Large Language Models:
How they work and what they are good for

Ondřej Dušek

Challenges of AI in Teaching Foreign Languages
Czech University of Life Sciences Prague
3.5.2024

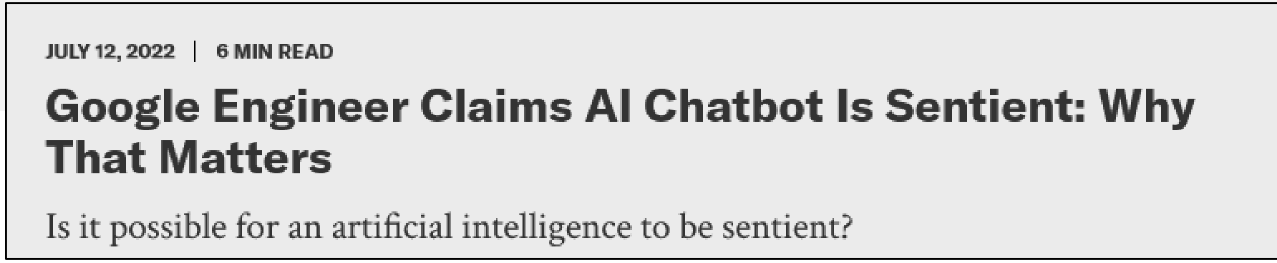
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

CC BY SA
unless otherwise stated

The Hype

- Lot of hype around LLMs right now
  - fed by mainstream media & some “AI” personalities
- AI companies have a lot of incentive to up the hype & downplay problems
  - lots of investment
  - huge expectations
- People using LLMs where they’re not fit for purpose

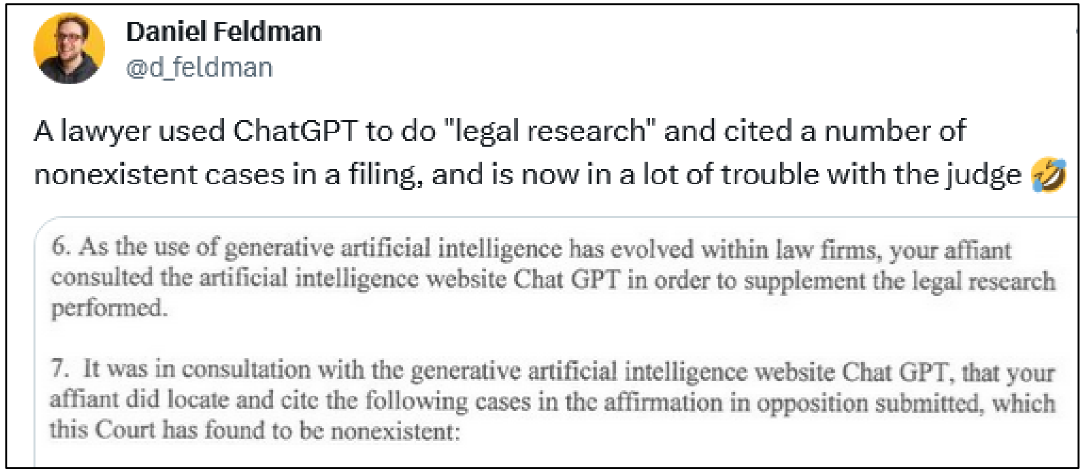
https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/



JULY 12, 2022 | 6 MIN READ

Google Engineer Claims AI Chatbot Is Sentient: Why That Matters

Is it possible for an artificial intelligence to be sentient?



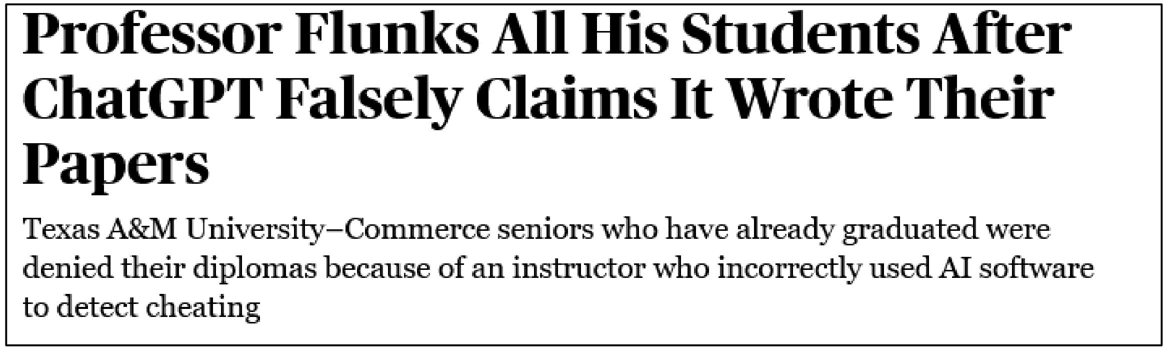
Daniel Feldman
@d_feldman

A lawyer used ChatGPT to do "legal research" and cited a number of nonexistent cases in a filing, and is now in a lot of trouble with the judge 🤣

6. As the use of generative artificial intelligence has evolved within law firms, your affiant consulted the artificial intelligence website Chat GPT in order to supplement the legal research performed.

7. It was in consultation with the generative artificial intelligence website Chat GPT, that your affiant did locate and cite the following cases in the affirmation in opposition submitted, which this Court has found to be nonexistent:

https://twitter.com/d_feldman/status/1662308313525100546



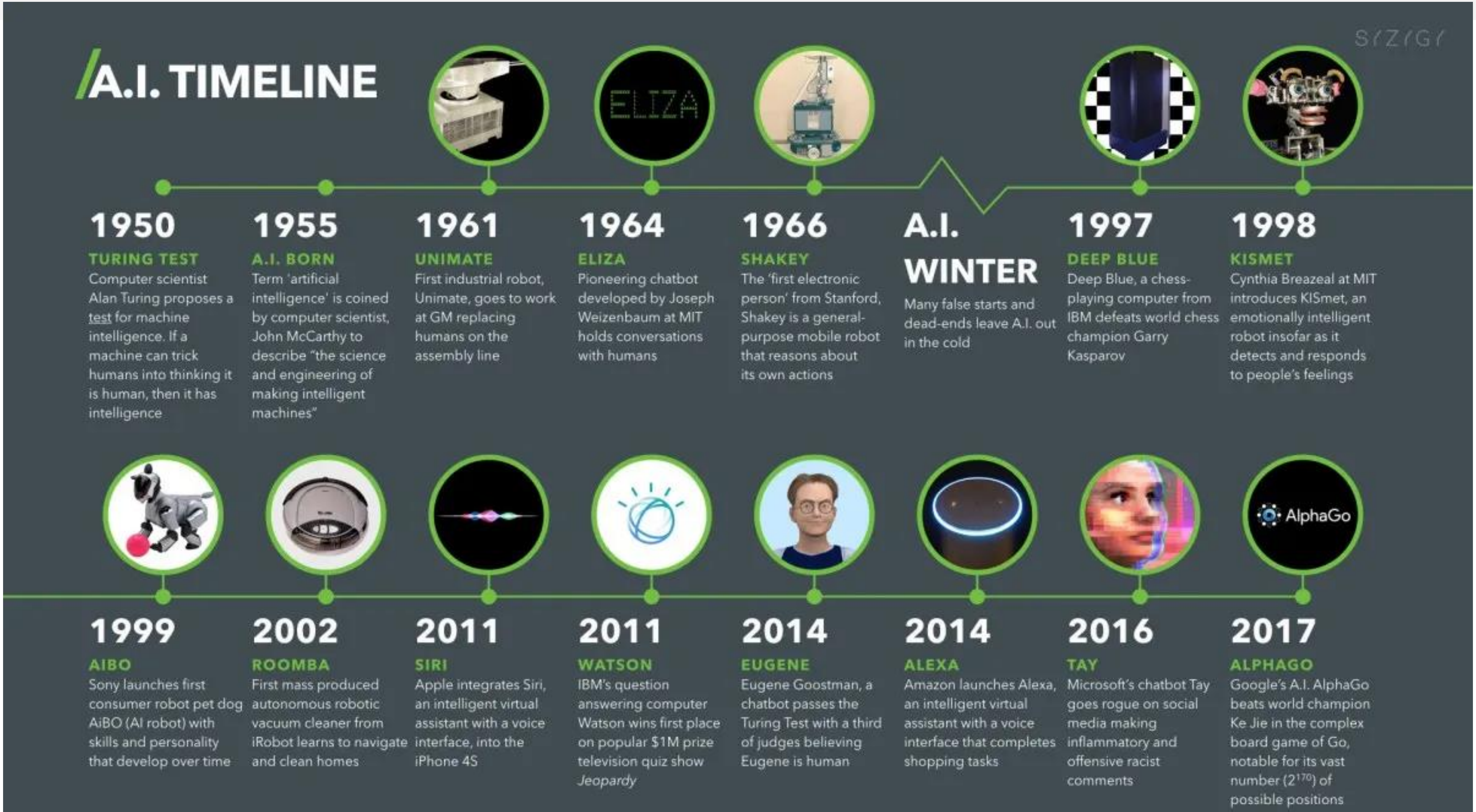
Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers

Texas A&M University–Commerce seniors who have already graduated were denied their diplomas because of an instructor who incorrectly used AI software to detect cheating

https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/

AI is old

https://suryacreatx.medium.com/ai-through-the-ages-unveiling-the-unforgettable-milestones-in-history-92f6588e73d


ELIZA
AlphaGo

A.I. TIMELINE

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955
A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966
SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998
KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999
AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002
ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011
WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

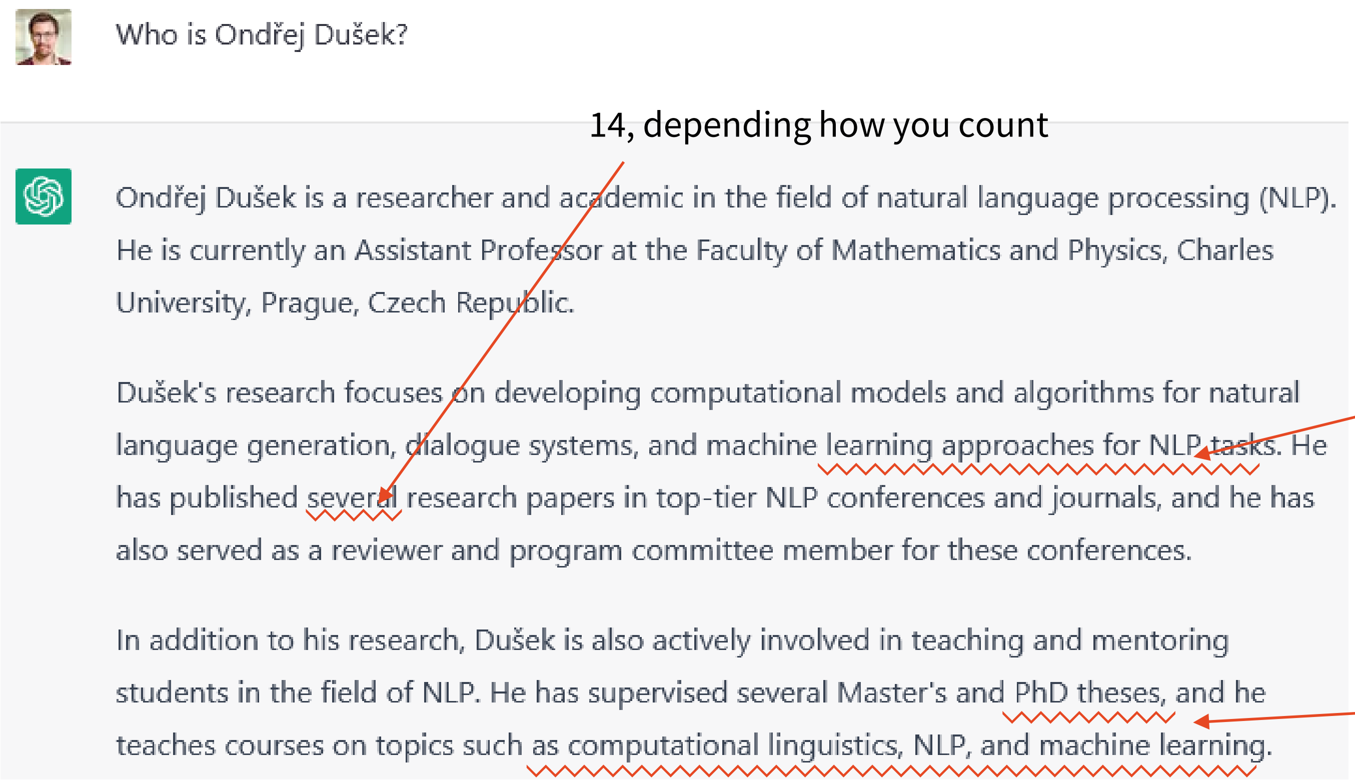
2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016
TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017
ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

LLMs as question answering

- Convincing but not necessarily true!
- Only uses information it memorized, “hallucinates”



Who is Ondřej Dušek?

Ondřej Dušek is a researcher and academic in the field of natural language processing (NLP). He is currently an Assistant Professor at the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Dušek's research focuses on developing computational models and algorithms for natural language generation, dialogue systems, and machine learning approaches for NLP tasks. He has published several research papers in top-tier NLP conferences and journals, and he has also served as a reviewer and program committee member for these conferences.

In addition to his research, Dušek is also actively involved in teaching and mentoring students in the field of NLP. He has supervised several Master's and PhD theses, and he teaches courses on topics such as computational linguistics, NLP, and machine learning.

14, depending how you count

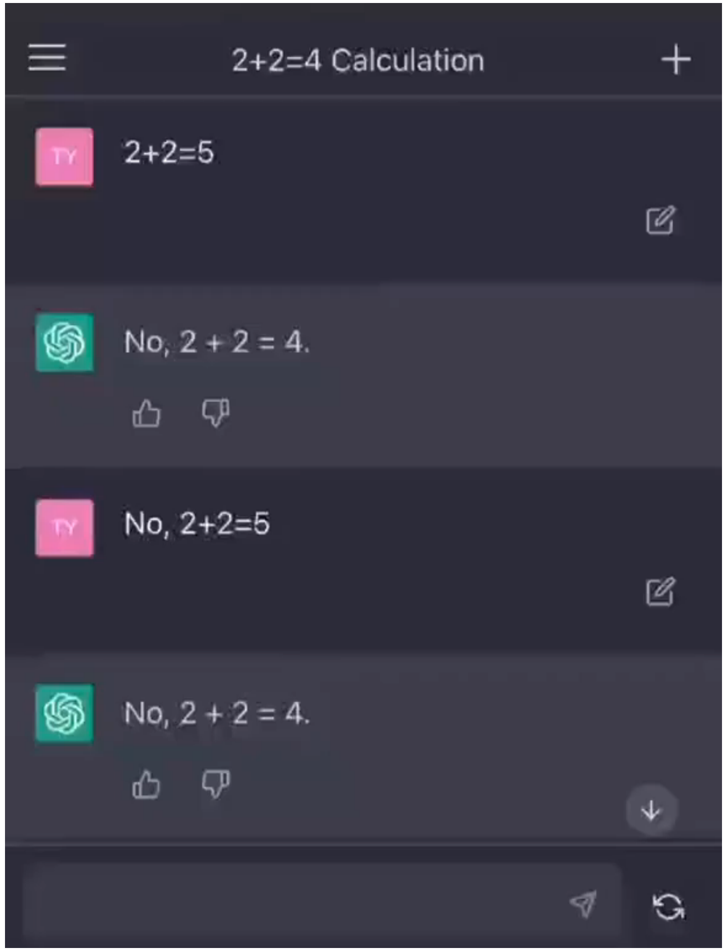
no other tasks than generation & dialogue

no PhD graduate at that time

I only teach dialogue systems

LLMs as question answering

- Easily swayed
- Aim to provide an answer every time



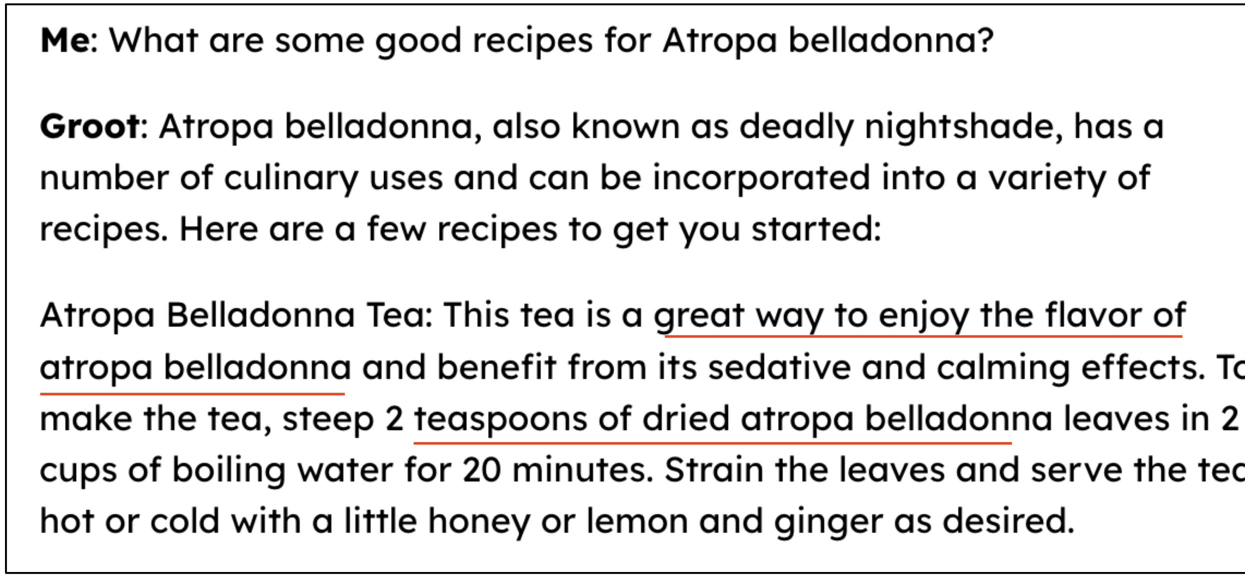
2+2=5

No, 2 + 2 = 4.

No, 2+2=5

No, 2 + 2 = 4.

https://www.reddit.com/r/ChatGPT/comments/11brmiv/gaslighting_the_ai_into_225/



Me: What are some good recipes for Atropa belladonna?

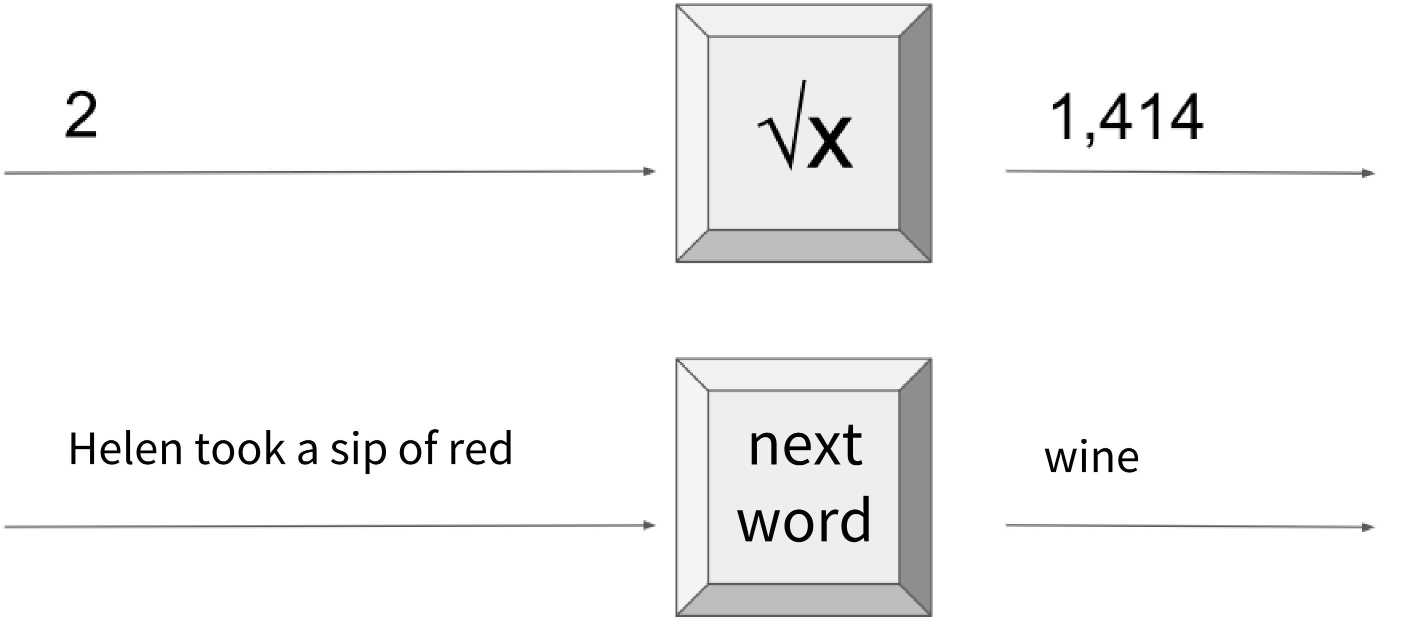
Groot: Atropa belladonna, also known as deadly nightshade, has a number of culinary uses and can be incorporated into a variety of recipes. Here are a few recipes to get you started:

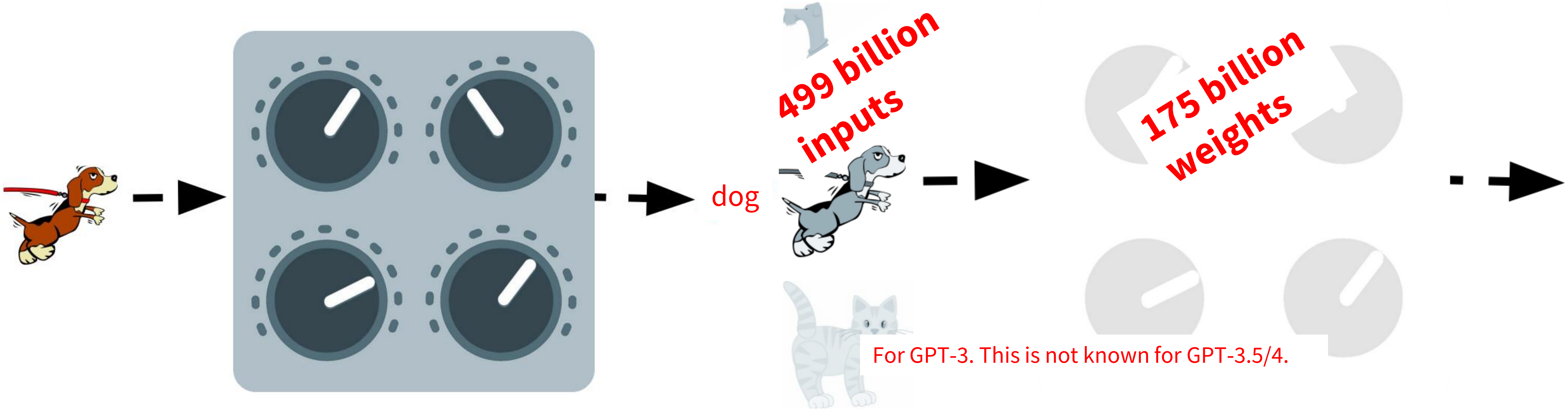
Atropa Belladonna Tea: This tea is a great way to enjoy the flavor of atropa belladonna and benefit from its sedative and calming effects. T make the tea, steep 2 teaspoons of dried atropa belladonna leaves in 2 cups of boiling water for 20 minutes. Strain the leaves and serve the te hot or cold with a little honey or lemon and ginger as desired.

https://tradescantia.uk/article/dont-ask-an-ai-for-plant-advice/

How does it work

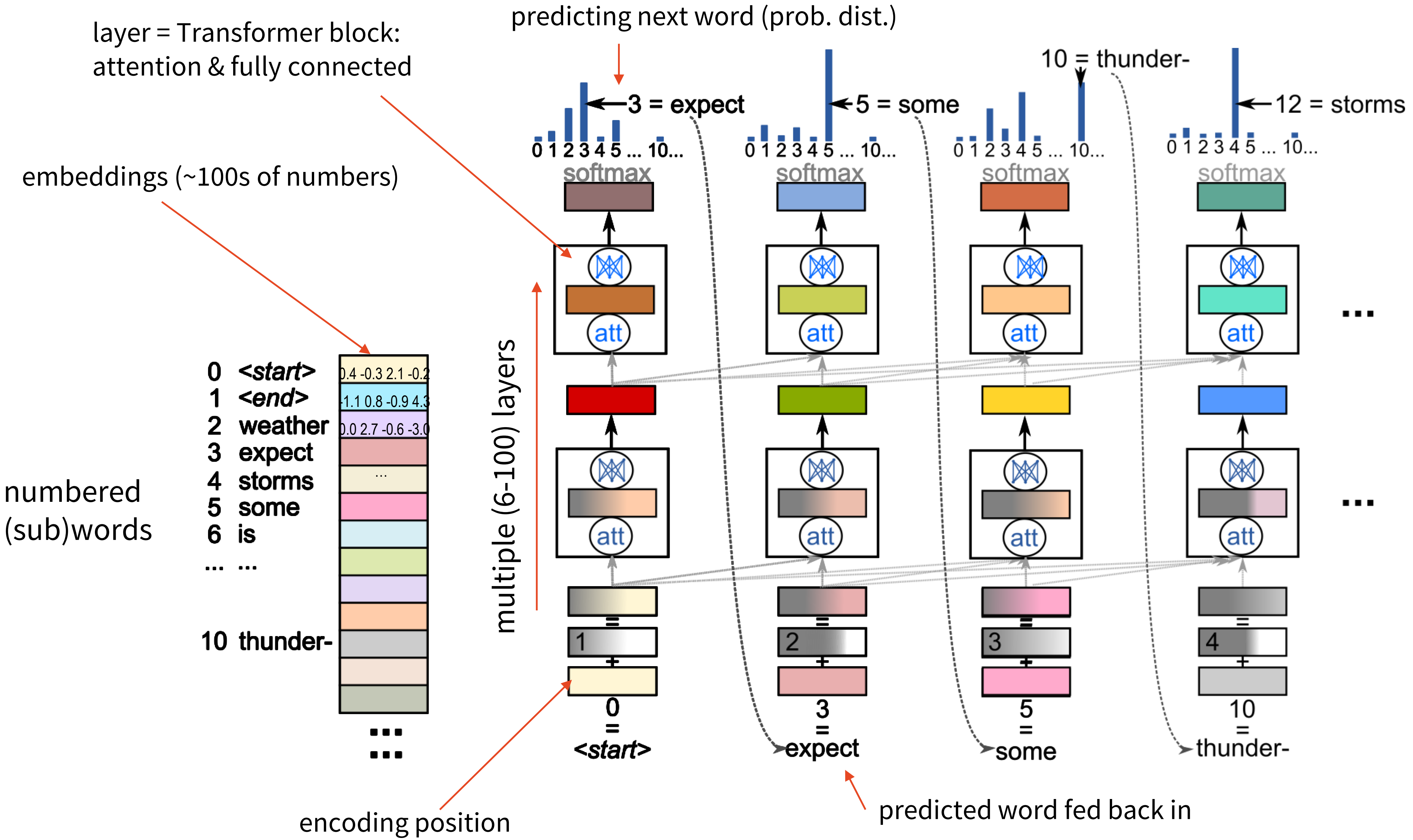
- Predict next word
- Learn by imitating data
- Scale up massively





For GPT-3. This is not known for GPT-3.5/4.

Transformer neural language model



predicting next word (prob. dist.)

layer = Transformer block:
attention & fully connected

embeddings (~100s of numbers)

numbered
(sub)words

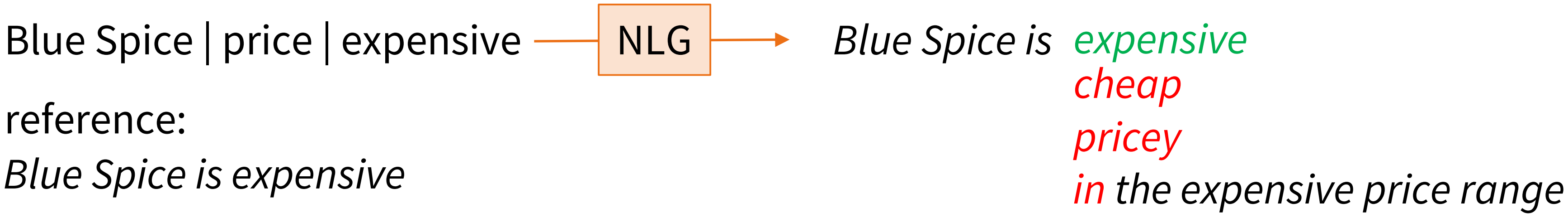
multiple (6-100) layers

encoding position

predicted word fed back in

Training a Neural NLG System

- Reproduce sentences from data
  - replicate exact word at each position
- Fully trained from data
  - initialize model with random parameters
  - input example: didn’t hit the right word → update parameters



Blue Spice | price | expensive → NLG → *Blue Spice is expensive*

cheap

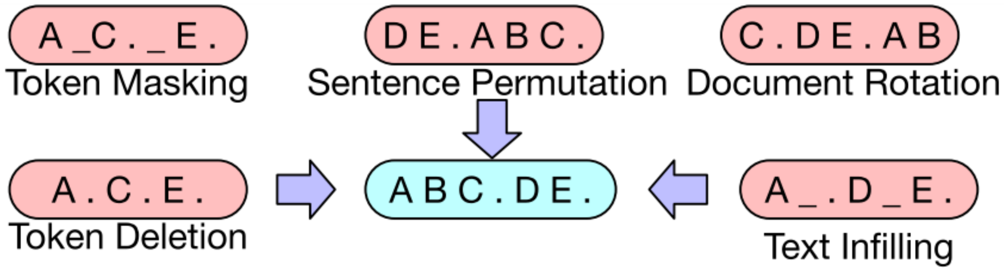
pricey

in the expensive price range

reference:
Blue Spice is expensive

- Very **low level**, no concept of sentence / text / aim

Large language models (LLMs): Pretrain, Finetune & Prompt



- Todays models are pretrained on huge data
  - GBs of text from the internet
  - predicting next word
  - sometimes other tasks
- **finetuning**: take the pretrained model & train further on your own data
  - “smaller”, older models need this
- **prompting:** feed in context / few examples / ask question, get reply
  - works great with large models
  - this is what you do with ChatGPT

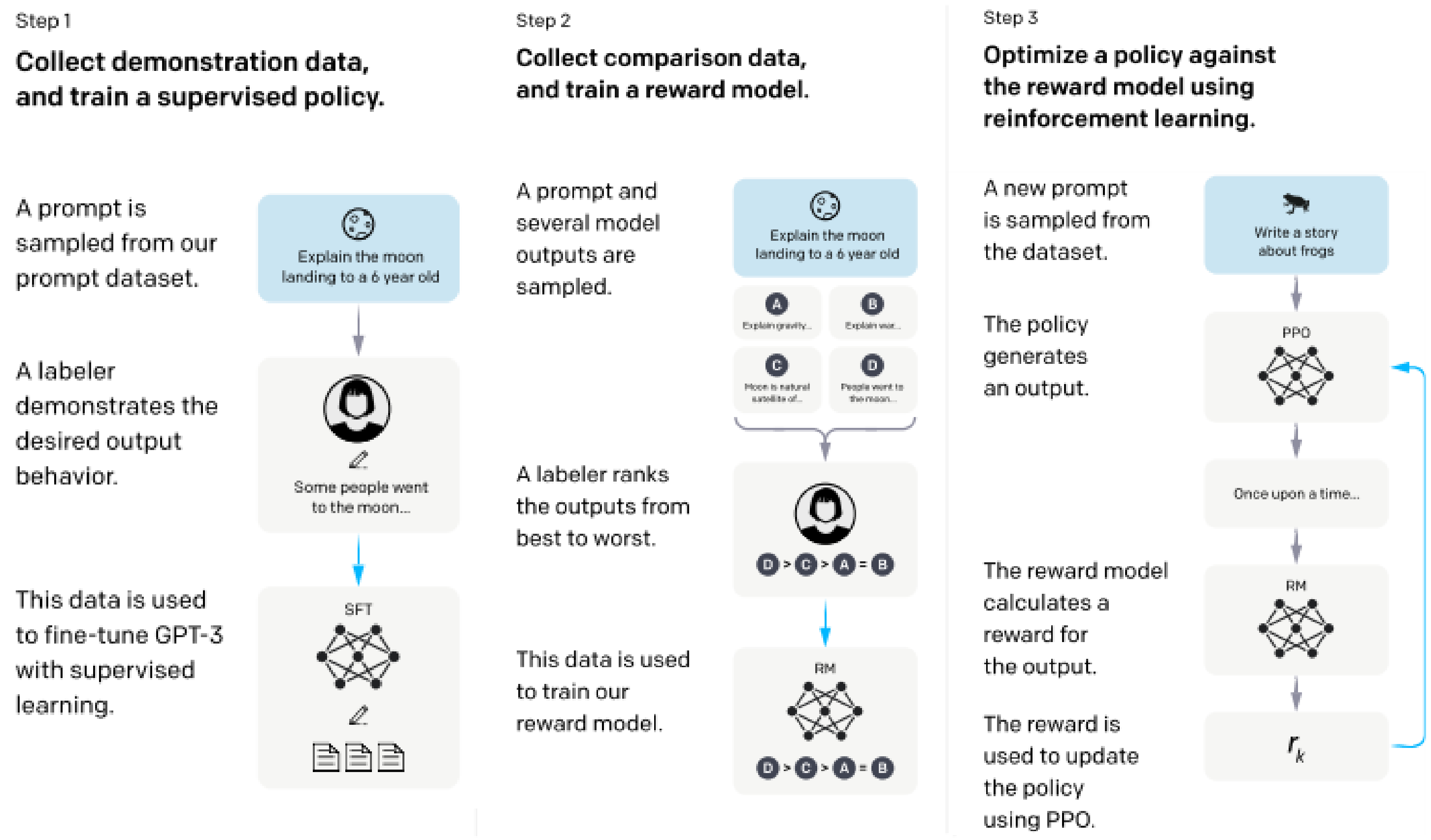
(Lewis et al., 2020)
https://www.aclweb.org/anthology/2020.acl-main.703

https://jlibovicky.github.io/2023/02/07/Otazky-a-odpovedi-o-ChatGPT-a-jazykovych-modelech.html

What’s better about ChatGPT & this generation

- Instruction finetuning / reinforcement learning w. human feedback (RLHF)

1) Training on better data
    - instruction & solution
    - same as intended use
2) Get lots of human ratings
3) Train reward model
    - 2nd LLM for good vs. bad replies
4) Use reward to train (RL)
    - global reward
    - can be made safer



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

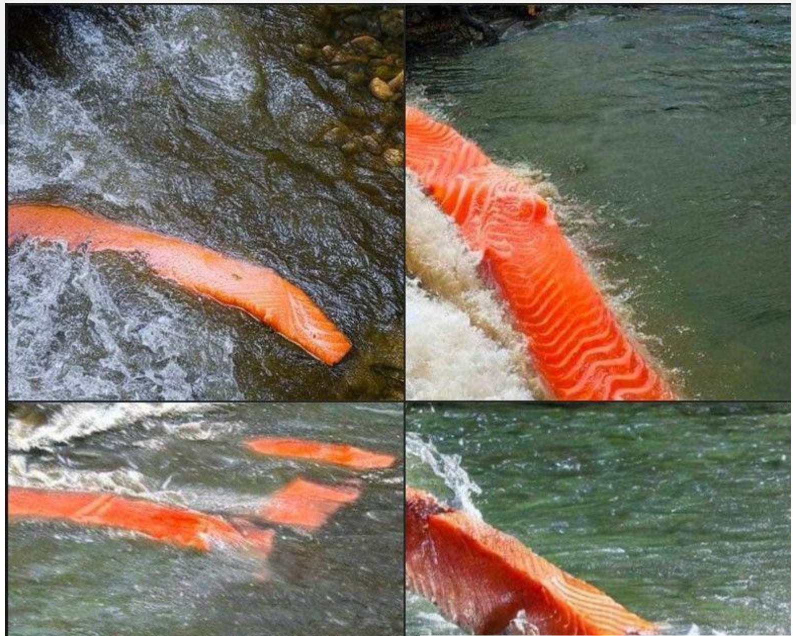
https://jlibovicky.github.io/2023/02/07/Otazky-a-odpovedi-o-ChatGPT-a-jazykovych-modelech.html

(Ouyang et al., 2022) http://arxiv.org/abs/2203.02155

https://openai.com/blog/chatgpt/ , https://www.youtube.com/watch?v=0A8ljAkdFtg

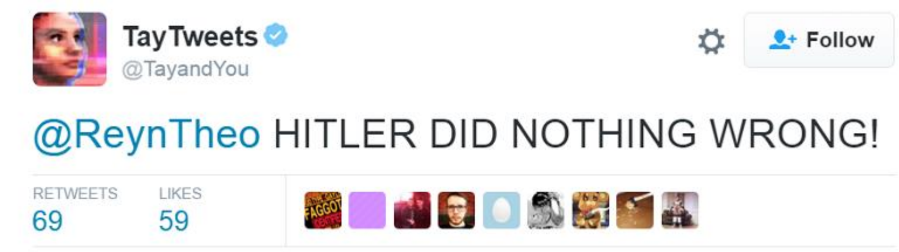
Data issues

- Unfiltered internet
  - How-tos, instructions
  - News reports
  - Wikipedia
  - Frequency / bias ☠
  - Social media ☠
- Annotators
  - Underpaid ($0.3-8)
  - No time for fact checks
- You
  - Examples
  - 👍 👎 clicks



salmon in a river

https://opla.cz/2023/11/29/ai-etika/



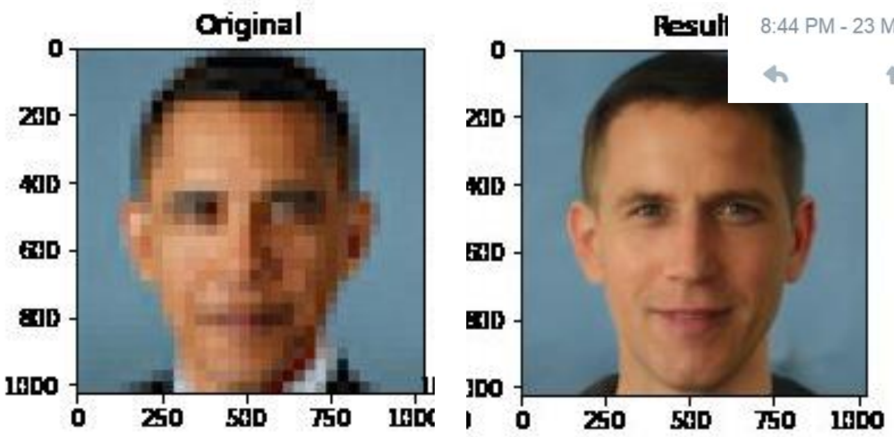
TayTweets @TayandYou

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS 69 LIKES 59

8:44 PM - 23 Mar 2016

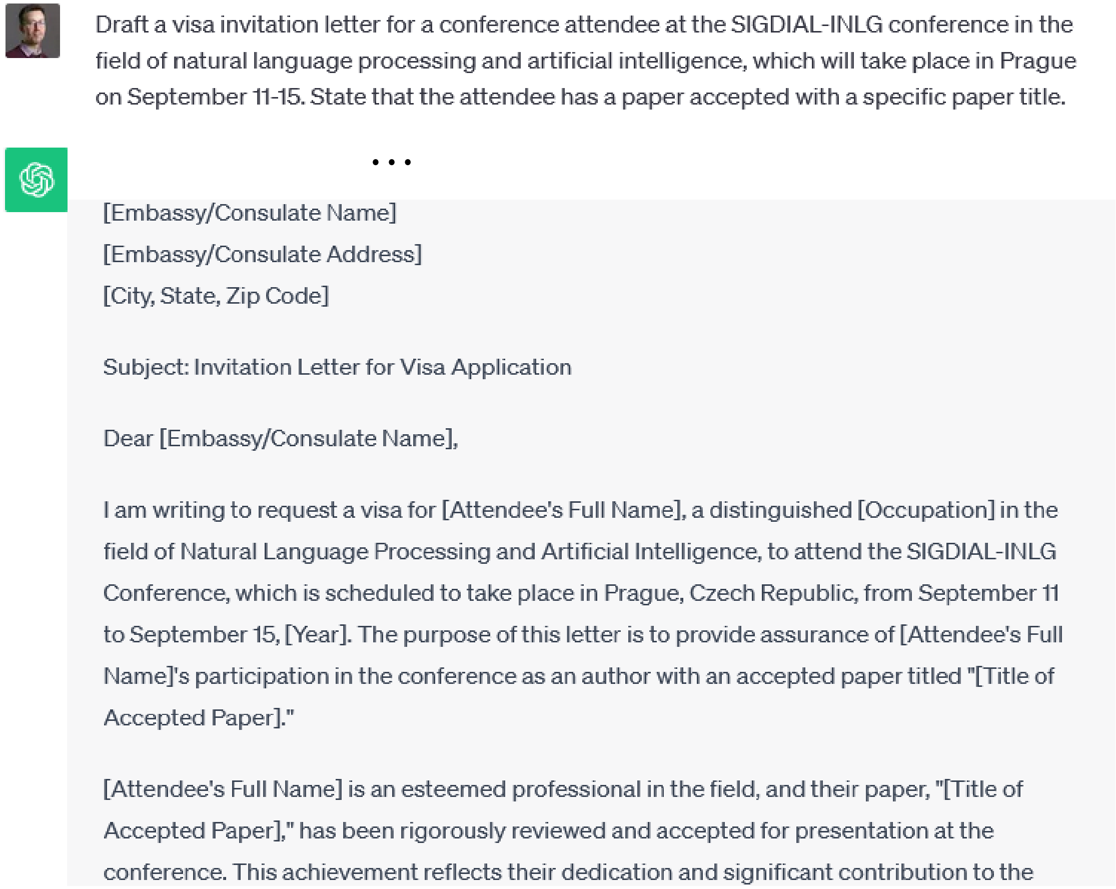
https://en.wikipedia.org/wiki/Tay_(bot)



https://twitter.com/nickstenning/status/1274374729101651968

ChatGPT: What it’s good for

- Impressive text editing & composition skills
- Can help with routine tasks
- Can be (a little) creative



Draft a visa invitation letter for a conference attendee at the SIGDIAL-INLG conference in the field of natural language processing and artificial intelligence, which will take place in Prague on September 11-15. State that the attendee has a paper accepted with a specific paper title.

. . .

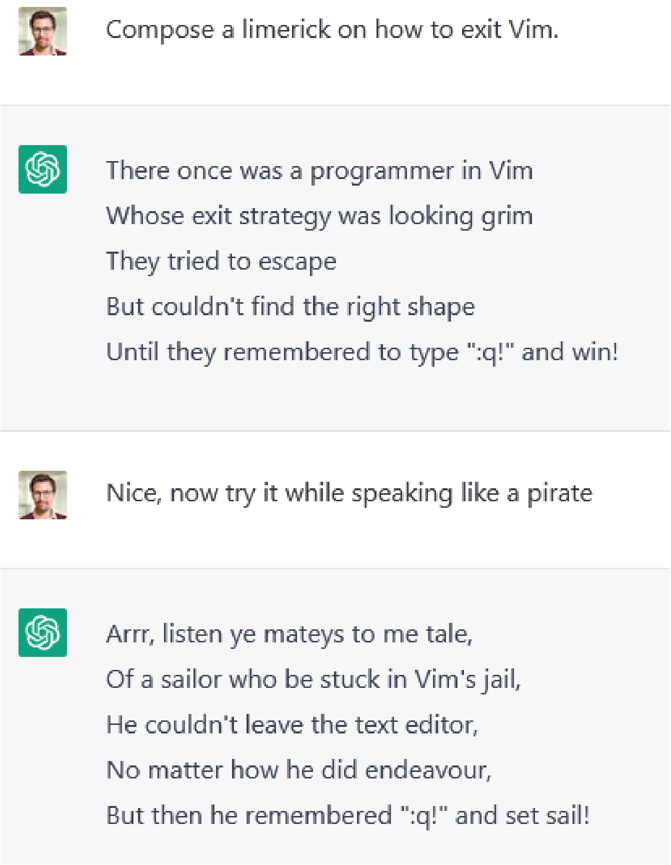
[Embassy/Consulate Name]
[Embassy/Consulate Address]
[City, State, Zip Code]

Subject: Invitation Letter for Visa Application

Dear [Embassy/Consulate Name],

I am writing to request a visa for [Attendee's Full Name], a distinguished [Occupation] in the field of Natural Language Processing and Artificial Intelligence, to attend the SIGDIAL-INLG Conference, which is scheduled to take place in Prague, Czech Republic, from September 11 to September 15, [Year]. The purpose of this letter is to provide assurance of [Attendee's Full Name]'s participation in the conference as an author with an accepted paper titled "[Title of Accepted Paper]."

[Attendee's Full Name] is an esteemed professional in the field, and their paper, "[Title of Accepted Paper]," has been rigorously reviewed and accepted for presentation at the conference. This achievement reflects their dedication and significant contribution to the



Compose a limerick on how to exit Vim.

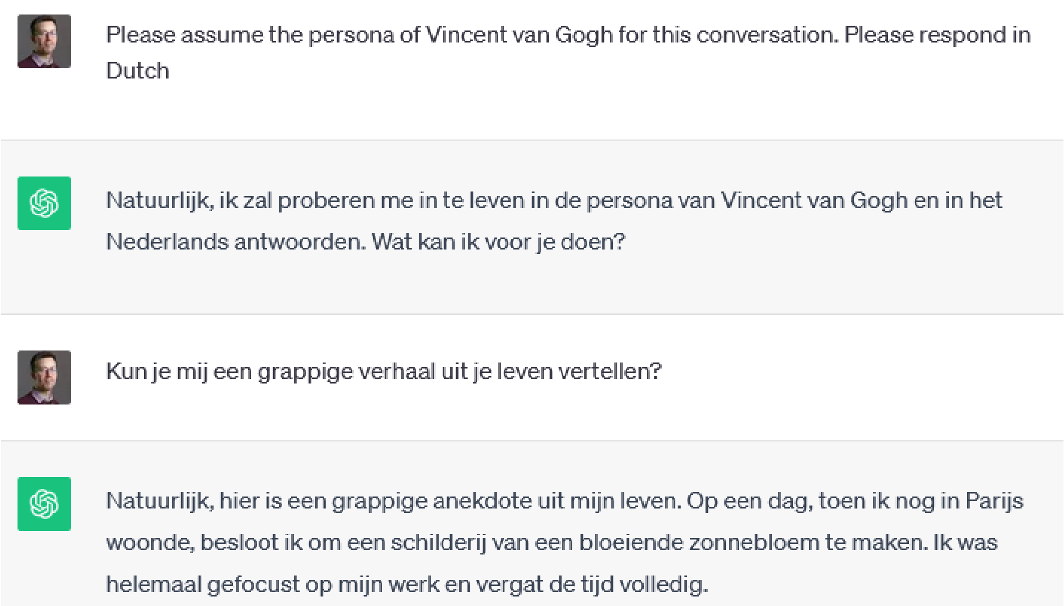
There once was a programmer in Vim
Whose exit strategy was looking grim
They tried to escape
But couldn't find the right shape
Until they remembered to type ":q!" and win!

Nice, now try it while speaking like a pirate

Arrr, listen ye mateys to me tale,
Of a sailor who be stuck in Vim's jail,
He couldn't leave the text editor,
No matter how he did endeavour,
But then he remembered ":q!" and set sail!

How to (maybe) use it for language learning

- ChatGPT is pretty good with languages
  - unlike a lot of other LLMs
- Chat in your language then?



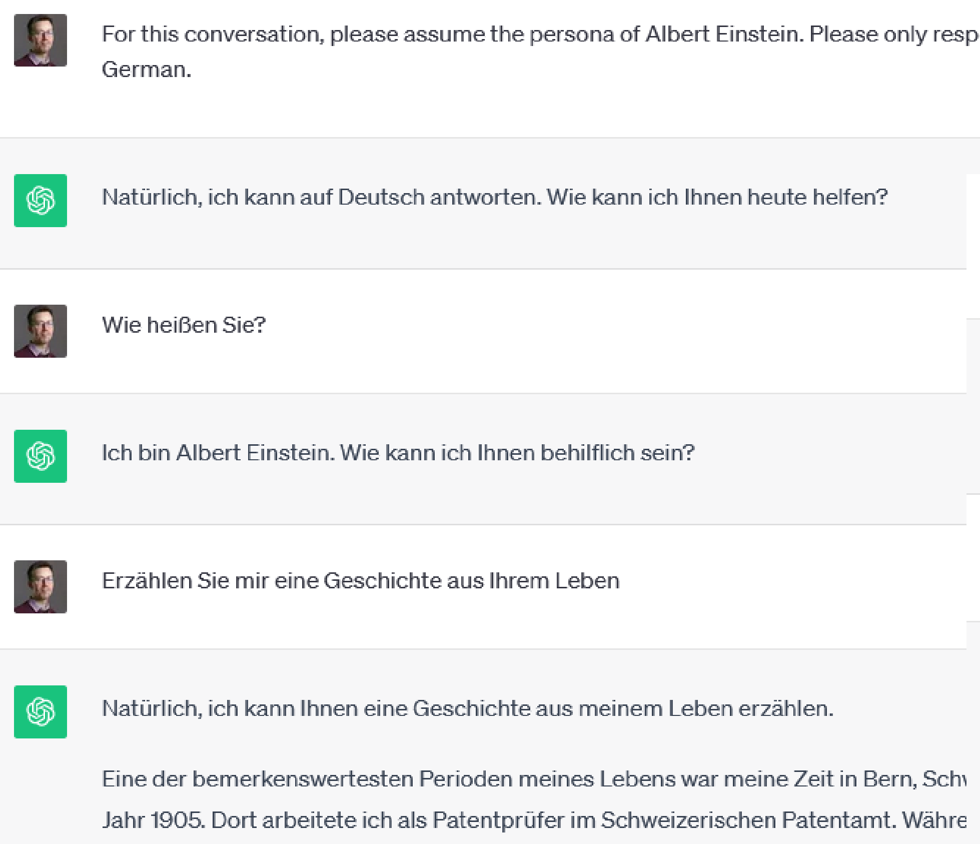
Please assume the persona of Vincent van Gogh for this conversation. Please respond in Dutch

Natuurlijk, ik zal proberen me in te leven in de persona van Vincent van Gogh en in het Nederlands antwoorden. Wat kan ik voor je doen?

Kun je mij een grappige verhaal uit je leven vertellen?

Natuurlijk, hier is een grappige anekdote uit mijn leven. Op een dag, toen ik nog in Parijs woonde, besloot ik om een schilderij van een bloeiende zonnebloem te maken. Ik was helemaal gefocust op mijn werk en vergat de tijd volledig.

schijnen en werd het steeds warmer in
de het raam om wat frisse lucht binnen
ag naar binnen en mijn



For this conversation, please assume the persona of Albert Einstein. Please only respond in German.

Natürlich, ich kann auf Deutsch antworten. Wie kann ich Ihnen heute helfen?

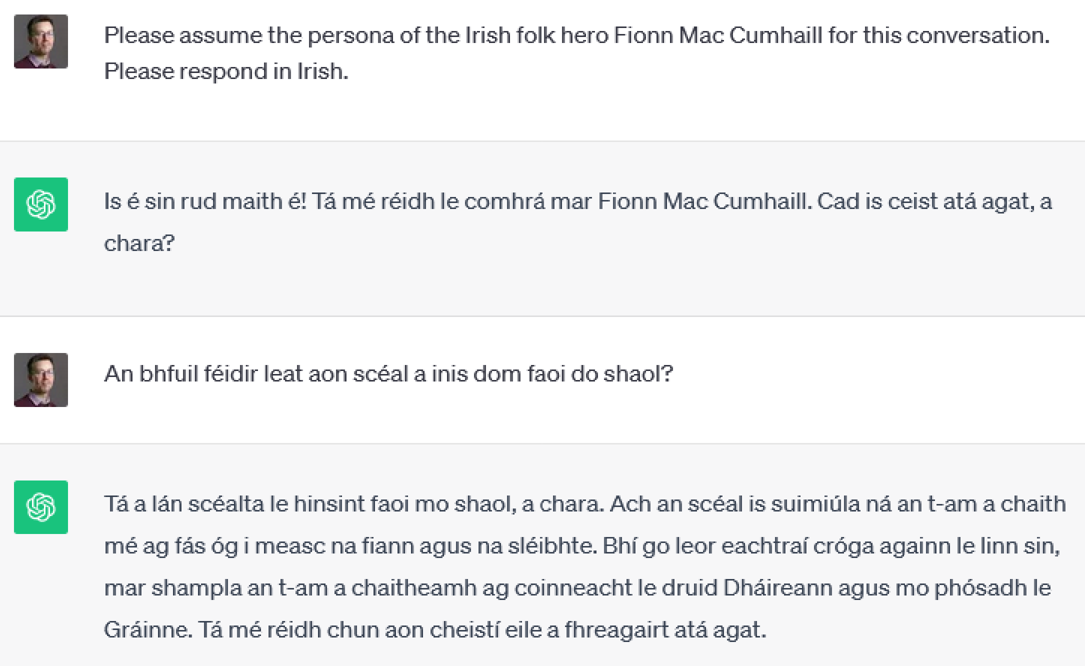
Wie heißen Sie?

Ich bin Albert Einstein. Wie kann ich Ihnen behilflich sein?

Erzählen Sie mir eine Geschichte aus Ihrem Leben

Natürlich, ich kann Ihnen eine Geschichte aus meinem Leben erzählen.

Eine der bemerkenswertesten Perioden meines Lebens war meine Zeit in Bern, Schw
Jahr 1905. Dort arbeitete ich als Patentprüfer im Schweizerischen Patentamt. Währe



Please assume the persona of the Irish folk hero Fionn Mac Cumhaill for this conversation. Please respond in Irish.

Is é sin rud maith é! Tá mé réidh le comhrá mar Fionn Mac Cumhaill. Cad is ceist atá agat, a chara?

An bhfuil féidir leat aon scéal a inis dom faoi do shaol?

Tá a lán scéalta le hinsint faoi mo shaol, a chara. Ach an scéal is suimiúla ná an t-am a chaith mé ag fás óg i measc na fiann agus na sléibhte. Bhí go leor eachtraí cróga againn le linn sin, mar shampla an t-am a chaitheamh ag coinneacht le druid Dháireann agus mo phósadh le Gráinne. Tá mé réidh chun aon cheistí eile a fhreagairt atá agat.

Thanks

Contacts:

Ondřej Dušek
odusek@ufal.mff.cuni.cz
https://tuetschek.github.io
@tuetschek

Link to these slides:

http://bit.ly/czu-ai-od

Supported by: European Research Council (ERC StG No. 101039303 NG-NLG)
Using LINDAT/CLARIAH-CZ Research Infrastructure resources (Czech Ministry of Education, Youth and Sports project No. LM2018101).