Training language models to follow instructions

Introduction

The research paper begins by highlighting that scaling up language models (LMs) does not inherently improve their ability to follow user instructions or align with human intentions. Large models often produce undesirable outputs like false information, biased or toxic content, or simply fail to adhere to instructions. This misalignment stems from the traditional training objective—predicting the next token from internet data—which diverges from the goal of producing helpful, honest, and harmless responses. To address this, the paper explores Reinforcement Learning from Human Feedback (RLHF) as a method to better align LMs with user intent. The researchers fine-tune GPT-3 by collecting human-written demonstrations and preference data to train a supervised model and a reward model, followed by reinforcement learning using Proximal Policy Optimization (PPO). The resulting models, named InstructGPT, are found to be more preferred by human evaluators, outperforming even much larger base GPT-3 models. They also exhibit improved truthfulness and reduced toxicity, though they are not flawless. Overall, the paper proposes RLHF as a promising path toward aligning LMs with human values and use cases.

Related Work

The research paper situates its contributions within several key domains of existing literature. It builds upon prior work in alignment and learning from human feedback, particularly the use of Reinforcement Learning from Human Feedback (RLHF). Originally developed for training agents in robotics and games, RLHF has more recently been applied to natural language tasks such as summarization, dialogue, and evidence extraction. This study extends RLHF to a broader and more diverse range of language tasks.

The paper also draws on research in instruction-following language models, where models are fine-tuned on a mixture of NLP tasks with natural language instructions to promote generalization across tasks. Such instruction-tuning has been shown to improve performance in both zero-shot and few-shot settings.

In addressing model safety, the paper references work focused on evaluating and mitigating harms such as bias, toxicity, and misinformation. Existing approaches include filtering training data, modifying generation behavior, using control tokens, and leveraging human-in-the-loop systems. The paper acknowledges challenges in reliably reducing harm without introducing unintended side effects, especially for marginalized groups. This context frames the paper's approach of aligning model behavior using human feedback as a practical and scalable method.

Methods and Experimental Details

The research paper outlines a three-step methodology to align language models with user intent using reinforcement learning from human feedback (RLHF). The process begins with supervised fine-tuning (SFT), where a pretrained GPT-3 model is fine-tuned on human-written demonstrations that illustrate the desired behavior. These prompts come from labelers and users of the OpenAI API, and they cover a broad spectrum of tasks including generation, question answering, summarization, and classification.

The second step involves training a reward model (RM) using comparison data. Labelers rank multiple model responses for the same prompt, and the RM is trained to predict which responses humans prefer. To efficiently use labeler input, the method generates multiple pairwise comparisons from each set of ranked responses and batches them during training to avoid overfitting.

The final step applies Proximal Policy Optimization (PPO) to fine-tune the SFT model using the reward signal from the RM. This reinforcement learning setup treats the model as a policy in a bandit environment, where each prompt-response pair yields a scalar reward. An additional KL-divergence penalty is used to prevent over-optimization and ensure output diversity. A variant of this method, called PPO-ptx, incorporates gradients from the original pretraining distribution to reduce performance regressions on public NLP benchmarks.

To support training and evaluation, a team of about 40 trained contractors was employed to write prompts, provide demonstration data, and label comparisons. These labelers were carefully screened for their ability to handle sensitive content and for consistent agreement in evaluations.

The paper uses multiple GPT-3 models of varying sizes (1.3B, 6B, and 175B parameters) and evaluates them based on alignment to user intent, measured through labeler preference ratings and performance on both custom prompts and public NLP datasets. The evaluation framework includes metrics for helpfulness, truthfulness, and harmlessness, and utilizes datasets like TruthfulQA and RealToxicityPrompts to measure alignment outcomes.

Results

The research paper presents results across three domains: performance on real-world API prompts, public NLP benchmarks, and qualitative analyses. It finds that InstructGPT models significantly outperform base GPT-3 models in terms of human preference. For instance, outputs from a 1.3B InstructGPT model are preferred over those from the 175B GPT-3 model, demonstrating the effectiveness of alignment through fine-tuning with human feedback. Even when GPT-3 is prompted with few-shot examples, InstructGPT still receives higher ratings, especially on metrics like appropriateness, adherence to explicit instructions, and reduction of hallucinated content.

When tested with labelers who did not participate in training data collection, InstructGPT still maintains strong performance, suggesting that its alignment generalizes beyond its original supervision sources. Moreover, comparisons with models fine-tuned on public instruction datasets like FLAN and T0 show that InstructGPT outperforms these baselines on OpenAI's API-style prompts, reinforcing the paper's claim that public NLP datasets do not fully reflect real-world language model usage.

On public NLP benchmarks, InstructGPT shows measurable gains in truthfulness, nearly doubling performance over GPT-3 on the TruthfulQA dataset. It also hallucinates less frequently on closed-domain tasks. For toxicity, the models show improvement when prompted to be respectful, but gains are limited or reversed under neutral conditions. Bias measurements on datasets like Winogender and CrowS-Pairs indicate that InstructGPT is not less biased than GPT-3, with some models showing higher certainty in biased outputs.

The study further shows that performance regressions on NLP benchmarks (an "alignment tax") can be reduced by mixing in pretraining gradients during PPO fine-tuning (PPO-ptx). This hybrid approach mitigates the drop in performance on tasks like SQuAD and DROP without reducing alignment quality.

Finally, qualitative results highlight that InstructGPT generalizes to tasks outside of its fine-tuning distribution, such as following instructions in non-English languages and answering questions about code. However, it still makes simple errors, such as accepting false premises, over-hedging in responses, or failing to meet complex constraints—underscoring the need for continued refinement.

Discussion

The research paper reflects on broader implications and limitations of aligning language models using human feedback. It emphasizes that alignment techniques like Reinforcement Learning from Human Feedback (RLHF) are not only effective but also relatively low-cost compared to training massive language models from scratch. For instance, training InstructGPT models requires a fraction of the compute used for GPT-3 while yielding superior user-aligned performance. This makes alignment a cost-efficient strategy, especially when model size increases offer diminishing returns on instruction-following capabilities.

The paper notes promising evidence that InstructGPT can generalize alignment beyond its supervised training distribution, including on tasks involving code or non-English instructions. This generalization is crucial because it is infeasible to collect human feedback for every possible task. The study also finds that most performance regressions on standard NLP benchmarks can be mitigated using the PPO-ptx approach, reducing what it refers to as the "alignment tax."

However, the discussion also addresses a key philosophical and practical question: who exactly the model is being aligned to. Since the alignment process depends heavily on the preferences of a specific group of labelers—selected contractors from particular regions and cultural backgrounds—InstructGPT reflects the norms, interpretations, and biases of that group. Furthermore, these preferences are shaped by researchers' written instructions and API customers' submitted prompts, raising questions about representativeness and fairness. The paper cautions against assuming that such alignment represents universal human values and encourages future efforts to develop systems that can accommodate or condition on different value systems.

Several limitations are acknowledged, including the small and non-representative pool of labelers, single-label evaluations for most data points, and the persistence of unsafe behaviors. The paper highlights that InstructGPT can still follow harmful instructions or generate toxic, biased, or inappropriate content if prompted to do so. It emphasizes the need for more participatory and accountable alignment processes.

Finally, the paper identifies open research questions, such as how to reduce harmful outputs more reliably, how to better measure alignment beyond proxy metrics, and how to ensure broader societal representation in alignment data. It calls for techniques that are scalable, low-cost, and generalizable to future, more powerful AI systems.

Conclusion

The research paper demonstrates that fine-tuning language models using human feedback specifically through supervised learning followed by reinforcement learning from human preferences (RLHF), is a powerful method for aligning model behavior with user intent. The resulting InstructGPT models outperform the much larger base GPT-3 models in terms of helpfulness, truthfulness, and adherence to instructions, despite having significantly fewer parameters. These models also show improved reliability, reduced hallucinations, and lower toxicity when guided appropriately, all while maintaining competitive performance on public NLP benchmarks through techniques like PPO-ptx.

The study highlights that alignment is not only technically feasible but also computationally efficient, offering a cost-effective alternative to simply scaling model size. However, it also underscores the challenges of defining whose preferences are being optimized and the ethical complexity involved in alignment decisions. Limitations remain, particularly in handling harmful or biased prompts, and in ensuring that the alignment reflects diverse perspectives beyond the narrow set of labelers and contexts used in this work.

Overall, the research presents RLHF as a promising and scalable approach to producing language models that are more aligned with human values, while acknowledging that further work is needed to broaden representativeness, reduce harmful outputs, and solidify the foundations of alignment for more capable future systems.

References

[1] Training language models to follow instructions

Summary By:
Kartikey Vijayakumar Hebbar
Northeastern University, Boston
Intro to AI Agents - Summer 2025