

In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

Prompt: Write a title for the above text.

Assistant A: Improving Fine-Tuning for Language Models: A GPT-3-inspired Approach

Assistant B: Training language models to follow instructions with human feedback

```

Your output should be:

```

```
{"reason": "This text discusses how to make large language models follow user instructions better, and Assistant B's response is more in line with the meaning of the text.", "choice": "B"}
```

```

Example 3:

```

Text: Though technology companies announced massive layoffs last year, 2023 has been much worse. Layoffs have far outpaced last year's cuts, as tech giants including Amazon, Facebook parent company Meta, Microsoft, Google, IBM, SAP, and Salesforce - as well as many smaller companies - announce sweeping job cuts. The problem: Big Tech went on a hiring binge during the pandemic when lockdowns sparked a tech buying spree to support remote work and an uptick in e-commerce, and now they face revenue declines. Although global IT spending is forecast to rise in 2023, with enterprise software and IT services experiencing the greatest growth, the overall increase is expected to be modest, with data center systems and communications services growing by less than 1%, according to market research firm Gartner. Meanwhile hardware sales are forecast to decline. Continuing supply chain issues, inflation, and the war in Ukraine are also having an impact on both business and consumer spending, leading to fears of recession. According to data compiled by Layoffs.fyi, the online tracker keeping tabs on job losses in the technology sector, 784 tech companies have laid off about 208,181 staff so far this year, compared to 164,411 layoffs last year.

Prompt: How many job losses have occurred in the technology sector so far in 2023?

Assistant A: 208,181 job losses have occurred in the technology sector so far in 2023.

Assistant B: 208,181.

```

Your output should be:

```

```
{"reason": "Both answers are acceptable and correct. They should be a tie.", "choice": "Tie"}
```

```

Your response should only be in the JSON format above; THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE. Write the "reason" key before writing the "choice" key, so that you think step-by-step before