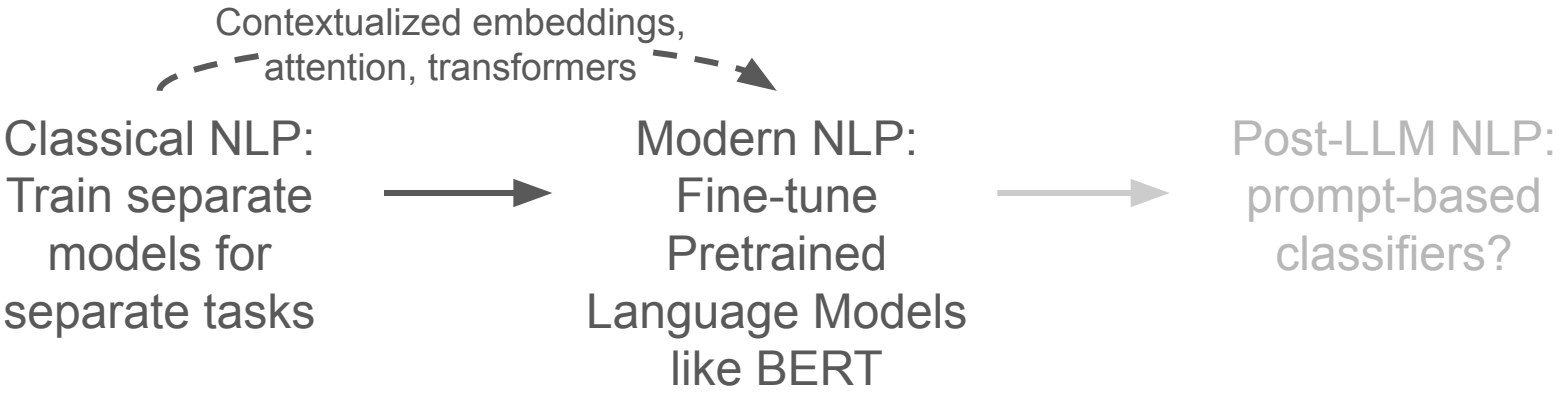
NLP Crash Course III

Indira Sen and David Garcia

The timeline so far: Supervised NLP



The timeline so far: Supervised NLP

Classical NLP: Train separate models for separate tasks → Modern NLP: Fine-tune Pretrained Language Models like BERT → Post-LLM NLP: prompt-based classifiers?

Natural Language Generation (NLG)

NLG focuses on systems that produce fluent, coherent and useful language output for human consumption

Natural Language Generation (NLG)

NLG focuses on systems that produce fluent, coherent and useful language output for human consumption

Open-ended generation: the output distribution still has high freedom

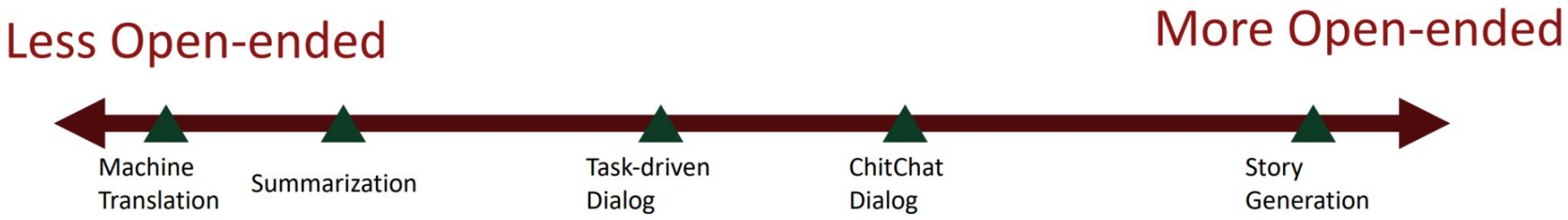
Non-open-ended generation: the input mostly determines the output generation.

Natural Language Generation (NLG)

NLG focuses on systems that produce fluent, coherent and useful language output for human consumption

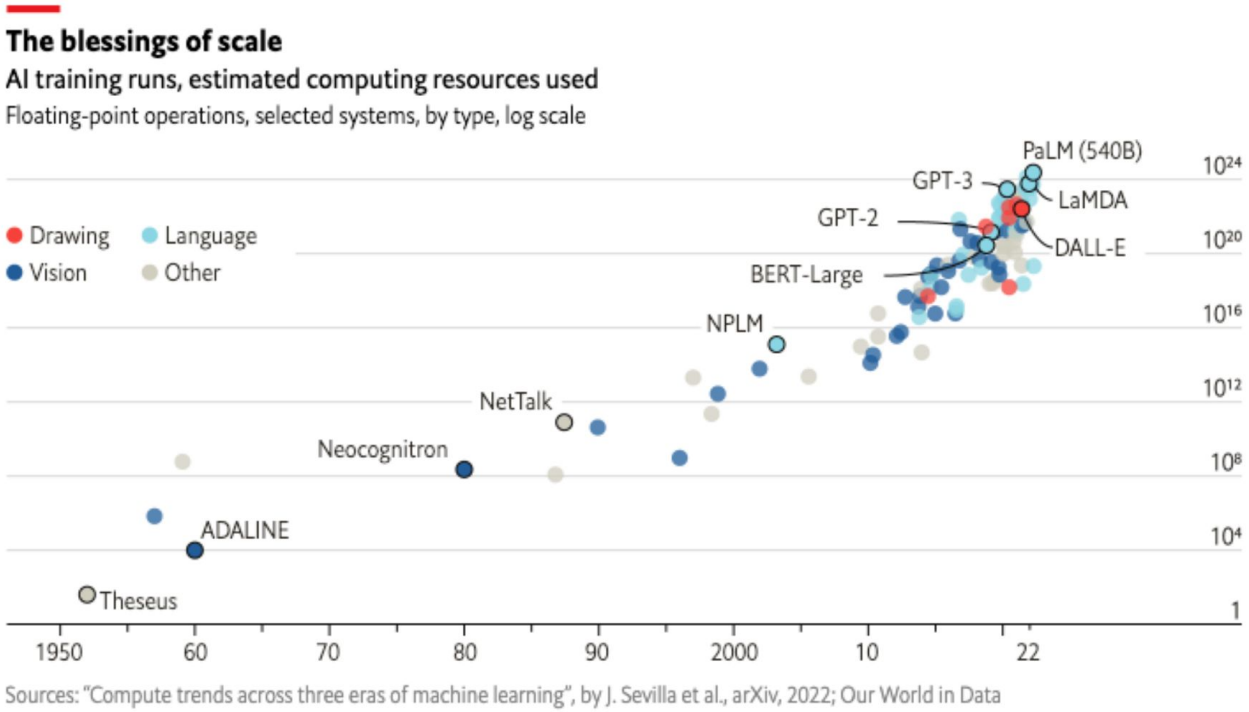
Open-ended generation: the output distribution still has high freedom

Non-open-ended generation: the input mostly determines the output generation.



https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/index.html

Emergent Properties of LMs



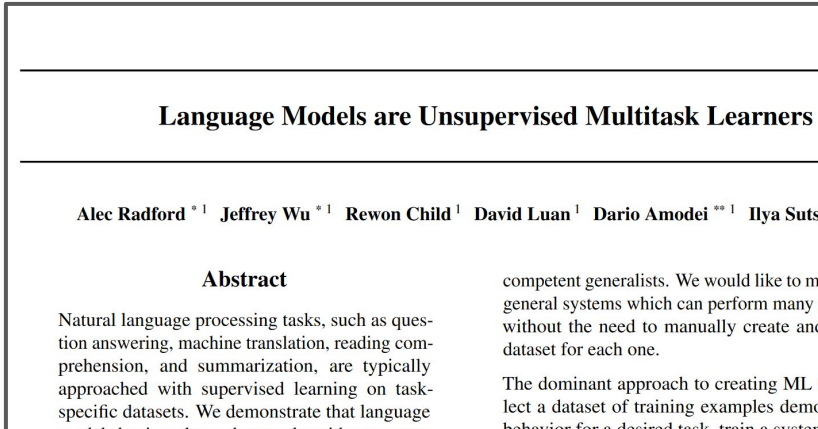
https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress

Emergent Properties of LMs

Let's revisit the Generative Pretrained Transformer (GPT) models from OpenAI as an example:

GPT-2 (1.5B parameters; Radford et al., 2019)

- Same architecture as GPT, just bigger (117M -> 1.5B)
- But trained on much more data: 4GB -> 40GB of internet text data (WebText)
- Scrape links posted on Reddit w/ at least 3 upvotes (rough proxy of human quality)



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Suts

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2,

competent generalists. We would like to m
general systems which can perform many
without the need to manually create an
dataset for each one.

The dominant approach to creating ML
lect a dataset of training examples demo
behavior for a desired task, train a syste
behaviors, and then test its performance
and identically distributed (IID) held-ou
has served well to make progress on nar
the often erratic behavior of captioning m
2017), reading comprehension systems (Ji
and image classifiers (Alcorn et al., 2018
and variety of possible inputs highlights
comings of this approach.

Our suspicion is that the prevalence of si
on single domain datasets is a major cont
of generalization observed in current sy

[Language models are unsupervised multitask learners]

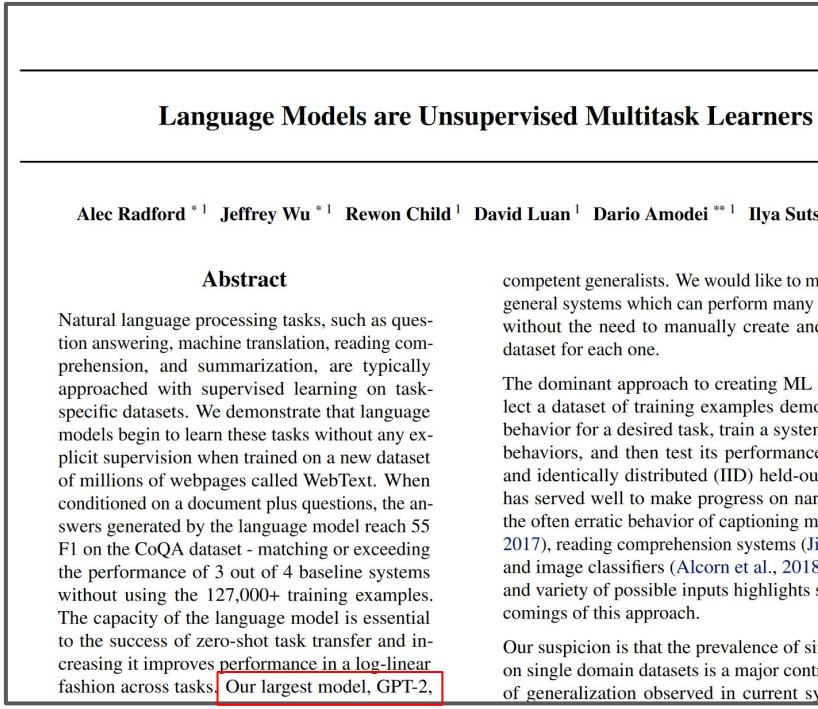
Emergent Properties of LMs

One key emergent ability in GPT-2 is **zero-shot** learning:

the ability to do many tasks with no examples, and no gradient updates, by simply:

- Specifying the right sequence prediction problem (e.g. question answering):

Passage: Tom Brady... Q: Where was Tom Brady born? A: ...



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Suts

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2,

competent generalists. We would like to m
general systems which can perform many
without the need to manually create an
dataset for each one.

The dominant approach to creating ML
lect a dataset of training examples demo
behavior for a desired task, train a syste
behaviors, and then test its performance
and identically distributed (IID) held-ou
has served well to make progress on nar
the often erratic behavior of captioning m
2017), reading comprehension systems (Ji
and image classifiers (Alcorn et al., 2018
and variety of possible inputs highlights
comings of this approach.

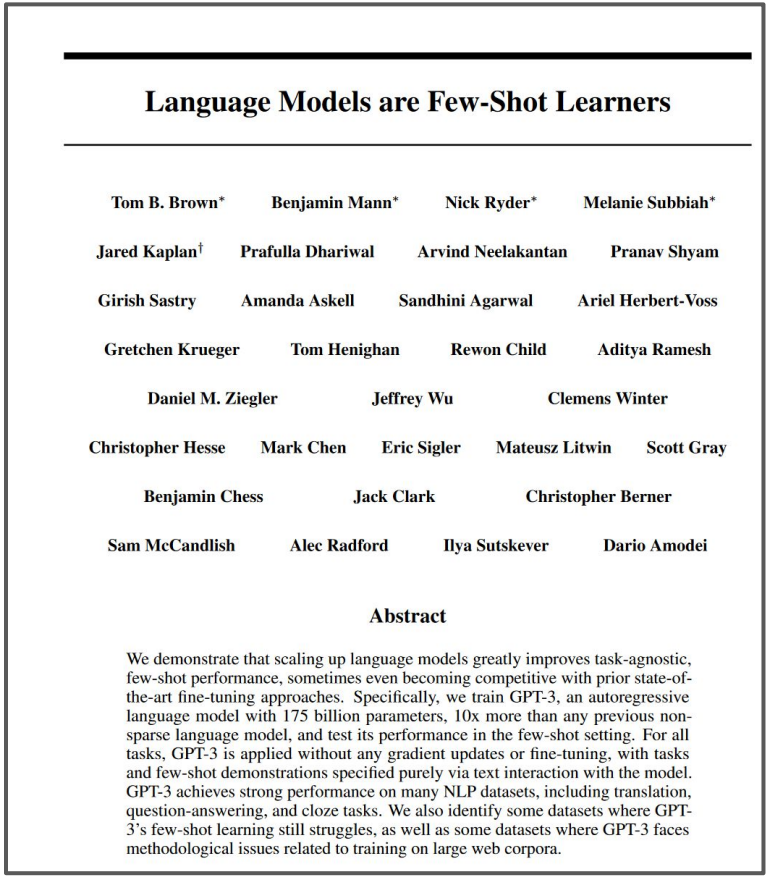
Our suspicion is that the prevalence of si
on single domain datasets is a major cont
of generalization observed in current sy

[Language models are unsupervised multitask learners]

Emergent zero-shot and few-shot learning

Specify a task by simply prepending examples of the task before your example

- Also called **in-context learning**, to stress that **no gradient updates** are performed when learning a new task



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam

Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss

Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh

Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

Abstract

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

Language models are few-shot learners

Instruction Finetuning

Before: sentence completion, not answering a prompt.

Language models are not aligned with user intent

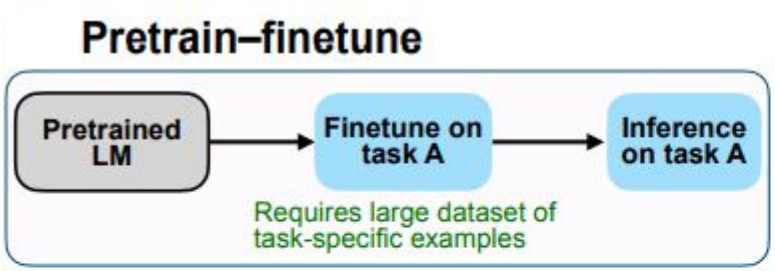
PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3
	`Explain the theory of gravity to a 6 year old.`
	`Explain the theory of relativity to a 6 year old in a few sentences.`
	`Explain the big bang theory to a 6 year old.`
	`Explain evolution to a 6 year old.`

Instruction Finetuning

Before: sentence completion, not answering a prompt.

Language models are not aligned with user intent

Instruction Finetuning!



Finetuned Language Models Are Zero-Shot Learners

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. 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 a language model 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

at fine-tuning with human feedback is a promising direction for aligning
e models with human intent.

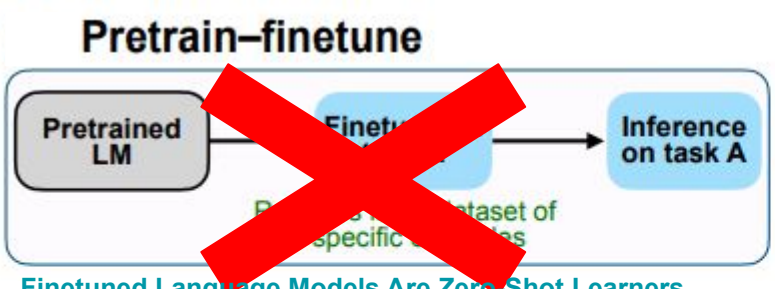
Training language models to follow instructions with human feedback

Instruction Finetuning

Before: sentence completion, not answering a prompt.

Language models are not aligned with user intent

Instruction Finetuning!



Finetuned Language Models Are Zero-Shot Learners

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. 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 a language model 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

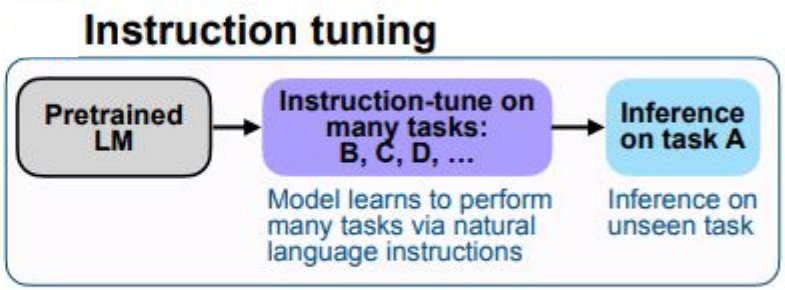
Training language models to follow instructions with human feedback

Instruction Finetuning

Before: sentence completion, not answering a prompt.

Language models are not aligned with user intent

Instruction Finetuning!



Finetuned Language Models Are Zero-Shot Learners

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. 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 a language model 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

Training language models to follow instructions with human feedback

Instruction Finetuning

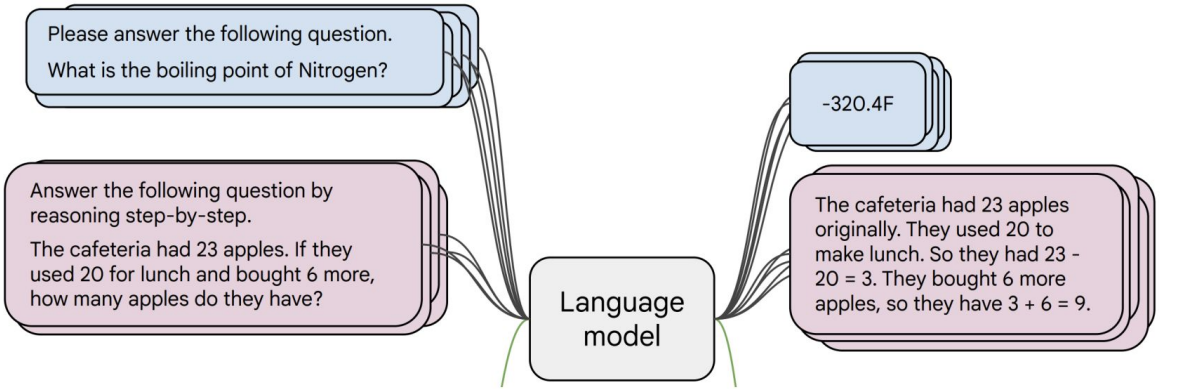
Also in Flan-T5

Scaling Instruction-Finetuned Language Models

Hyung Won Chung* Le Hou* Shayne Longpre* Barret Zoph† Yi Tay†
William Fedus† Yunxuan Li Xuezhi Wang Mostafa Dehghani Siddhartha Brahma
Albert Webson Shixiang Shane Gu Zhuyun Dai Mirac Suzgun Xinyun Chen
Aakanksha Chowdhery Alex Castro-Ros Marie Pellat Kevin Robinson
Dasha Valter Sharan Narang Gaurav Mishra Adams Yu Vincent Zhao
Yanping Huang Andrew Dai Hongkun Yu Slav Petrov Ed H. Chi
enny Zhou Quoc V. Le

nstructions has been shown to improve
per we explore instruction finetuning
g the model size, and (3) finetuning on
above aspects dramatically improves
pting setups (zero-shot, few-shot, CoT),
nded generation, RealToxicityPrompts).
performs PaLM 540B by a large margin
mance on several benchmarks, such as
oints,[1] which achieve strong few-shot
52B. Overall, instruction finetuning is a
trained language models.

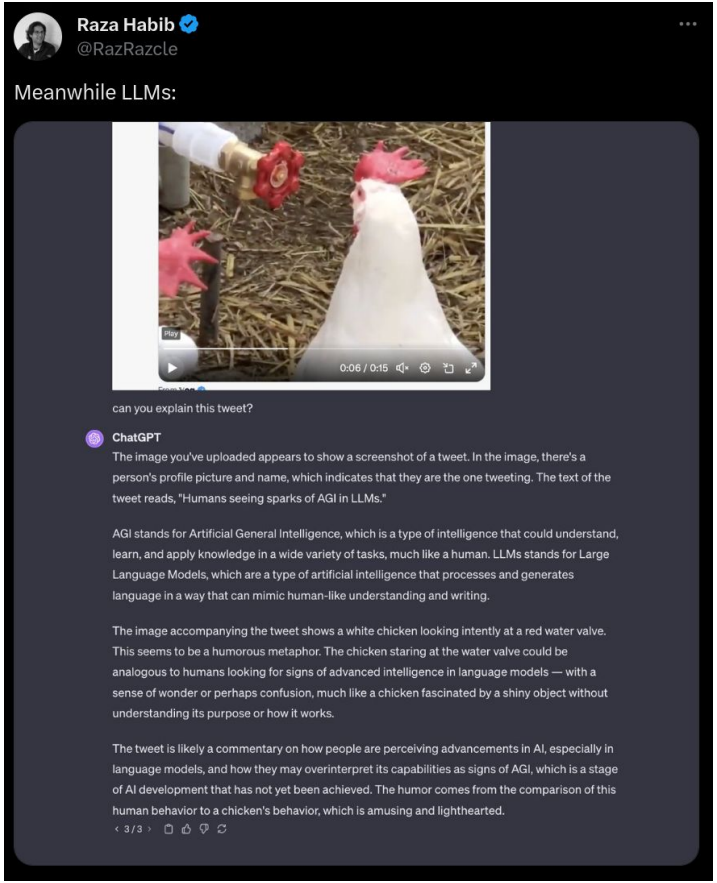
- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



- Evaluate on **unseen tasks**

Scaling instruction-finetuned language models

Sparks of AGI



https://twitter.com/deliprao/status/1724163062830153814
https://twitter.com/RazRazcle/status/1724174924367974827

Reinforcement Learning from Human Feedback (RLHF)

Instruction finetuning is great but has some limitations

Even with instruction finetuning, there a mismatch between the LM objective and the objective of “satisfy human preferences”!

Can we explicitly attempt to satisfy human preferences?

Reinforcement Learning from Human Feedback (RLHF)

Instruction finetuning is great but has some limitations

Even with instruction finetuning, there a mismatch between the LM objective and the objective of “satisfy human preferences”!

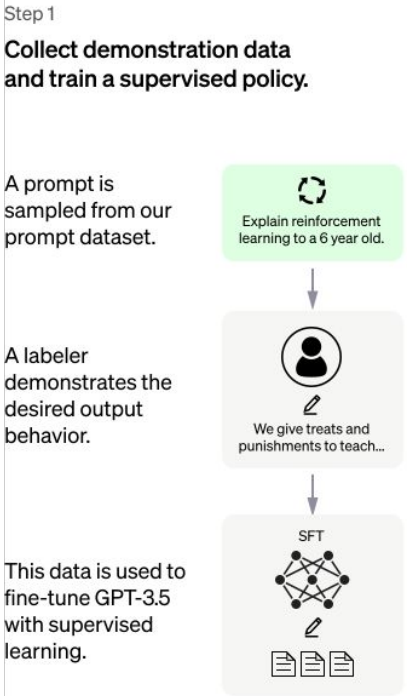
Can we explicitly attempt to satisfy human preferences?

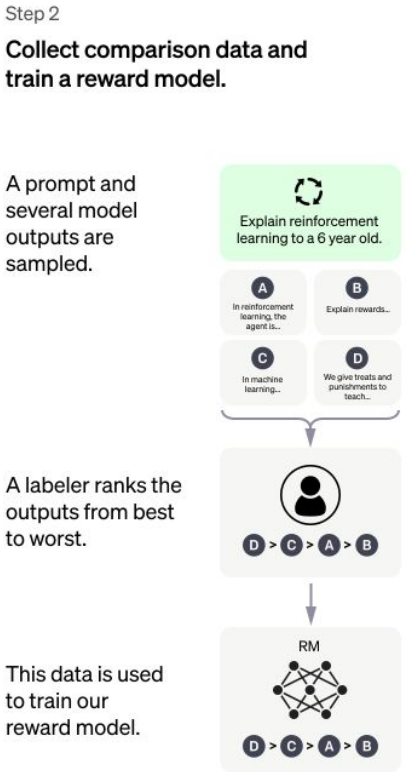
Reinforcement Learning: area of machine learning and optimal control concerned with how an intelligent agent ought to take actions in a dynamic environment in order to maximize the cumulative reward.

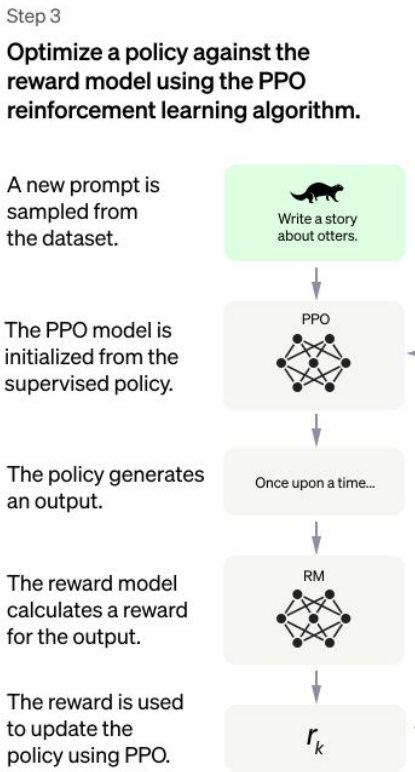
https://en.wikipedia.org/wiki/Reinforcement_learning

In the InstructGPT paper,they also had RLHF

[Training language models to follow instructions with human feedback]







Training language models to follow instructions with human feedback

lmeida* Carroll L. Wainwright*

wal Katarina Slama Alex Ray

Luke Miller Maddie Simens

Paul Christiano*†

Ryan Lowe*

y make them better at following
ls can generate outputs that are
In other words, these models are
an avenue for aligning language
ks by fine-tuning with human
rompts and prompts submitted
set of labeler demonstrations of
e-tune GPT-3 using supervised
model outputs, which we use to
forcement learning from human
GPT. In human evaluations on
rameter InstructGPT model are
having 100x fewer parameters.
ts in truthfulness and reductions
rformance regressions on public
es simple mistakes, our results
romising direction for aligning

Tasks collected from labelers

- **Plain:** We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
- **Few-shot:** We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
- **User-based:** We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.

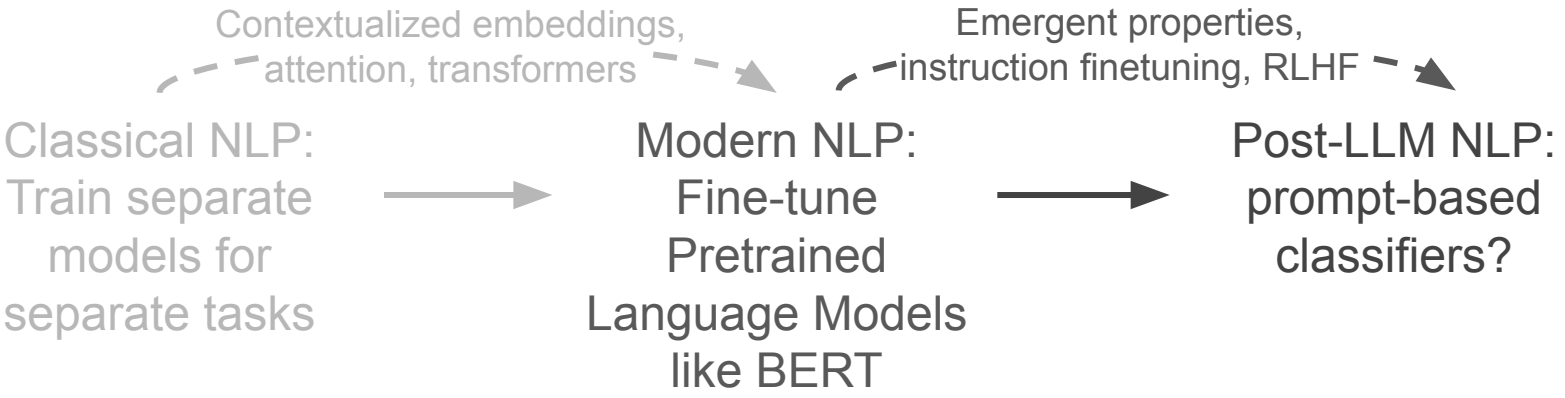
Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

Training language models to follow instructions with human feedback

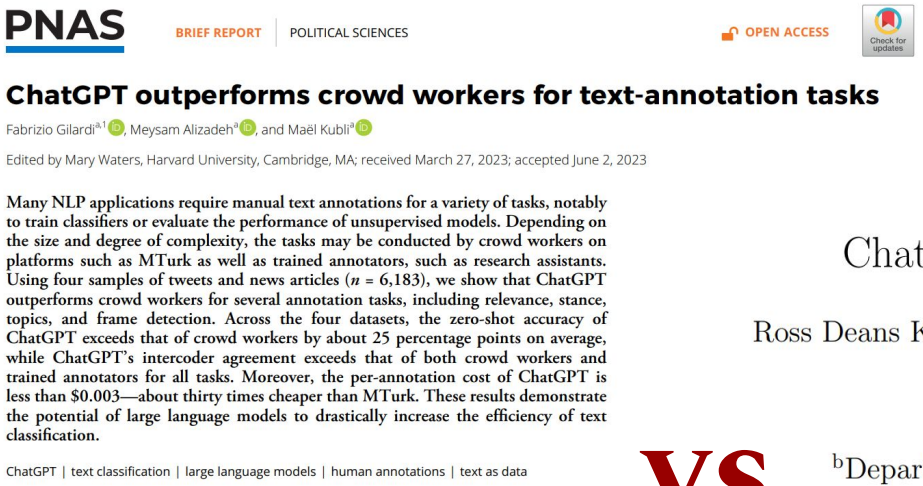
The results

PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3
	`Explain the theory of gravity to a 6 year old.`
	`Explain the theory of relativity to a 6 year old in a few sentences.`
	`Explain the big bang theory to a 6 year old.`
	`Explain evolution to a 6 year old.`
	InstructGPT
	`People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.`

The timeline so far: Supervised NLP



Next week: debating content analysis with LLM



PNAS BRIEF REPORT | POLITICAL SCIENCES OPEN ACCESS

ChatGPT outperforms crowd workers for text-annotation tasks

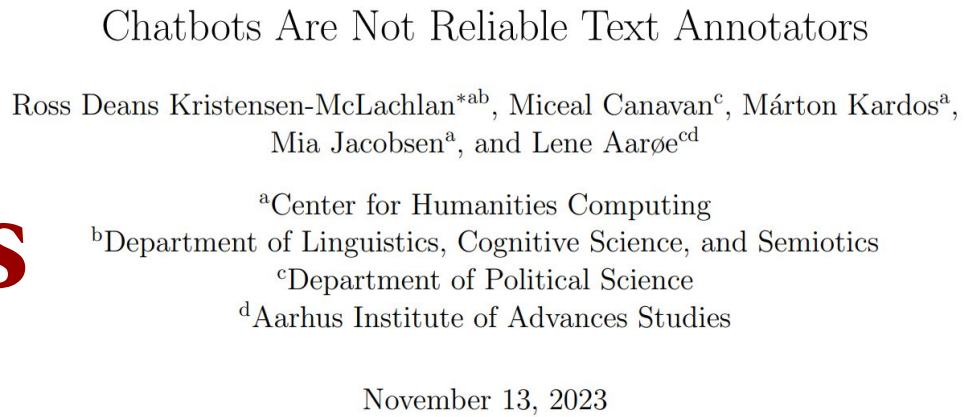
Fabrizio Gilardi[a,1], Meysam Alizadeh[a], and Maël Kubli[a]

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles (n = 6,183), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

VS



Chatbots Are Not Reliable Text Annotators

Ross Deans Kristensen-McLachlan*[ab], Miceal Canavan[c], Márton Kardos[a], Mia Jacobsen[a], and Lene Aarøe[cd]

[a]Center for Humanities Computing
[b]Department of Linguistics, Cognitive Science, and Semiotics
[c]Department of Political Science
[d]Aarhus Institute of Advances Studies

November 13, 2023