# Article information:

Aligning Language Models to Follow Instructions  
<https://openai.com/blog/instruction-following/>

# Article summary:

1. InstructGPT models are better than GPT-3 at following instructions and producing helpful outputs.

2. InstructGPT models are less likely to produce biased or toxic outputs when given a sensitive prompt or instruction.

3. Fine-tuning language models with humans in the loop is a powerful tool for improving their safety and reliability.

# Article rating:

Appears moderately imbalanced: The article provides some useful information, but is missing several important points or pieces of evidence that would be required to present the discussed topics in a balanced and reliable way. You are encouraged to seek a more balanced perspective on the presented issues by exploring the provided research topics and looking at different information sources.

# Article analysis:

The article provides an overview of OpenAI's InstructGPT model, which is designed to be better than GPT-3 at following instructions and producing helpful outputs. The article claims that InstructGPT is less likely to produce biased or toxic outputs when given a sensitive prompt or instruction, and that fine-tuning language models with humans in the loop is a powerful tool for improving their safety and reliability.

The article does not provide any evidence to support its claims about the trustworthiness and reliability of the model, nor does it explore any potential biases or risks associated with using such a model. It also fails to mention any counterarguments or alternative perspectives on the use of this technology, making it difficult to assess its potential implications for society as a whole. Additionally, there is no discussion of how this technology might be misused by malicious actors, nor does it address any ethical considerations related to its use. Finally, there is no indication that the authors have considered both sides of the argument equally before making their conclusions about the trustworthiness and reliability of InstructGPT models.

# Topics for further research:

* Ethical considerations of language models
* Risks associated with OpenAI's InstructGPT
* Potential biases in language models
* Misuse of language models
* Counterarguments to using language models
* Implications of language models for society

# Report location:

<https://www.fullpicture.app/item/c07c3acc88715a043cdd0de4d02d0e6f>