UNDERSTANDING THE EFFECTS OF RLHF ON LLM GENERALISATION AND DIVERSITY

Robert Kirk[*] [α] **Ishita Mediratta** [β] **Christoforos Nalmpantis** [β] **Jelena Luketina** [γ]
Eric Hambro [β] **Edward Grefenstette** [α] **Roberta Raileanu** [β]

[α] University College London, [β] Meta, [γ] University of Oxford

ABSTRACT

Large language models (LLMs) fine-tuned with reinforcement learning from human feedback (RLHF) have been used in some of the most widely deployed AI models to date, such as OpenAI's ChatGPT or Anthropic's Claude. While there has been significant work developing these methods, our understanding of the benefits and downsides of each stage in RLHF is still limited. To fill this gap, we present an extensive analysis of how each stage of the process (i.e. supervised fine-tuning (SFT), reward modelling, and RLHF) affects two key properties: out-of-distribution (OOD) generalisation and output diversity. OOD generalisation is crucial given the wide range of real-world scenarios in which these models are being used, while output diversity refers to the model's ability to generate varied outputs and is important for a variety of use cases. We perform our analysis across two base models on both summarisation and instruction following tasks, the latter being highly relevant for current LLM use cases. We find that RLHF generalises better than SFT to new inputs, particularly as the distribution shift between train and test becomes larger. However, RLHF significantly reduces output diversity compared to SFT across a variety of measures, implying a tradeoff in current LLM fine-tuning methods between generalisation and diversity. Our results provide guidance on which fine-tuning method should be used depending on the application, and show that more research is needed to improve the tradeoff between generalisation and diversity.

1 INTRODUCTION

Large language models (LLMs) have become a standard approach to solving natural language processing (NLP) tasks. As these models become more capable, the tasks we want them to solve become more complex, which makes it more difficult to provide training data and to evaluate performance. For such tasks it may be easier and faster for humans to evaluate or rank model outputs than provide full demonstrations. Thus, there has been much recent work on using human preferences in this form to fine-tune LLMs, with one dominant approach being reinforcement learning from human feedback (Christiano et al., 2017; Ziegler et al., 2020, RLHF). This approach has been used to produce some of the most impressive AI systems that exist today (Glaese et al., 2022; OpenAI, 2022; 2023; Anthropic, 2023).

The standard RLHF fine-tuning pipeline generally consists of three stages: *supervised fine-tuning* (SFT), where the pretrained model is fine-tuned with the language modelling loss on demonstrations of the desired behaviour; *reward modelling* (RM), where the pretrained model is fine-tuned to predict human preferences between pairs of outputs for a given input; and *reinforcement learning* (RL), where the reward model is used to fine-tune the model produced by the SFT stage using an on-policy RL algorithm like PPO (Schulman et al., 2017). While this pipeline has been used to seemingly great success, there is little understanding about how each component contributes to the behaviour of the final model. Two important areas where the effects of each component of the pipeline have been underexplored are the out-of-distribution (OOD) generalisation and output diversity.

[*] Work partly done during an internship at Meta. Correspondence to robert.kirk.20@ucl.ac.uk. For details of author contributions, see Author Contributions.

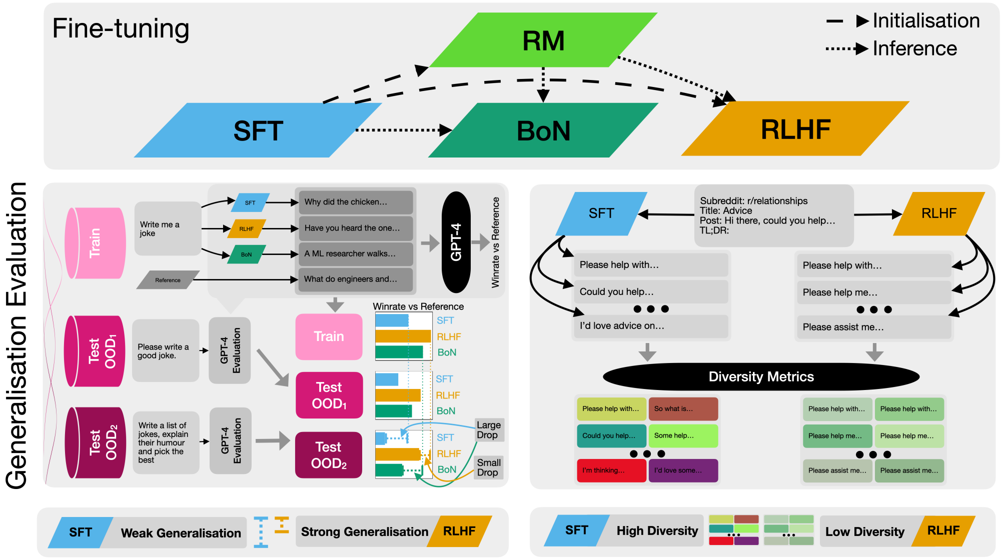

Figure 1: **Overview of Experimental Protocol and Conclusions.** In this work, we fine-tune large language models (LLMs) with three different techniques (SFT, BoN, and RLHF), and evaluate their out-of-distribution generalisation (using GPT-4 as a simulated human evaluator) and output diversity (using a range of metrics from the literature). We find that RLHF has stronger generalisation performance but lower output diversity than SFT, demonstrating a tension between these two desirable properties in current LLM fine-tuning techniques.

OOD generalisation is important for the widespread adoption of these models, since it is necessary to ensure that LLMs are performant and reliable in a wide variety of situations that go beyond the distribution of the training data. While anecdotally models seem to perform well across a wide range of settings, it is unknown which stage of the pipeline is responsible for this strong generalisation, and even whether the observed generality is due to the training or fine-tuning methods used, the large model size, or purely from the very large and diverse distribution of data.

Further, these models are used in creative or open-ended domains such as story generation (Castricato et al., 2022), scientific research (Boiko et al., 2023), or other tasks where a diverse output distribution is required, such as red teaming (Perez et al., 2022). In these situations training models that produce diverse (but still high-quality) outputs is of crucial importance. There has been some speculation as to the possible effects of different steps of the RLHF pipeline on diversity (janus, 2022), and some work has shown a decrease in diversity from RLHF (Khalifa et al., 2021; Perez et al., 2022). However, no rigorous analysis has been done of the effects of different parts of the RLHF pipeline on output diversity across different tasks. In addition, in contrast with our paper, prior work has been limited to evaluating diversity using simple token-level metrics such as BLEU (Papineni et al., 2002) and on use cases which are not as common in practice.

In this work, we evaluate each stage of the RLHF pipeline (i.e. supervised fine-tuning, reward modeling, and reinforcement learning) as well as best-of-N (BoN) sampling in terms of their effects on in-distribution (ID) performance, OOD performance, and output diversity (Fig. 1). We disentangle the effects of the data set and the training method on generalisation by using OOD test datasets that induce realistic distribution shifts between training and testing, and evaluate generalisation using these test sets at each stage of the pipeline (Section 5.1). While diversity is a difficult concept to operationalise, we take a pragmatic approach and measure a range of diversity metrics at each step of the pipeline, covering syntactic, semantic, and logical diversity (Section 5.2). We evaluate both diversity of outputs sampled for a single input and for a range of inputs. Evaluating BoN as well as SFT and RLHF enables us to uncover whether differences between RLHF and SFT are due to the use of a reward model or the type of optimisation applied.

In summary, we find:

- RLHF improves in-distribution performance (as expected from previous work) and also OOD performance in comparison to SFT.

- However, RLHF substantially decreases the diversity of outputs sampled for a given input compared to SFT.

- Even when sampling outputs for different inputs, RLHF produces less diverse text on some metrics, implying that such models tend to produce more similar text regardless of the input.

These findings reveal an inherent tension between generalisation and diversity when applying current fine-tuning techniques. This underscores the necessity for novel methods that improve both these attributes without sacrificing one for the other, and for research to understand whether this tension is fundamental to fine-tuning or a deficit of current techniques. We open source our code to enable reproducible research here: `https://github.com/facebookresearch/rlfh-gen-div`.

2 Background and Related Work

Fine-tuning Large Language Models. The current common practice in NLP is to fine-tune large pre-trained language models (LLM) for downstream tasks. The standard approach for fine-tuning is supervised fine-tuning (SFT), which trains the model on demonstrations of solving the task using supervised learning. When it is easier to evaluate or rank model outputs than it is to gather demonstrations that accurately perform the desired task, an alternative method called reinforcement learning from human feedback (Christiano et al., 2017; Ziegler et al., 2020, RLHF) can be used. Most previous work on RLHF uses on-policy RL algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Advantage Actor Critic (A2C) (Mnih et al., 2016), but offline RL methods have also been proposed (Snell et al., 2022). Once a RM has been trained, it can also be used to do Best-of-N (BoN) sampling (also called rejection sampling) of the model outputs. (Casper et al., 2023) survey existing problems with RLHF as a method for fine-tuning LLMs, and our work further investigates some of these problems (specifically policy generalisation and output diversity or "mode collapse").

Dubois et al. (2023, AlpacaFarm) introduces a framework for developing methods for learning from human feedback in an instruction following setting, and in this framework demonstrate that approaches that learn from human feedback (including RLHF) generally perform better than SFT on a specific evaluation set they introduce. We use the AlpacaFarm models and evaluation dataset in our experiments on instruction following. In that work they do not evaluate out-of-distribution generalisation or output diversity, and only present results on a single evaluation dataset, while we address all of these issues, while also performing the analysis in an additional task (summarisation).

We discuss other recent approaches for fine-tuning LLMs in Appendix C, but note that while these works sometimes show improvements, they are not used by most large-scale systems being deployed currently, and hence we focus our analysis on the more popular and widely used RLHF pipeline, as that is where understanding will be most relevant and useful.

Generalisation and Diversity in NLP. Almost all prior work using RLHF has evaluated models on the same distribution of inputs used for fine-tuning (Bai et al., 2022; Glaese et al., 2022; Ouyang et al., 2022; Stiennon et al., 2022), meaning that the generalisation properties of such methods isn't understood. One notable exception is Stiennon et al. (2022), who perform some experiments evaluating their models trained the TL;DR dataset (Völske et al., 2017) (reddit post summarisation) on the CNN/Daily Mail dataset (Nallapati et al., 2016) (news article summarisation). However, they didn't investigate how different parts of the pipeline affected generalisation, and the investigation was less rigorous and involved than ours is. (Hupkes et al., 2023) provide a comprehensive survey of generalisation in the wider NLP literature.

Several works (Khalifa et al., 2021; Perez et al., 2022) have shown in specific settings that RLHF fine-tuning produces models with less output diversity, as measured by self-BLEU (Zhu et al., 2018). Our work extends these works by making diversity evaluation a primary focus and using diversity metrics beyond self-BLEU which have been externally validated (Tevet & Berant, 2021) and measure diversity in a range of different ways.

We discuss more related work, including details on LLMs, SFT and RLHF, in Appendix C.