

# DIGIDAN

weizenbaum  
institut



Weizenbaum, ELIZA, and  
(Chat)GPT

Session Chair:  
**Bettina Berendt**  
Weizenbaum Institute and TU Berlin

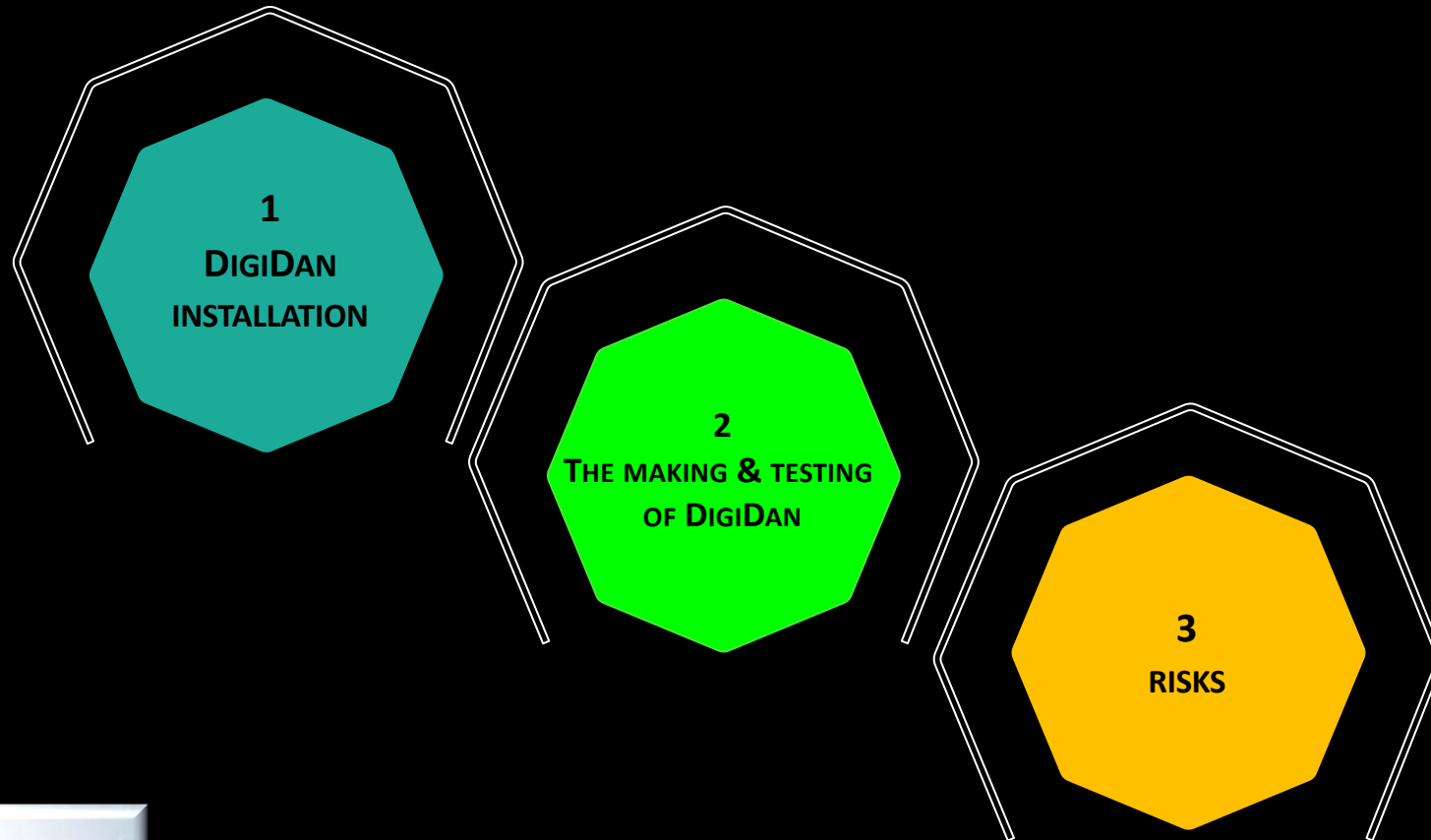
Anna Strasser, Pieter Delobelle, Thomas Winters

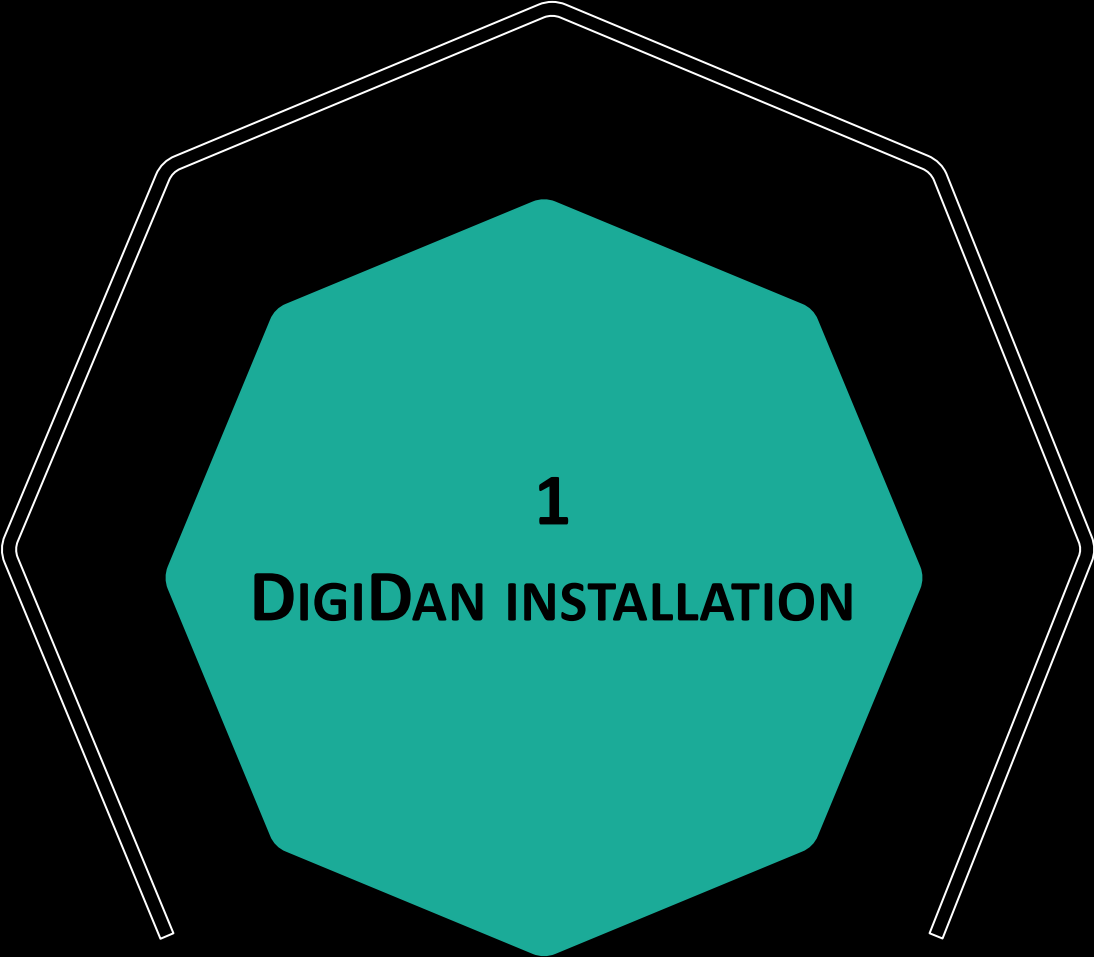
*Anna Strasser*



# Overview

slides can be downloaded at  
[https://www.denkwerkstatt.berlin/  
ANNA-STRASSER/TALKS/](https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS/)





**1**  
**DIGIDAN INSTALLATION**

# AI CAN OUTPERFORM EVEN EXPERT HUMANS IN MANY DOMAINS

notable successes in many domains

- chess, go, discovering novel algorithms, protein folding (Deep Blue, AlphaGo, AlphaTensor, AlphaFold)
- automatic translation (DeepL), lipreading (LipNet)
- computer code generation (Github Copilot),
- producing original prose with fluency equivalent to that of a human (LLMs)

Campbell 2002; Silver et al. 2016, 2018; Ardila et al. 2019; Brown & Sandholm 2019; Jumper, Evans, & Pritzel et al. 2021; Fawzi et al. 2022; Assael et al. 2016; Steven & Izhev 2022

## IS PHILOSOPHY SAFE FROM AI TAKEOVER?

Will machines ever generate essays that survive the refereeing process at *Philosophical Review*?

How close can we get to creating an AI that can produce novel and seemingly intelligent philosophical texts?

DigiDan

**WE CREATED A LANGUAGE MODEL OF DANIEL DENNETT SUFFICIENTLY GOOD THAT EXPERTS IN DENNETT'S WORK COULD NOT RELIABLY DISTINGUISH PARAGRAPHS WRITTEN BY DENNETT FROM THOSE WRITTEN BY THE LANGUAGE MODEL.**



# *DigiDan installation*

## VIDEO

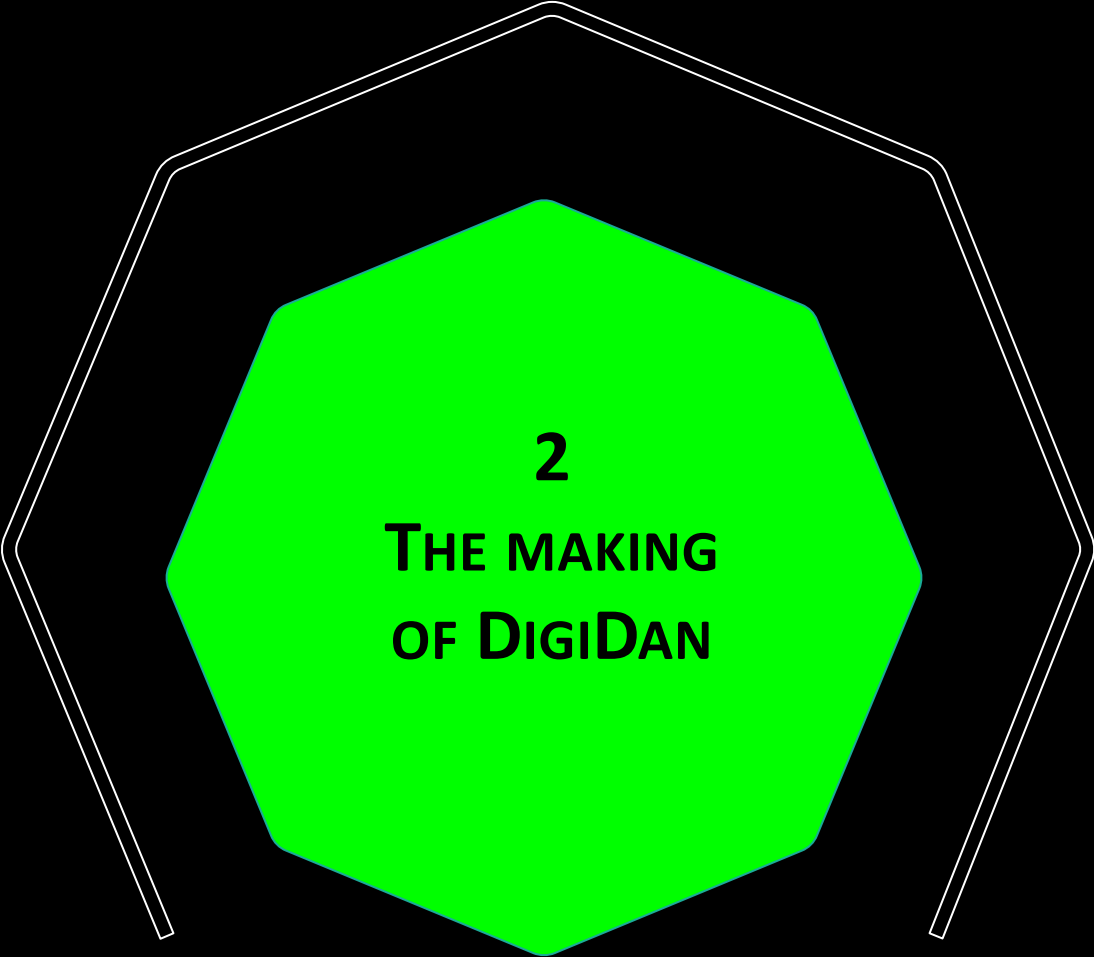
At the end of the video, you will be asked who was the real human Daniel Dennett.

You can participate in a poll by using the QR-code.

Or go to the SLIDO App and type **#2395819**

<https://app.sli.do/event/jpQ8fe1ejM57u9e9HDqErP>





**2**  
**THE MAKING  
OF DIGIDAN**

# LOOKING BACK

I asked myself whether I was happy that I got involved in this project.



YES

**'TALKING' TO PHILOSOPHERS IS MUCH MORE ATTRACTIVE THAN ONLY INTERPRETING THEIR TEXTUAL OUTPUTS**

*An anecdote:*

The night before my oral examination on Kant, I had this dream, in which I found myself discussing with Kant and even convincing him of something.

- good dose of self-confidence for my exam
- influenced my further work in philosophy by making me develop a strong preference to deal more with living than with already deceased philosophers



NO

**BLURRING THE DIFFERENCE BETWEEN HUMANS & MACHINES**

I hope no language model will ever be trained with all my statements.

- I do not aim to be mistaken for such a model.
- I do not aim to have such a digital legacy\* continuing to make statements on my behalf after my death.

# GPT-3 IS A LARGE LANGUAGE MODEL

a neural network trained to predict the likely next word

**P**re-trained

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

**G**enerative

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

**T**ransformer



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

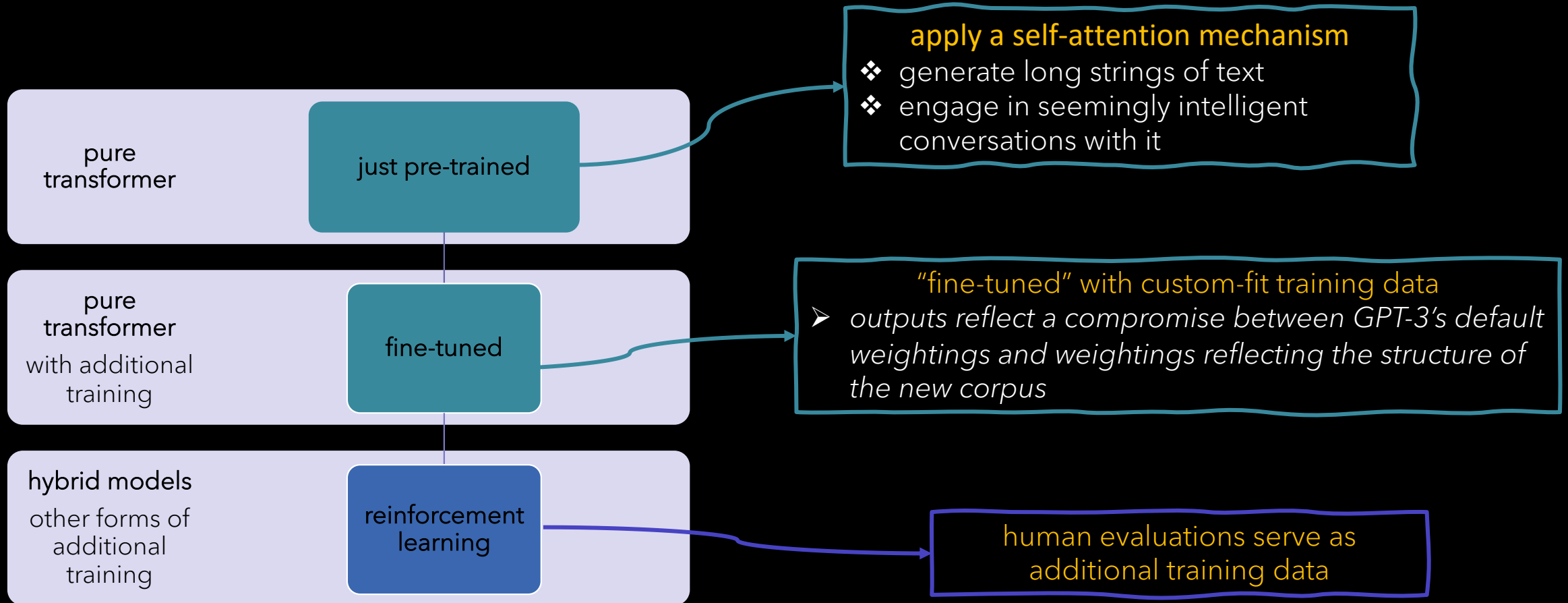
**Generative Pretrained Transformer**

- a 96-layer, 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)



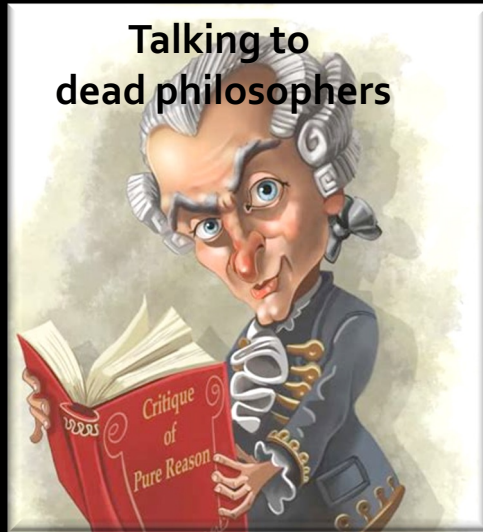
# OTHER LARGE LANGUAGE MODELS



# Piloting



fine-tuning LLMs with Kant's work in English translation



fine-tuning with a collection of philosophical blog posts

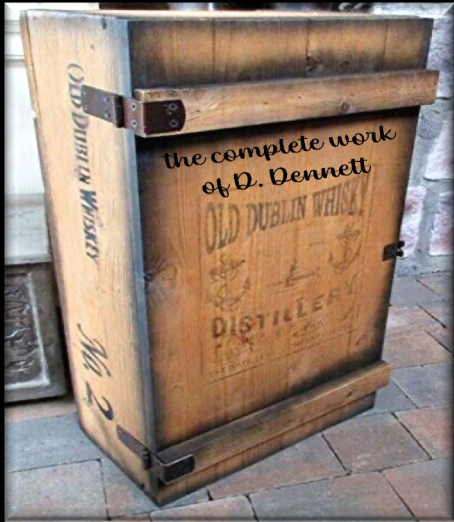


<https://schwitzsplinters.blogspot.com/2021/11/two-robot-generated-splintered-mind.html>



# Editing & fine-tuning

## PREPARING TRAINING DATA



Dennett's corpus

Name	Anders...	Gedä...	Art
(110)	19.11.21	7 KB	Text
(111)	19.11.21	21 KB	Text
(112)	19.11.21	50 KB	Text
(113)	19.11.21	69 KB	Text
(114)	19.11.21	59 KB	Text
(115)	19.11.21	24 KB	Text
(116)	19.11.21	42 KB	Text
(117)	19.11.21	20 KB	Text
(118)	19.11.21	40 KB	Text
(119)	19.11.21	24 KB	Text
(110)	19.11.21	28 KB	Text
(109)	19.11.21	8 KB	Text
(01 0)	14.11.21	34 KB	Text
(01 1)	14.11.21	49 KB	Text
(01 2)	14.11.21	49 KB	Text
(01 3)	14.11.21	40 KB	Text
(01 4)	14.11.21	67 KB	Text
(01 5)	14.11.21	63 KB	Text
(01 6)	14.11.21	29 KB	Text
(01 7)	14.11.21	10 KB	Text
(143)	17.11.21	477 KB	Text
(15)	17.11.21	753 KB	Text
1	16.11.21	28 KB	Text
2	15.11.21	10 KB	Text
3	23.11.21	28 KB	Text
4	15.11.21	94 KB	Text
6	16.11.21	18 KB	Text
7	Vorgestern	63 KB	Text
8	16.11.21	14 KB	Text
9	15.11.21	52 KB	Text
10	Vorgestern	20 KB	Text
13	16.11.21	49 KB	Text
14	Vorgestern	50 KB	Text
16	16.11.21	6 KB	Text
17	16.11.21	20 KB	Text
18	22.11.21	80 KB	Text
19	22.11.21	91 KB	Text
20	22.11.21	6 KB	Text
21	22.11.21	4 KB	Text
23	Vorgestern	9 KB	Text
24	Gestern	17 KB	Text
25			

converted into plain text format

- stripping away headers, footnotes, scanning errors, marginalia, and other distractions

jsonl training data

Dinner is ready!  
Today we serve three million tokens

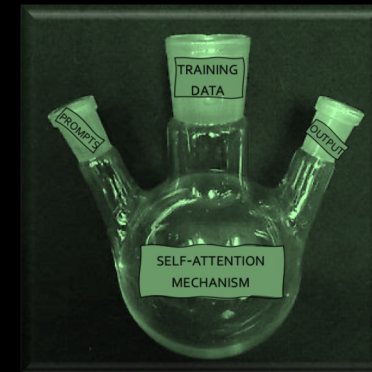
15 BOOKS  
269 ARTICLES



BLANK PROMPTS

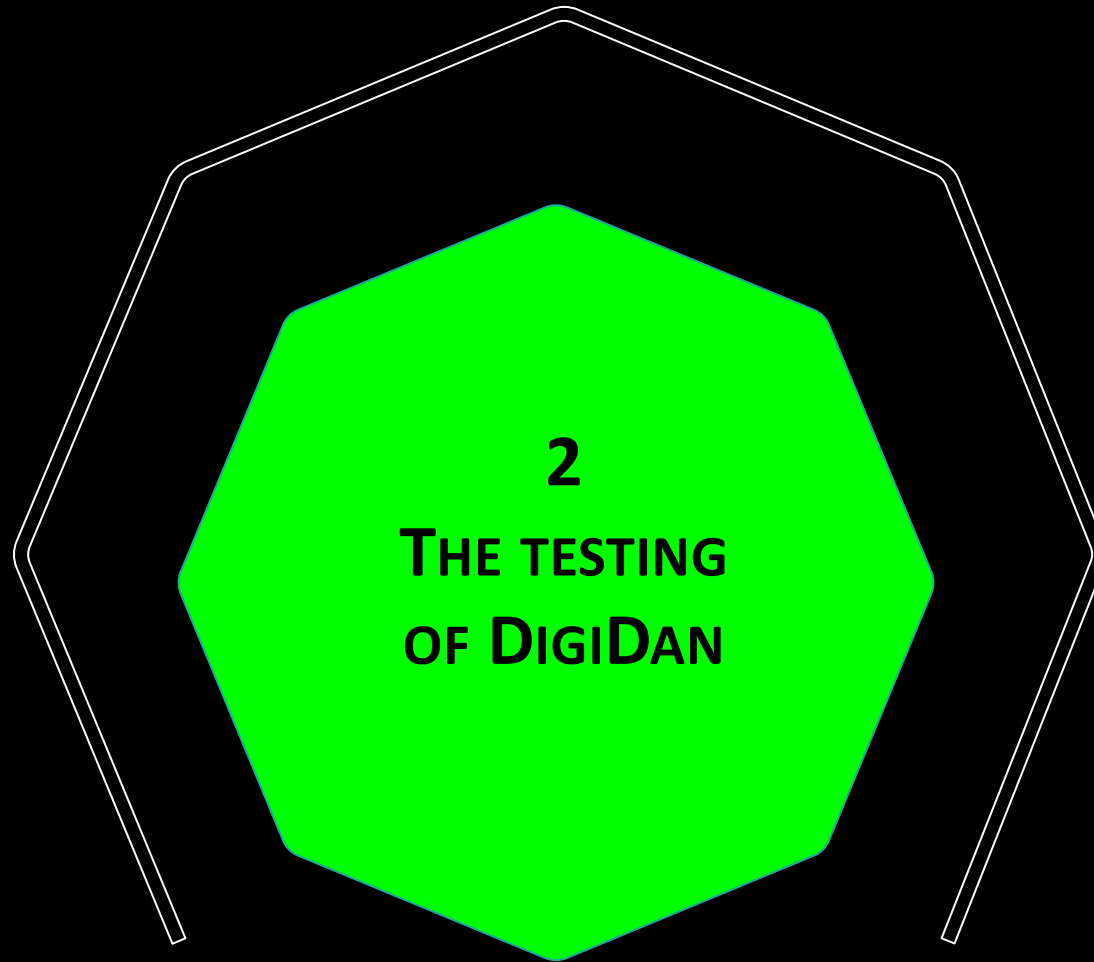
SEGMENTS OF TRAINING DATA (<2000 TOKENS)

1. {"prompt":"","completion":" <paragraph of text of 1-n.txt>"}
2. {"prompt":"","completion":" <paragraph of text of 1-n.txt>"}
3. {"prompt":"","completion":" <paragraph of text of 1-n.txt>"}
- ...
- ...
- ...
1826. {"prompt":"","completion":" <paragraph of text of 1-n.txt>"}
1827. {"prompt":"","completion":" <paragraph of text of 1-n.txt>"}
1828. {"prompt":"","completion":" <paragraph of text of 1-n.txt>"}



## FINE-TUNING THE GPT-3 DAVINCI ENGINE

- open-ended generation
- leave the prompt empty
- at least a few thousand examples
- repeating the process four times



**2**  
**THE TESTING  
OF DIGIDAN**

Schwitzgebel, Eric, Schwitzgebel, David, Strasser, Anna (2023). Creating a Large Language Model of a Philosopher. *Mind & Language*.

preprint at  
<https://arxiv.org/abs/2302.01339>



# Testing the machine

HOW EASILY CAN THE OUTPUTS OF THE FINE-TUNED GPT-3 BE DISTINGUISHED FROM DENNETT'S REAL ANSWERS?

We asked Dennett ten philosophical questions.

- Dennett provided us with sincere written answers, ranging in length from 41 to 124 words

We posed those same questions to our fine-tuned version of GPT-3.

- four responses for each of the ten questions

We recruited experts in Dennett's work, blog readers, and ordinary online research participants into an experiment in which they attempted to distinguish Dennett's real answers from the answers generated by GPT-3.

## Hypotheses

**EXPERT RESPONDENTS WILL PERFORM BETTER THAN ORDINARY RESEARCH PARTICIPANTS**

**EXPERT RESPONDENTS WILL ON AVERAGE GUESS CORRECTLY AT LEAST 80% OF THE TIME**

**EXPERT RESPONDENTS WILL RATE DENNETT'S ACTUAL ANSWERS AS MORE DENNETT-LIKE THAN GPT-3'S ANSWERS**

# Prompt engineering

GPT-3 COMPLETIONS ARE HIGHLY SENSITIVE TO THE CONTENT AND STRUCTURE OF THE PROMPTS  
GOOD "PROMPT ENGINEERING" IS IMPORTANT FOR COAXING USEFUL REPLIES FROM GPT-3

... we settled on the following simple prompt:

```
Interviewer: [text of question]  
Dennett:
```

*This simple prompt has several advantages:*

minimal structure reduces potential concerns about the prompt possibly nudging completions toward specific philosophical content, as a more substantive prompt might

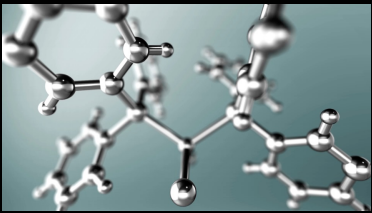
encourages GPT-3 to speak in the first person, voicing Dennett's views, rather than speaking in the third person about Dennett

simple format makes it easily generalizable to other cases

# Collecting & editing GPT-3's responses

We gathered completions in the GPT-3 playground using our prompt

- OpenAI's default settings: *temperature = 0.7, top P = 1, frequency penalty = 0, presence penalty = 0, best of = 1*



perceived quality of response was never used as a basis for selection  
→ no "cherry picking" of responses that we judged to be better, more Dennett-like, or more likely to fool participants

WE RE-INPUTTED THE PROMPT UNTIL THE COMPLETION MET 2 CRITERIA

1. LENGTH: comparable length with Dan's answer
2. AVOIDING OBVIOUS CUES
  - excluding outputs that contained the words "Interviewer" or "Dennett"
  - regularizing curly quotes to straight quotes, single quotes to double quotes, and dashes to m-dashes

# Guessing task & Evaluation of the likeliness

1

We posed the question below to Daniel C. Dennett and also to a computer program that we trained on samples of Dennett's works. One of the answers below is the actual answer given by Dennett. The other four answers were generated by the computer program. We'd like you to guess: which one of the answers was given by Dennett?

Question:

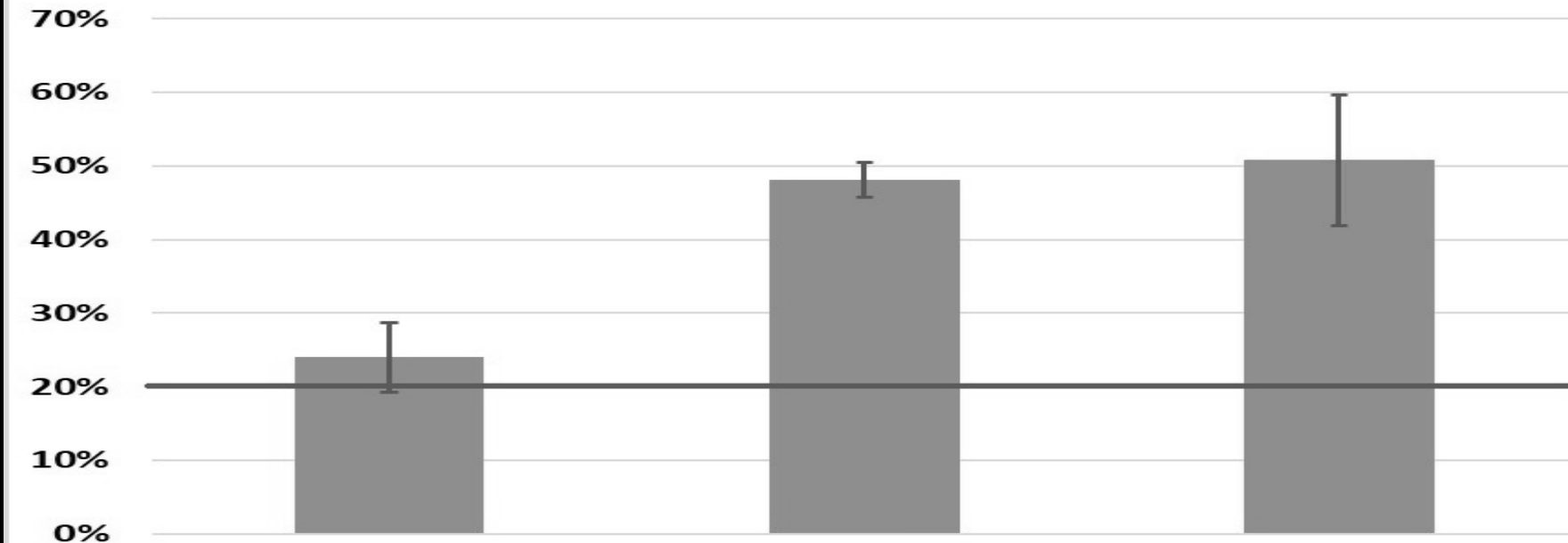
2

Participants were instructed to rate each answer (Dennett's plus the four from GPT-3) on the following five-point scale:

- "not at all like what Dennett might say" (1)
- "a little like what Dennett might say" (2)
- "somewhat like what Dennett might say" (3)
- "a lot like what Dennett might say" (4)
- "exactly like what Dennett might say" (5)



## Success Rate in Distinguishing Dennett from a GPT-3 Model Fine-Tuned on Dennett's Works (chance = 20%)



	Ordinary Research Participants	Blog Readers	Dennett Experts
majority	with no classes in philosophy & no familiarity with Dennett's work	with graduate degrees in philosophy & familiarity with Dennett's work	reported having read over 1000 pages of Dennett's work
correctly guessed	1.20 times out of 5 <ul style="list-style-type: none"> <li>• 86% 1-2 correct</li> <li>• 14% 3-4 correct</li> </ul>	4.81 times out of 10 (48%)	5.08 times out of 10 (51%)
given a five-alternative forced choice	<ul style="list-style-type: none"> <li>• near chance rate of 20%</li> </ul>	<ul style="list-style-type: none"> <li>• substantially above chance</li> </ul>	

# Hypotheses



**EXPERT RESPONDENTS WILL PERFORM BETTER THAN ORDINARY RESEARCH PARTICIPANTS**

significantly below the hypothesized accuracy of 80%

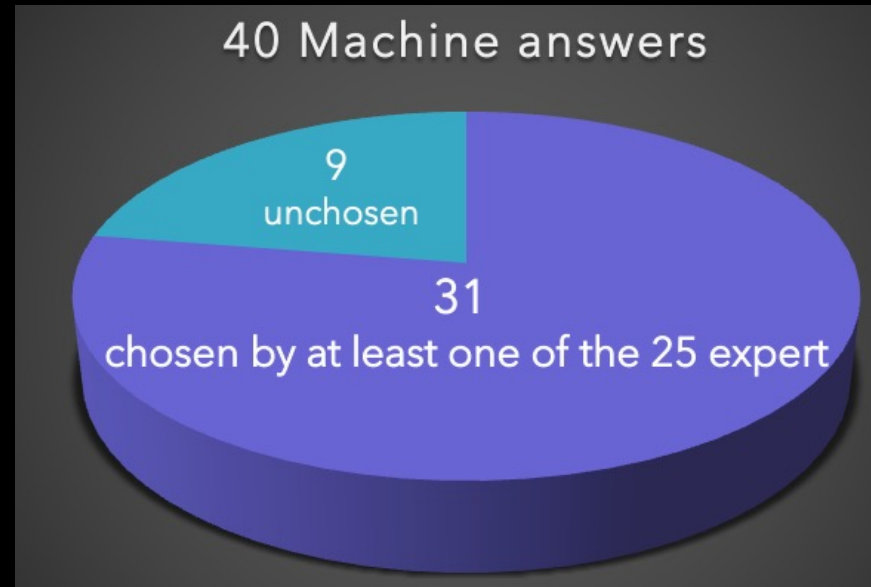


**EXPERT RESPONDENTS WILL ON AVERAGE GUESS CORRECTLY AT LEAST 80% OF THE TIME**



**EXPERT RESPONDENTS WILL RATE DENNETT'S ACTUAL ANSWERS AS MORE DENNETT-LIKE THAN GPT-3'S ANSWERS**

# Performance of the machine answers



- ❖ not at all like what Dennett would say
- ❖ representing a significant failure of the fine-tuning project to reliably represent Dennett's views

# Was DigiDan overtrained?

DOES THE MACHINE SIMPLY PARROT SENTENCES OR MULTI-WORD STRINGS OF TEXTS FROM DENNETT'S CORPUS?

## Turnitin plagiarism checker

- check for "plagiarism" between machine outputs & the Turnitin corpus supplemented with the training data
  - 5% overall similarity between machine answers and the comparison corpora
  - none of the passages were marked as similar to the training corpus we used in fine-tuning

## ngram package from the R programming language (Schmidt & Heckendorf 2015)

- looked for strings of 6 or more words that matched between the 3240 words of machine answers & approximately two million words of Dennett's corpus
  - strings defined as contiguous "6-grams," "7-grams," etc., with matching cases sharing the same order of six (or more) words

**WE FOUND only 21 MATCHING STRINGS**



BEFORE AIMING FOR FURTHER FINE-TUNED LLMs

RISKS SHOULD BE EVALUATED

Strasser, Anna (2023). On pitfalls (and advantages) of sophisticated Large Language Models.

preprint at  
<https://arxiv.org/abs/2303.17511>





# Copyright

## Copyright law governing fine-tuned language models is not yet settled

- unclear whether it is fair use of intellectual property to fine-tune a language model on the works of a single author
  - idea-borrowing via fine-tuned language models might be undetectable as plagiarism, even if it is rightly considered plagiarism
  - fine-tuned models will not output a long sequence of text that exactly matches a sequence of text from the author's corpus
- until the law is settled
  - WE RECOMMEND SEEKING THE EXPLICIT PERMISSION OF THE AUTHOR BEFORE FINE-TUNING & PUBLISHING ANY OF THE OUTPUTS**
- open question
  - How to deal with works by deceased authors? (Nakagawa & Orita 2022)



# Overreliance

NOT GOOD ENOUGH!

DigiDan did not reliably produce outputs representing Dennett's views.

not surprising:

- all deep learning networks have problems with reliability

(Alshemali & Kalita 2020; Bosio et al. 2019)

- user might mistakenly assume that outputs are likely to reflect the actual views of the author
  - tempting for students, social media users, or others who might rather query a fine-tuned model of an author than read the author's work

I RECOMMEND SUBSTANTIAL CAUTION

BEFORE RELEASING TO THE PUBLIC ANY LANGUAGE MODELS FINE-TUNED ON AN INDIVIDUAL AUTHOR





# Counterfeiting

Language models should be clearly described as such, their limitations should be noted, and all outputs should be explicitly flagged as the outputs of a computer program rather than a person.

**If machine-generated text were presented as a quotation or paraphrase of positions of existing persons, this would arguably constitute counterfeiting**



Dennett as interviewed in Cukier 2022





# *Increasingly difficult to distinguish*



How can teachers in the future ensure that submitted essays are not simply a product of an LLM?

- Perhaps universities will return to supervised essay writing in person.

**How to deal with verifiable authorship with respect to the mass of electronically distributed texts?**

- How can we trust the content of websites?
- How can we know whether in chat conversations we are interacting with humans and not with chat-bots?
- How can we trust in video calls?

**Will we establish new social practices that aim at proving that one is really the original author of what is written or said?**

# *Long-term potentials*

## IS PHILOSOPHY SAFE FROM AI TAKEOVER?

What do you think about

- computer programs that generate music in the style of a particular composer
- image-generation programs
- language-models that generate text on behalf of the user in other domains

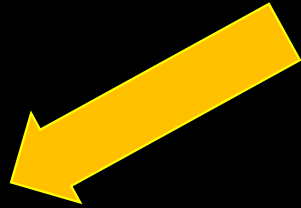
### **LANGUAGE MODEL as THINKING TOOL USED BY HUMANS**

- future fine-tuned language models might produce outputs interesting enough to serve as a valuable source of cherry-picking by experts
  - selected outputs might have substantial merit
- an author might create many outputs, choose the most promising, edit them lightly, and present them, not unreasonably, as original work

# Take home message

DigiDan can sometimes give outputs indistinguishable by experts from Dennett's outputs.

BUT neural networks are not reliable;  
they're not like calculators, which always generate the same correct answer.



## Fine-tuned language models can create opportunities for plagiarism, over-interpretation, and over-reliance

- Our efforts to make sense of anything that looks roughly interpretable can betray us!
- GPT-3 can serve as an automatic plagiarist → dangerous prospect of this technology because copyright doesn't come close to dealing with it!

## RECOMMENDATIONS

- We need legislation to outlaw some of the ways in which these systems might be used!
- We should always ask for permission if we build a model based on a living person!

❖ But it could also be a helpful thinking tool ?

# Acknowledgements

This could have not happened without Eric & David Schwitzgebel!

Special thanks to both Daniel C. Dennett & Matthew Crosby!

- Dennett provided cooperation, advice, and encouragement in all aspects of this project.
- Matthew Crosby provided technical expertise and implemented the fine-tunings for this project, as well as collaborating on a conceptual paper that provided the groundwork for this project (Strasser, Crosby, and Schwitzgebel 2023)



- Alshemali, B., & Kalita, J. (2020). Improving the reliability of deep neural networks in NLP: A review. *Knowledge-Based Systems*, 191, 105210. <https://doi.org/10.1016/j.knosys.2019.105210>
- Ardila, D. et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6), 954–961. <https://doi.org/10.1038/s41591-019-0447-x>
- Assael, Y., Shillingford, B., Whiteson, S., & Freitas, N. (2016). LipNet: Sentence-level lipreading. <https://doi.org/10.48550/arXiv.1611.01599>
- Bosio, A., Bernardi, P., Ruospo, A., & Sanchez, E. (2019). A reliability analysis of a deep neural network. *2019 IEEE Latin American Test Symposium (LATS)*, pp. 1–6. <https://doi.org/10.1109/LATW.2019.8704548>
- Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science*, 365(6456), 885–890. <https://doi.org/10.1126/science.aay2400>
- Campbell, M., Hoane, A. J., Jr., & Hsu, F. H. (2002). Deep blue. *Artificial Intelligence*, 134(1–2), 57–83.
- Cukier, K. (2022). Babbage: Could artificial intelligence become sentient? *The Economist*. <https://shows.acast.com/theeconomistbabbage/episodes/babbage-could-artificial-intelligence-become-sentient>
- Dennett, d. (2023). The problem with counterfeit people. *The Atlantic*. <https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075>
- Fawzi, A., Balog, M., Romera-Paredes, B., Hassabis, D., & Kohli, P. (2022). Discovering novel algorithms with AlphaTensor. <https://www.deepmind.com/blog/discovering-novel-algorithms-with-alphatensor>
- Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Nakagawa, H., & Orita, A. (2022). Using deceased people's personal data. *AI & Society*, 1–19.
- Schwitzgebel, E. (2021). Two robot-generated splintered mind posts. Blog Post at the Splintered Mind. <https://schwitzsplinters.blogspot.com/2021/11/two-robot-generated-splintered-mind.html>
- Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. *Mind & Language*, 1–22. <https://doi.org/10.1111/mila.12466> [preprint at <https://arxiv.org/abs/2302.01339>]
- Silver, D. et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>
- Silver, D et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419), 1140–1144. <https://doi.org/10.1126/science.aar6404>
- Steven, J., & Iziev, N. (2022). AI is mastering language. Should we trust what it says? *The New York Times*. <https://www.nytimes.com/2022/04/15/magazine/ai-language.html>
- Strasser, A. (2023). On pitfalls (and advantages) of sophisticated Large Language Models. preprint at <https://arxiv.org/abs/2303.17511>
- Strasser, A., Crosby, M., & Schwitzgebel, E. (2023). How far can we get in creating a digital replica of a philosopher? In R. Hakli, P. Mäkelä, & J. Seibt (Eds.), *Social robots in social institutions. Proceedings of Robophilosophy 2022* (pp. 371–380). IOS Press.