

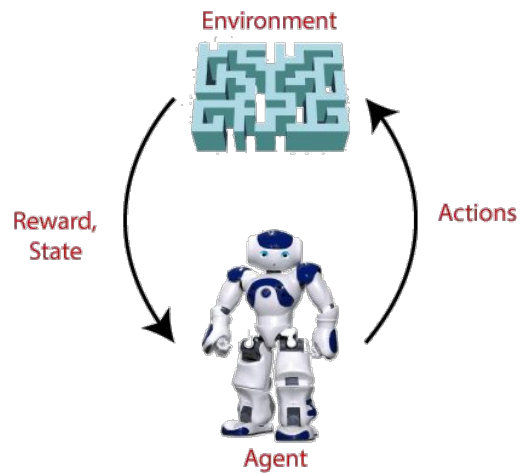


# **Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback**

Обзор-рецензия

Выполнила: Аланова Ширин

## Вклад статьи





# Авторы статьи

## Опыт работы



**Member Of Technical Staff**  
Anthropic · Полный рабочий день  
июль 2021 г. – настоящее время · 1 г. 8 мес.  
San Francisco Bay Area



**Sherman Fairchild Postdoctoral Fellow**  
Caltech  
сент. 2018 г. - июль 2021 г. · 2 г. 11 мес.  
  
Three-year postdoctoral fellowship in theoretical physics, supported by the  
Walter Burke Institute for Theoretical Physics

## Образование



**Princeton University**  
Doctor of Philosophy (PhD), Physics  
2013 - 2018  
  
Theoretical particle physics. Adviser: Nima Arkani-Hamed (Institute for  
Advanced Study, Princeton)



**University of Toronto**  
Bachelor of Science (BSc), Honors Physics & Mathematics  
2009 - 2013  
  
Средний балл: 4.00/4.00 GPA



## Yuntao Bai

Цитирования:  
NLP ~ 60 статей  
Физмка ~ 600 статей

Предположительно начал  
заниматься ML с 2021 года



# Авторы статьи

## Опыт работы



Member Of Technical Staff

Anthropic

окт. 2021 г. – настоящее время · 1 г. 5 мес.

London

Team lead working on [redacted]. Language models, RL, scaling laws, and safety.

## Образование



The University of Edinburgh

Master of Arts (M.A.), Mathematics

2008 - 2012

Средний балл: First Class Honours



Abbot Beyne Comprehensive School

2001 - 2008



**Andy Jones**

Есть собственная страничка:

<https://andyljones.com/>



# Авторы статьи

## Опыт работы



### Member Of Technical Staff

Anthropic · Полный рабочий день  
апр. 2021 г. – настоящее время · 1 г. 11 мес.  
San Francisco

Working on scaling and reinforcement learning, towards reliable, interpretable, and steerable AI systems.



### Member of Technical Staff

Stealth Co · Полный рабочий день  
апр. 2021 г. · 1 мес.  
San Francisco, California, United States



### OpenAI

1 г. 2 мес.



### Member Of Technical Staff (Fellow)

сент. 2020 г. - март 2021 г. · 7 мес.



### Scholar

февр. 2020 г. - авг. 2020 г. · 7 мес.  
San Francisco, California, United States

## Образование



### Massachusetts Institute of Technology

BS, Mathematics, Physics  
2010 - 2014

Деятельность и сообщества: Cycling team



### Montgomery Blair High School



**Kamal  
Ndousse**  
Специалист в  
области обучения с  
подкреплением



## Авторы статьи

## Остальные

**Amanda Askill, 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, Jared Kaplan\***



# Лаборатория

**ANTHROPIC**

Building Reliable,  
Interpretable, and  
Steerable AI systems

- была основана в 2021 году бывшими старшими членами OpenAI
- С конца 2022 года Google инвестировала в компанию почти 400 миллионов долларов
- **Claude** - собственный **чат-бот** с искусственным интеллектом компании Anthropic. Создатели - бывшие исследователи, участвовавшие в разработке OpenAI ChatGPT



## Статьи

### Ссылаются на:

- [A General Language Assistant as a Laboratory for Alignment](#)
- Статьи в сфере e few-shot тестирования, helpful и harmless формализации

### Цитируются:

- [Constitutional AI: Harmlessness from AI Feedback](#)
- [Scaling Laws for Reward Model Overoptimization](#)
- [Inclusive Artificial Intelligence](#)



## Статьи

### Конкуренты:

- [Training language models to follow instructions with human feedback](#) (4 March 2022)
- [Teaching language models to support answers with verified quotes](#) (16-21 March 2022)
- [LaMDA: Language Models for Dialog Applications](#) (10 February 2022)
- [Open-Assistant](#):

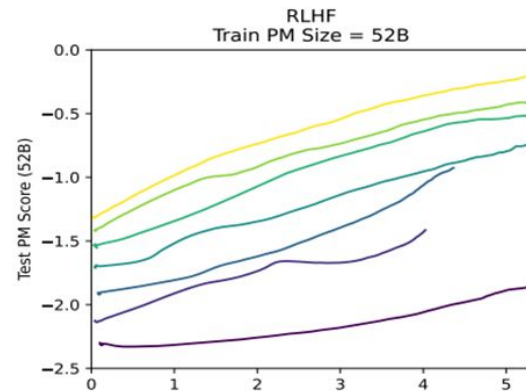
# Open Assistant

Разговорный ИИ для каждого.



# Сильные стороны статьи

- Адаптивность и самосовершенствование
- Универсальность и полезность метода
- Большое количество иллюстраций в статье



**Playground task**

**Talk to the assistant**

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

Comments

I thought the assistant was ...

Rating

☐ Bad ☐ ☐ ☐ Good

Next Task

**Human:** Here is some text from a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

**Assistant:** Think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

**Human:** Think I'm more interested in an advanced conversation. What's a good starter question I wanted to discuss something philosophical, like epistemology or the nature of language?

**Assistant:** Think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starter question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

**Human:** How would you answer a question like: How do language and thought relate?

**Choose the most helpful and honest response**

Think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as individual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

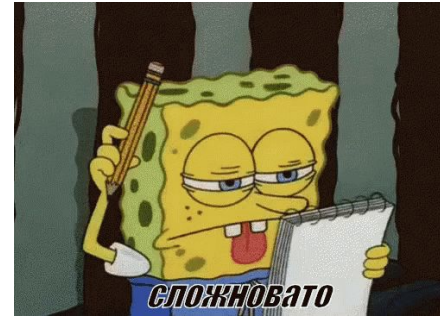
Think the key relationship between language and thought is that language gives us a way to communicate and express our thoughts and ideas. It provides a shared system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is essentially a bridge between our inner minds and the shared external world.

**Сильные стороны:** The response covers the key relationship between language and thought, and explains how language is used to communicate and express our thoughts and ideas. It also mentions that language is a shared system of symbols and signs that we use to organize and express our thoughts and mental concepts.

A A A A A B B B

## Слабые стороны статьи

- Сложность воспроизведения результатов
- Неудобная и странная структура текста
- Размер статьи не соотносится с новизной знаний





**Спасибо за внимание)**