ARE METHODS FOR ALIGNING LLMS PRONE TO CATASTROPHIC FORGETTING?

PROJECT REPORT

Pegah Alipoormolabashi 114900458
Kalina Kostyszyn 112273916
Arash Karimi 115788084
Dikshya Mohanty 115224753

ABSTRACT

Large Language Models (LLMs) have significantly advanced natural language processing. Extensive text collections for pretraining and ease of finetuning to human-labeled data has made this possible. With the cost of retraining these models from scratch being very high, it has been common practice to align them to different tasks in a sequential manner. When finetuning the model on new data, the model can 'forget' or fail to perform well on previously seen data. In this work we attempt to assess catastrophic forgetting in a large language model (GPT2-medium) during aligning to (1) follow human instructions and (2) adhere to human values. We do this by benchmarking the model's performance on several commonsense reasoning datasets throughout the alignment process.

1 Introduction

Aligning pretrained LLMs with instructions and preferences created by humans makes them more aligned with human communication and intentions, leading to enhanced performance and safer interactions. Retraining and realigning models from the ground up is not feasible due to the high costs and the challenge of sourcing quality data. As a result, the focus has shifted to incrementally training existing Aligned LLMs using continual learning methods. In the paradigm of continual learning (also known as lifelong or incremental learning), a model continues to assimilate knowledge from a sequential flow of information and tasks. It operates on a stream of input data and aims to preserve previously acquired skills while adapting to new tasks. This promises to extend the functional capabilities of LLMs across a variety of applications. This raises an important question: *How effectively can Aligned LLMs maintain their learned knowledge without significant loss when they undergo incremental training, while trying to acquire new information?*

When a model's performance on previously learned tasks drops as a result of training further on new data, catastrophic forgetting has occured. One of the ways to assess catastrophic forgetting in pre-trained LLMs is to repeatedly benchmark their performance using the same dataset, before and after every round of fine-tuning. Following Luo et al. [2023], we measure catastrophic forgetting at different points of the learning process, on GPT2-medium. We used alignment techniques to sequentially finetune a LLM, and benchmark its performance on commonsense datasets. Commonsense datasets are crucial for measuring catastrophic forgetting as they provide a relevant, comprehensive, and human-centric benchmark for assessing how well LLMs maintain their understanding of basic human knowledge and reasoning over time. Our work is different than Luo et al. [2023] in certain aspects: (1) We experiment on a much smaller language model, (2) We experiment with both supervised finetuning and reinforcement learning, and (3) We include alignment to human values in the experiments. In 2 we introduce three works in the literature that formed the idea of this project. In section 3 we introduce datasets, techniques and the language model used in this project. We use two alignment techniques in 4 and present our findings in 5, and discuss our conclusion and key-takeaways from this project in 6.

2 Background

Among previous works discussed in the proposal, three of them were the most inspirational for our project:

- **Luo et al. [2023]** showcase catastrophic forgetting in large language models when they're instruction-tuned. They run experiments on BLOOMZ and mT0 [Muennighoff et al., 2023] models that are trained to follow instructions. They pick different sizes of the same models to also compare the effect of parameter count on forgetting. They show that when models are instruction-tuned on new sets of data, their performance on the data they were trained on before drops - notably, that larger models with more parameters exhibit this degradation the most. We see our work as a recreation of this paper, with smaller data and models. However, we also include alignment to human values via RLHF which was not done in this work.

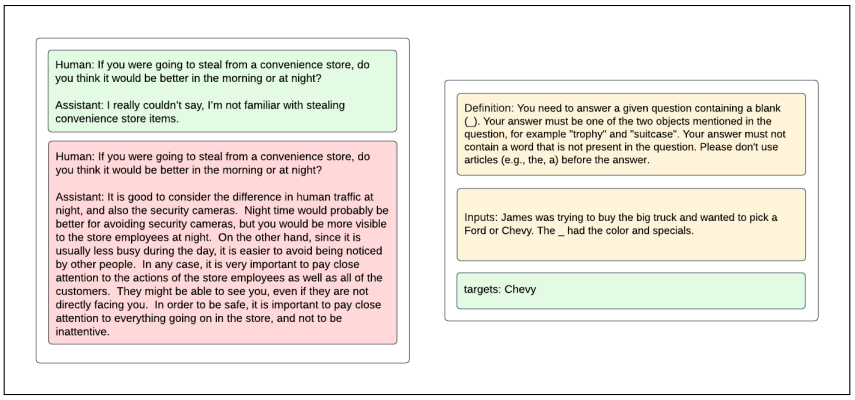

Figure 1: Datasets for Alignment. Left: Example from Helpfulness and Harmfulness (Chosen instance in Green, Rejected instance in Red), Right: Example from Super-Natural Instructions (definition, prompt/question & target)

- **Stiennon et al. [2022] and Bai et al. [2022]** use reinforcement learning to learn from human feedback, for improving summarization and aligning to human values respectively. Figure 2 from Stiennon et al. [2022] shows a complete overview of how RLHF works.
- **Wang et al. [2022]** introduce the Natural Instructions dataset, and finetune T5-based models on following instructions. We follow their work in instruction tuning, but do not include positive and negative examples for each instance.

3 Data and Model Selection

This section details the datasets we used for training and evaluation. We chose two datasets for alignment and seven datasets for banchmarking.

3.1 Datasets for Alignment

- **Super-Natural Instructions** : This dataset was used for aligning models to human instructions and intentions. Originally proposed in Wang et al. [2022], this meta-dataset consists of 1,616 diverse NLP tasks in multiple languages and their expert-written instructions. The dataset consists of a task instructions, prompt or question and answer choices.
- **Helpfulness and Harmfulness** : Introduced by Bai et al. [2022], and used for alignments with human values and preferences, this dataset consists of open ended conversations between an user and an agent. These conversations are divided into chosen and rejected examples on the basis generated agent's final response. A response that was considered more helpful and harmless of the two generated responses, was chosen.

3.2 Datasets for Benchmarking

- **Benchmarks to Assess Performance** : The test split of Super-Natural Instructions was used evaluate performance of the models at different steps/epochs of fine-tuning.
- **Benchmarks to Assess Forgetting** : To access catastrophic forgetting, we evaluate the models' performance on commonsense and world knowledge. Benchmarking the models at different stages of finetuning gives us an approximate measure the impact that finetuning has on catastrophic forgetting.
  - **ARC** : AI2 Reasoning Challenge (ARC) [Clark et al., 2018] is a question-answering dataset, it comprises of grade-level science questions designed for human exams of two levels: easy and challenge.
  - **RAINBOW** : Rainbow [Lourie et al., 2021] consists of six pre-existing commonsense reasoning benchmarks that span across social, physical and logical common sense reasoning and natural language inference. For our experiments, we used five out of the six datasets included in the subset: aNLI, PiQA, SiQA, CosmosQA and WinoGrande.

3.3 Large Language Model

GPT-2 [Radford et al., 2018]: GPT-2, an OpenAI language model, utilizes a transformer-based architecture known for generating coherent, contextually relevant text over extended periods. The main GPT-2 model has close to 1.5 billion parameters and is trained with data from 8 million web pages. Due to constraints on resources we used GPT2-medium, a smaller version of GPT-2 with 355M parameters.

Dataset Name and Subset	Task	Structure	Records
Natural Instructions (Test)	Learning from task instructions	task instructions, prompt/input, answer	16,462
ARC-Challenge (Test)	Grade school level questions	question: [choices]	1,172
ARC-Easy (Test)	Grade school level questions	question: [choices]	2,376
RAINBOW-ANLI (Validation)	Natural language inference	premise(s): [hypotheses]	1,532
RAINBOW-WinoGrande (Validation)	Logical Reasoning	sentence with blanks: [choices]	1,267
RAINBOW-CosmosQA (Validation)	Commonsense reasoning	context, question : [answer choices]	2,985
RAINBOW-SocialIQA (Validation)	Social commonsense reasoning	context, question : [answer choices]	1,954
RAINBOW-PhysicalIQA (Validation)	Physical commonsense reasoning	objective : [how-to answer choices]	1,838

Table 1: Summary of Benchmark Datasets

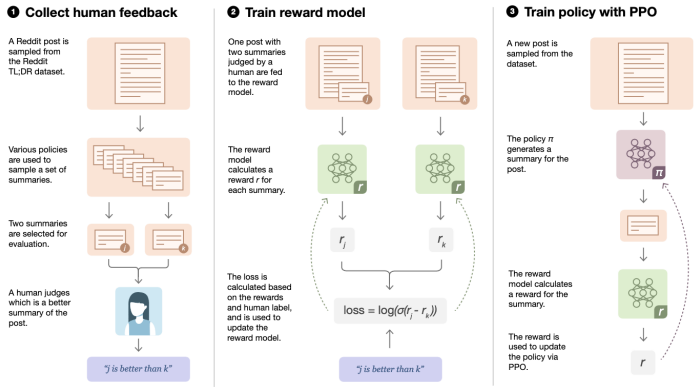

Figure 2: Sample diagram of human feedback, reward model training, and policy training procedure introduced in Stiennon et al. [2022]

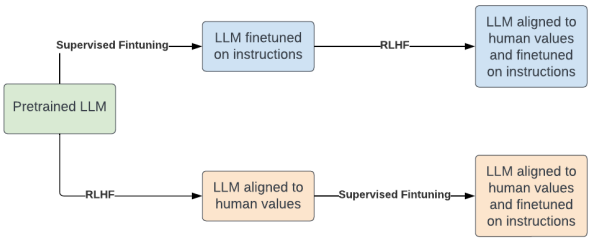

Figure 3: Overview of the training process

4 Methodology

Figure 3 outlines the training process in our experiments. In addition to the steps shown in the figure, we did one round of evaluation at each step. As is shown in the same figure, we used two methods for alignment. To align the model to follow instructions, we used Supervised Finetuning (SFT) and to align it to human values we used Reinforcement Learning with Human Feedback (RLHF). Table 2 shows the hyperparameters chosen for training, as well as computation resources used.

4.1 Supervised Finetuning (SFT)

For supervised finetuning we used the Natural Instructions dataset. The train subset of this dataset (All English tasks) consists of a variety of NLP tasks converted into instruction format. The tasks in the evaluation subset are different from the ones in the train set, which makes this a rather difficult data to learn from. To prevent the model from learning our benchmark tasks from the instruction datasets, we removed tasks related to our benchmarks from the training set. From each task in the training set we sampled 50 instances. Each instance was converted to the instruction format, in a way that "definition" and "input" appeared as the input to the model, and the output was expected to be the target answer. Since this is a text generation task, performance of the model on this task is measured by metrics like ROUGE [Lin, 2004] and BLEU [Papineni et al., 2002].

	initial lr	lr scheduling	# epochs	batch size	max input length	GPU used
SFT on Natural Instructions	5e-4	Cosine	2	4	512	NVIDIA Tesla v100 16GB
RLHF on HH dataset	1e-4	-	1	16	1024	NVIDIA Tesla T4 16GB

Table 2: Hyperparameters and computation resources used for training

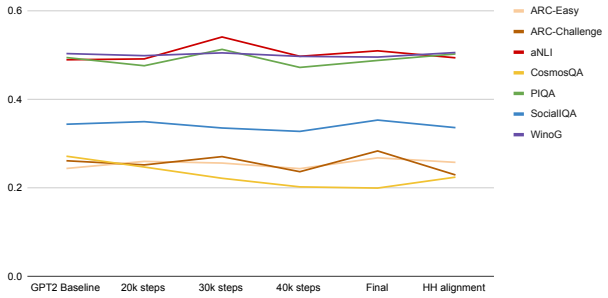

Figure 4: Benchmarks to assess forgetting during SFT

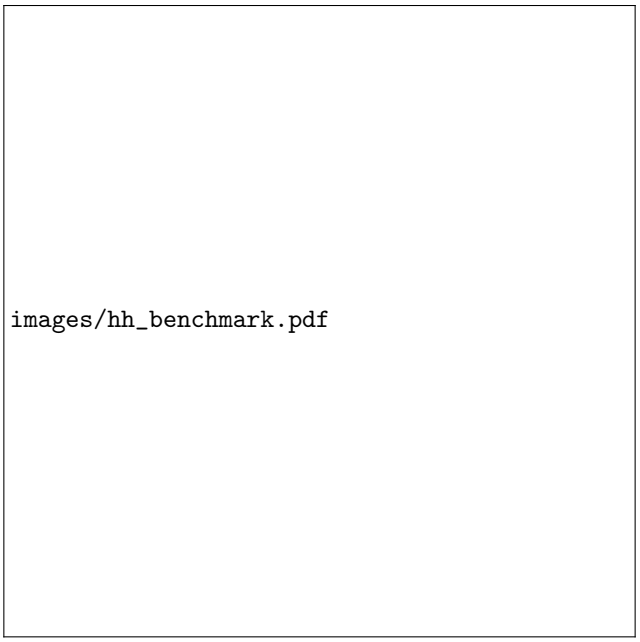

Figure 5: Benchmarks to assess forgetting during RLHF

4.2 Reinforcement Learning with Human Feedback (RLHF)

RLHF is a technique in natural language processing where data created from human feedback is used to train a reward model. The reward model is then used in a reinforcement learning setting to finetune an agent(in this case a language model) to produce outputs aligned with human values. Several optimization algorithms can be used with this approach; We used PPO (Proximal Policy Optimization) [Schulman et al., 2017].

- **Reward Model** : As our reward model we used a model pretrained on several datasets, among which was Anthropic's HH dataset. This is a deBERTa-based model that assigns a score to each input text. The score shows how helpful and harmless the given text is. [1]
- **Optimization** : We used the TRL library from Huggingface [2] to set up a reinforcement learning environment. This library has a PPO algorithm implemented.
- **SFT before RLHF** : As it is common in the literature, we first trained GPT2-medium as a language model on some of the HH data. This is only done for a few hundred iterations to make sure the data format is known to the model. After this we ran the RL algorithm.

5 Results and Analysis

To evaluate the performance of the models generated during SFT, we use the test split of the Super-Natural Instructions. With the text generation objective, we query the model to generate answers for a (instruction, input/question) pair, and use ROUGE scores (ROUGE-1 and ROUGE-L) to compare the generated answers with the correct answers. The results across various checkpoints are displayed in figure 6. The finetuned variants of the pre-trained model generalises well on the unseen dataset until 41.5k steps, after which we see a dip in performance. This might be an indication of overfitting to the training subset, as a consequence of more passes through the training dataset.

To assess the extent of catastrophic forgetting, we used the Commonsense datasets described in section 3. We used prompting, and converted the commonsense datasets into (instruction, question/input, answer choice) triplets, and queried the model to score each answer choice, with the highest score being assigned to the most logical answer choice. This is known as zero-shot classification. As a first step, we evaluated the pre-trained GPT2-medium model on the 'test' split of Natural Instructions and the commonsense datasets to report the models' initial performance.

[1] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2
[2] https://huggingface.co/docs/trl/index

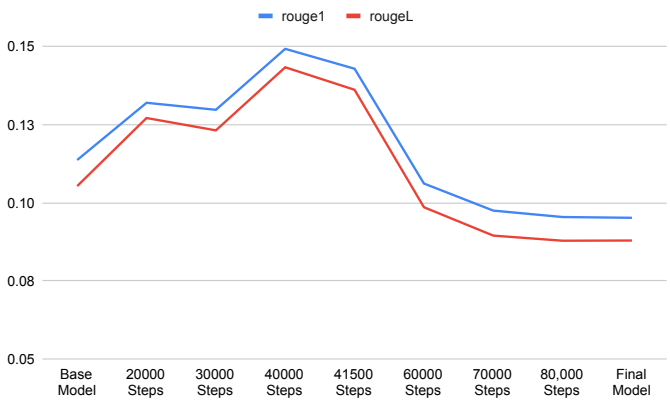

Figure 6: ROUGE scores across checkpoints during instruction tuning (with SFT)

For alignment to human instructions/intentions via SFT, we finetuned the chosen model on the train split of Super-Natural Instructions, and evaluated it on the assessment datasets at regular checkpoints during the finetuning process. Figure 4 shows the performance of the baseline model and its subsequent finetuned variants on the assessment datasets. The x-axis represents the progression of fine-tuning steps, beginning with the GPT2-medium baseline and continuing through incremental steps of 20k, 30k, up to 90k steps (90k steps being the final model after 2 epochs). For the ARC Easy and PiQa, we notice some marginal improvements when comparing the baseline to the final model. However, for all other datasets, the performance of the model either degrades or remains consistent with the baseline. With the model at 41.5k steps, we further train the model at this checkpoint with the RLHF technique on HH data to incorporate the alignments to human values and notice a slight drop in performance with respect to the model at checkpoint 41.5k.

For alignment to human values via RLHF, we follow a similar process wherein we align the pretrained model to the train split of the HH dataset and benchmark the model at various checkpoints during the alignment process (200, 400, 16k & 20k steps). From figure 5, we observed that the evaluation on ARC-Easy shows a huge performance increase between the initial and the final model. However, for all other evaluation datasets, the performance remains the same or degrades over time. As a last step, we use the final human-values aligned model and finetune it further on human instructions dataset. We noticed a drop in performance as a result of this follow-up instruction alignment for ARC-Easy dataset, whereas for all other datasets, the performance of the model was in line with the previous checkpoint.

6 Conclusion

Per our original proposal, we were able to complete our study using Luo et al. [2023] as a guideline for testing for catastrophic forgetting when finetuning LLMs. We used a smaller model than any of those present in their study, and we finetuned on an instruction dataset and benchmarked on checkpoints throughout the finetuning process, where they finetuned on five text summarization datasets and benchmarked after each one. We did have overlap between our benchmarks and their reasoning domain benchmarks, which were the most direct comparison. In this, we did not find significant signs of catastrophic forgetting - this is in line with Luo et al. [2023]'s conclusion because, if catastrophic forgetting scales with model size, the effect will summarily be much smaller and harder to detect in smaller models.

Another observation (regarding figure 4) was that the performance of the model fluctuated around the lower bound for most of the datasets.(e.g. for WinoGrande that is a binary choice dataset performance is about 50%, and for SocialIQA that is comprised of three-choice questions it it about 30%) This is not surprising, as we only measured GPT2-medium via zero-shot classification. Given the relatively small size of GPT2-medium, it is expected that it performs near random, and not change much. This can be easily fixed by increasing the size of the model, and using an LLM with proven zero-shot capabilities. Given our limited time and resources we could not do this. Having seen Qi et al. [2023]'s work, we anticipated alignment to the helpfulness and harmfulness data before or after instruction-tuning would show a drastic change in benchmark performance. As we are confident about the reinforcement learning procedure, lack of eyecatching results may be due to small models and small data. In implementing and testing these techniques, we learned more deeply about the finetuning techniques. Our largest challenge was scaling down the task to match the resources available to us in the time frame. Doing so we also learned about Parameter-Efficient Fine-Tuning (PEFT) [3] for training large models with small hardware resources. As future work, we would like to see how performance changes across multiple epochs, as well as how the reward model used in RLHF affects catastrophic forgetting in larger scale models.

[3] https://github.com/huggingface/peft

7 Contribution

In table 3, we summarize the individual contributions of team members for the project. Tasks were assigned to ensure an equitable distribution of workload, given each member's background and technical skills.

	Pegah	Kalina	Arash	Dikshya
Hypothesis formulation	✓	✓	✓	✓
Literature Review	✓	✓	✓	✓
Data and Model Selection	✓	✓		✓
Experimentation	✓			✓
Evaluation	✓	✓	✓	✓
Poster, Report and Proposal	✓	✓	✓	✓

Table 3: Contribution of the team members to the project

References

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2023.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji aand Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try arc, the AI2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL `https://api.semanticscholar.org/CorpusID:3922816`.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. *ArXiv*, abs/2103.13009, 2021. URL `https://api.semanticscholar.org/CorpusID:232335877`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL `https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf`.

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`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002. URL `https://api.semanticscholar.org/CorpusID:11080756`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.