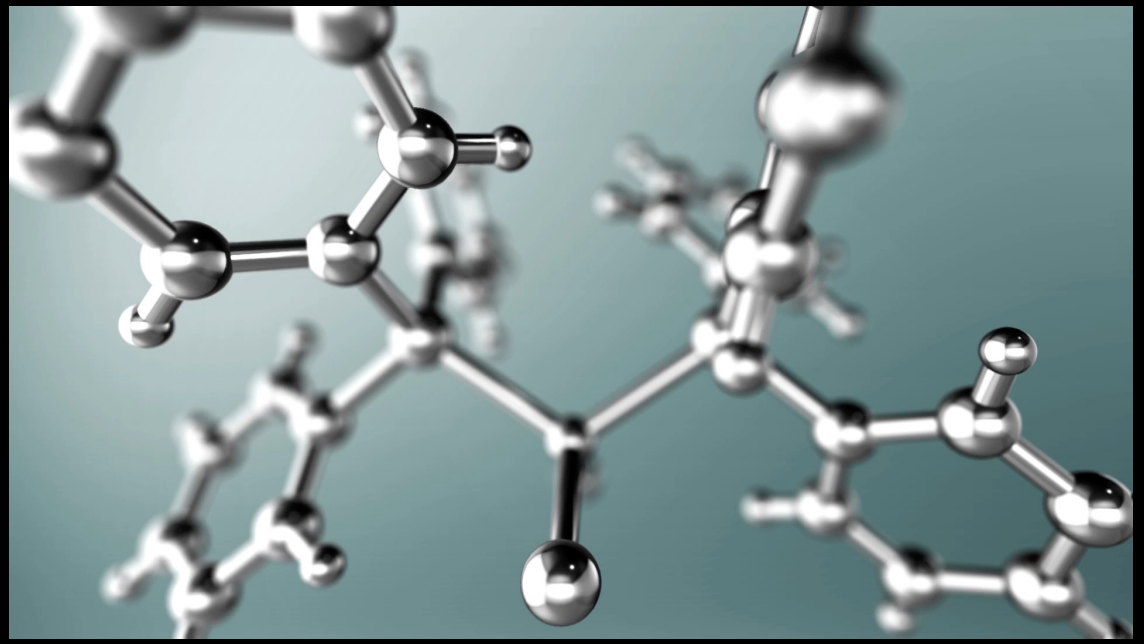
How far can we get in creating a digital replica of a philosopher?



ANNA STRASSER, MATTHEW CROSBY, ERIC SCHWITZGEBEL

DENKWERKSTATT BERLIN

OVERVIEW

1 **MOTIVATION**
Why do philosophers get involved with AI?
Can a language model represent a philosopher?

2 **TECHNIQUE**
What is a GPT-3?
How to fine-tune GPT-3 & use a fine-tuned GPT-3?

3 **LIMITS & RISKS**
Limitations of neuronal networks
Legal and ethical questions

4 **WHAT WE DID**
Fine-tuning | prompt engineering |evaluating

5 **RESULTS**
Distinguishing & evaluating machine-generated output

DENKWERKSTATT BERLIN

MOTIVATION

Why do philosophers get involved with AI?

motivating factors

- understanding how the human mind works
- proving that humans are special
- building new thinking tools
- …

WHENEVER A MACHINE SUCCESSFULLY SOLVES A TASK
– of which we think that humans solve it because of their intelligence –

then *no 'intelligence'* in the true (deep) sense of the word seems to be necessary for solving this task

BUT

Investigating multiple realizations of solving tasks is a good starting point for further philosophical investigations.

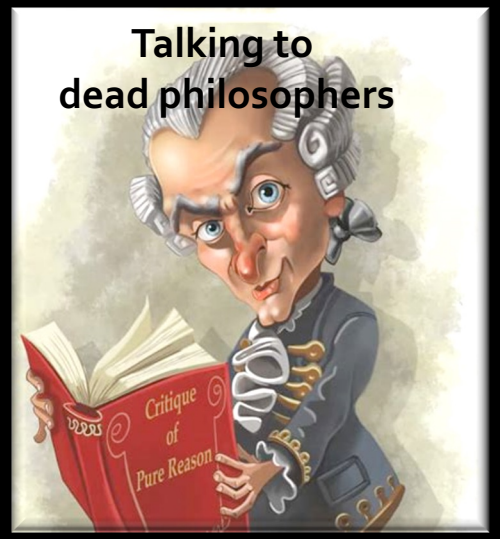
Can we build machines with which we can have interesting conversations?

AMBITIOUS GOAL:

We want to find out how far we can get in creating a digital replica of a philosopher.

→ investigating of how the best model can be built & used

→ exploring the limits and risks which are accompanied by the creation of digital replicas



several pilot experiments with GPT-3's Currie engine

- fine-tuning with Kant's work in English translation
- fine-tuning with a collection of philosophical blog posts

Schwitzgebel, Eric (2021)

What is a GPT-3?

a neural network trained to predict the next likely word in a sequence

Pre-trained

- 499 billion tokens*
  (Common Crawl / WebText / Books / Wikipedia)

Generative

- can generate long sentences
- not just yes or no answers or simple sentences

Transformer

- calculating the probability of the next word appearing surrounded by the other ones

Generative Pretrained Transformer

- a 175 billion parameter language model which shows strong performance on many NLP tasks

*1 token = significant fractions of a word (on average 0,7 words per token)

DENKWERKSTATT BERLIN

Applications using GPT-3

AI|Writer-App

allows people to correspond with replicas of historical figures via e-mail (Mayne A. https://www.aiwriter.app)

Applications useful for code generation

Codex model by GitHub Copilot using GPT-3 to translate conventional language into formal computer code (Langston 2021)

GPT-3 can be used to convert common natural language expressions into legal language and vice versa

Online documentation of Open-AI.

chatbots can be trained to speak in the tone of particular corporations

Online documentation of Open-AI.

For an overview of applications, see https://beta.openai.com/examples

According to an opinion piece in *MIT Technology Review,* GPT-3 is "shockingly good – and completely mindless" (Heaven 2020)

Beware of limits & risks

GPT-3 IS A DEEP LEARNING SYSTEM

- deals only with text
- limited input & output sizes (2048 linguistic tokens, roughly 1500 words)
- lacks any form of memory
  → we cannot expect GPT-3 to succeed in text-related tasks which require a larger amount of context knowledge than can be captured by the limited input size
- problems with reliability & interpretability
  → we cannot presuppose that all outputs of a GPT-3 will be acceptable

→ OFTEN, WE CANNOT PREDICT HOW SUCH A SYSTEM WILL REACT TO NEW INPUT

→ BIASES IN THE DATA LEAD THE MODEL TO GENERATE STEREOTYPED OR PREJUDICED CONTENT

- BE AWARE OF BIASES & OFFENSIVE LANGUAGE OF THE TRAINING DATA
- REMAIN SKEPTICAL BECAUSE OUTPUTS CAN BE SUBTLY FLAWED OR UNTRUE

↓

NEVER USE THEM IN CONTEXTS WHERE AN INCORRECT OUTPUT IS ETHICALLY QUESTIONABLE

ABOUT LIMITS & RISKS
Bommasani et al. (2021), Floridi & Chiriatti (2020)

More Limitations

Discrete language tasks

- notable gaps in reading comprehension & comparison tasks

Text synthesis

- starts repeating itself

Tasks that empirically benefit from bidirectionally

- fill-in-the-blank tasks

tasks presuppose looking back and comparing two pieces of content, tasks that require re-reading or carefully considering a long passage

Expensive & inconvenient

- very computing-power hungry model
- become "overqualified" for some specific tasks

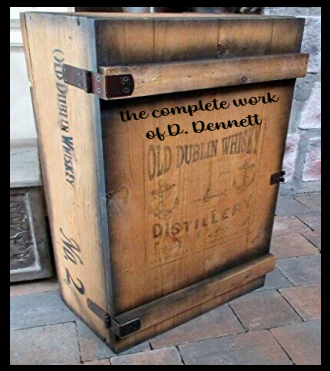
ANALYZING ALL KINDS OF SHORT-COMINGS MAY TELL US SOMETHING ABOUT OURSELVES.

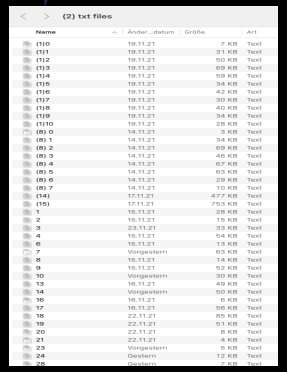
MY DIRECT OBJECTIVE IS NOT TO DEVELOP A SUPER-INTELLIGENT MACHINE.

What we did

TRAINING DATA:
13 books & 269 articles / interviews

EDITING





txt

max. 1024 tokens

{"prompt":"", "completion":" <paragraph of text of 1-n.txt>"}
{"prompt":"", "completion":" <paragraph of text of 1-n.txt>"}
{"prompt":"", "completion":" <paragraph of text of 1-n.txt>"}

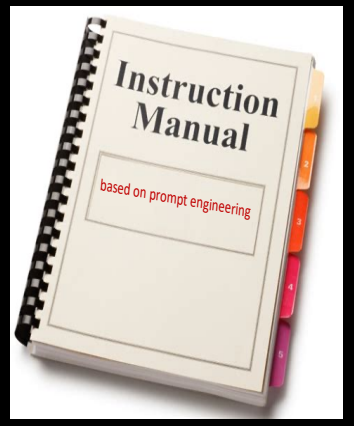
jasonl training data

FINE-TUNING

1828 PROMPTS → 3,275,000 TOKENS

FORMAT OF OUR PROMPTS
Interviewer: [text of the question]
Dennett:

PROMPT ENGINEERING



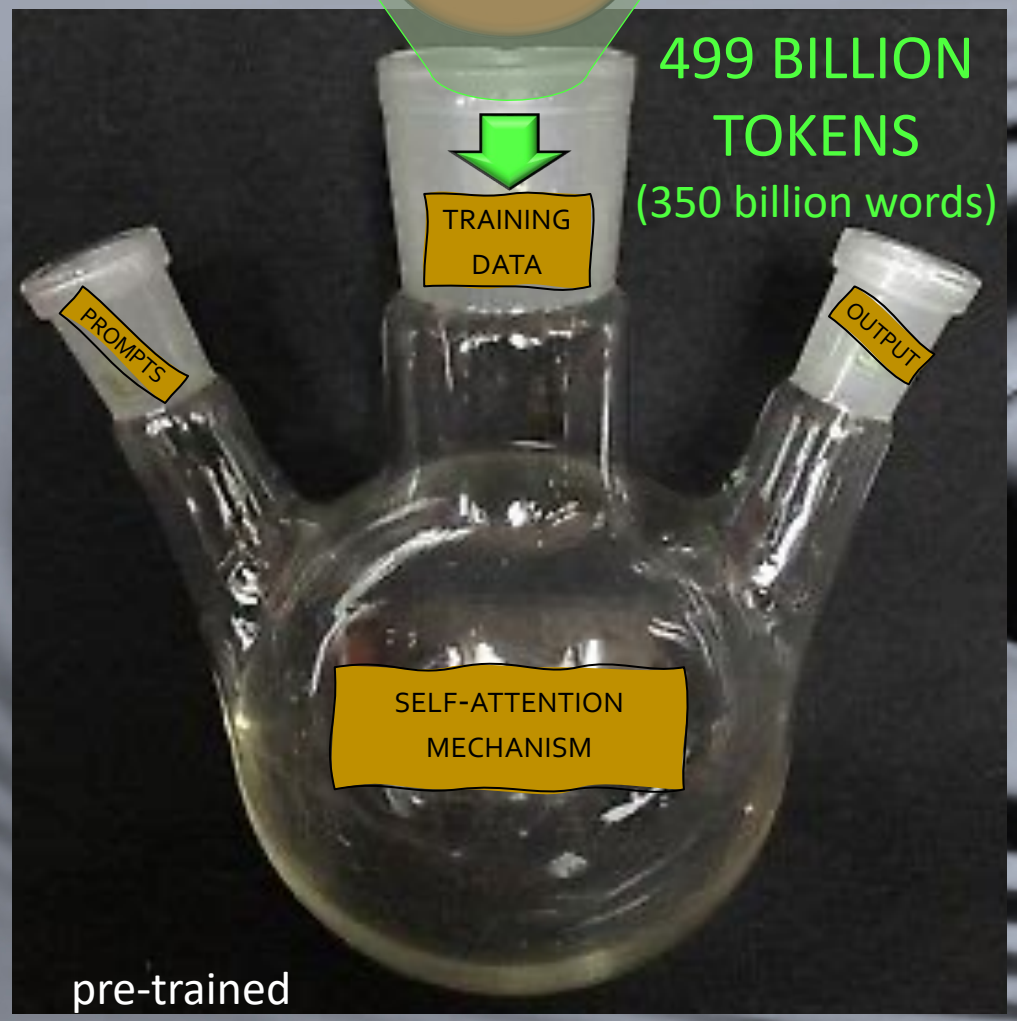
OPEN ENDED GENERATION

- leave the prompt empty
- at least a few thousand examples

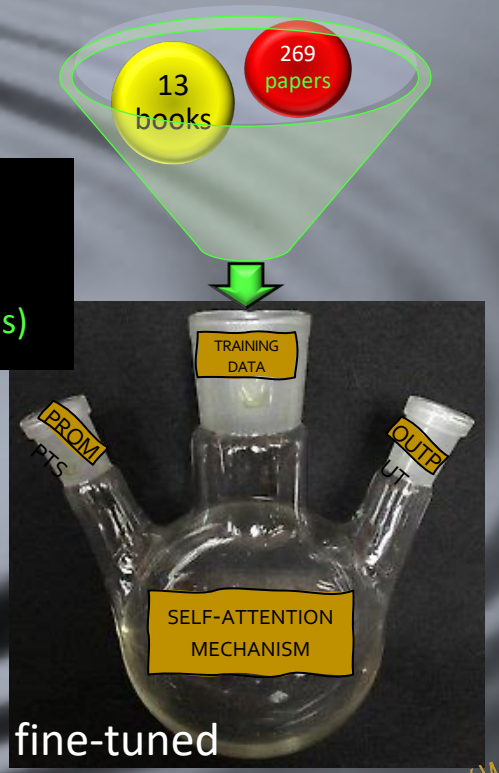
Anna Strasser

DENKWERKSTATT BERLIN

JUST A DROP IN THE OCEAN



3 MILLION TOKENS
(2,5 million words)



Anna Strasser