3): Model evaluation.**
  - Evaluate for accuracy, conciseness, and preference alignment.
  - Use ROUGE, METEOR, and BERTScore.
  - Perform qualitative comparison of summaries pre/post fine-tuning.
- **Week 5 (5/4 – 5/10): Results analysis and report.**
  - Analyze quantitative and qualitative results.
  - Finalize report with findings and visualizations.

Challenges: Setting up and training the reward model required significant time and compute resources. Configuring and deploying on MSI also introduced additional delays due to environment/setup dependencies and some compatibility constraints. Furthermore, instability in BART

training due to input token limitations and hardware crashes led us to refocus our benchmarking efforts solely on T5. We are actively working on optimizing training performance to ensure smoother PPO fine-tuning in the coming weeks as well as exploring BART a bit further.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.