CSU33061 AI 1

Introduction

www.scss.tcd.ie/Tim.Fernando/AI

Outline

Can machines think?

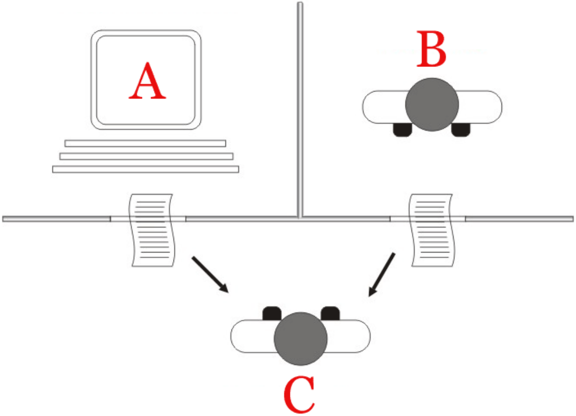
- Turing test & the ELIZA effect
- Generative AI (ChatGPT ...)

Improving intelligence

- agent & environment
- search & learning

Can machines think? (Turing 1950)

Turing test: can C tell A from B?

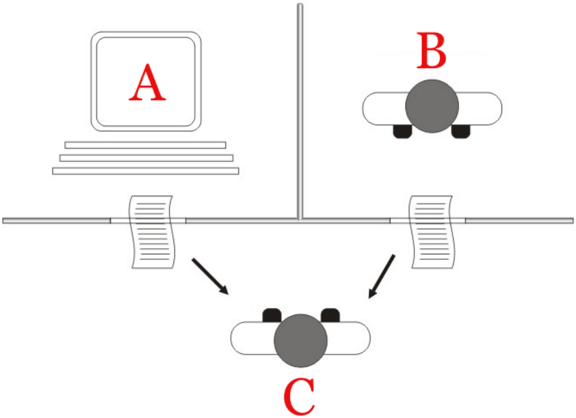


From Wikipedia, (Juan Alberto Sáánchez Margallo)

Intelligence operationalized: subject to testing

Can machines think? (Turing 1950)

Turing test: can C tell A from B?



From Wikipedia, (Juan Alberto Sáánchez Margallo)

Intelligence operationalized: subject to testing and cheating

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):
When you get enraged because Siri just doesn't get it.

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it   (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):

When you get enraged because Siri just doesn't get it.

Fooling fools is easy

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):
When you get enraged because Siri just doesn't get it.

Fooling fools is easy
though even fools can get angry.

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):
When you get enraged because Siri just doesn't get it.

Fooling fools is easy
though even fools can get angry.

Programs have improved hugely

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it   (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):
When you get enraged because Siri just doesn't get it.

Fooling fools is easy
though even fools can get angry.

Programs have improved hugely,
humans largely not.

ChatGPT 3.5

You

Is ChatGPT bigger than Taylor Swift?

ChatGPT

No, ChatGPT is not a physical entity and cannot be compared in terms of size to a person like Taylor Swift. ChatGPT is a computer program, specifically a language model developed by OpenAI, while Taylor Swift is a human being and a well-known singer, songwriter, and actress. The concept of "size" doesn't apply in the same way to a computer program and a person. If you have any specific questions about either ChatGPT or Taylor Swift, feel free to ask!

You

But hasn't ChatGPT's impact been greater than Taylor Swift's? Shouldn't Time Magazine have made

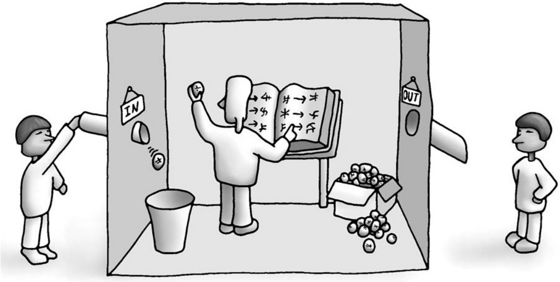
ChatGPT

As of my last knowledge update in January 2022, I cannot provide information on events or developments that occurred after that date. Therefore, I don't have information about Time Magazine's decisions or any comparisons between ChatGPT and Taylor Swift after that time.

However, it's worth noting that comparing the impact of a computer program like ChatGPT to that of a well-known and influential individual like Taylor Swift is challenging. Their impacts are fundamentally different and may not be directly comparable. Time Magazine's decisions for their annual recognitions often depend on a variety of factors, including cultural impact, influence, achievements, and societal contributions.

If there have been significant developments or events related to ChatGPT and Taylor Swift after my last update, I recommend checking the latest sources or news for the most up-to-date information.

Chinese room argument (Searle's thought experiment)

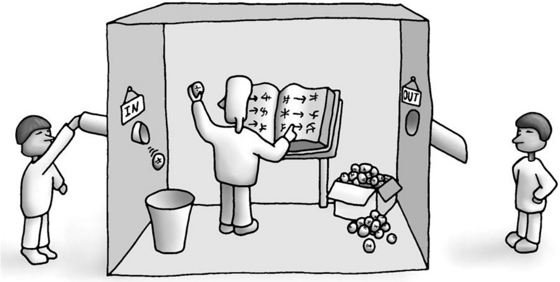


From http://america.pink/images/9/6/3/2/5/4/en/2-chinese-room.jpg

- a clerk can follow instructions for communicating in Chinese without understanding Chinese

Strong vs Weak AI

Chinese room argument (Searle's thought experiment)



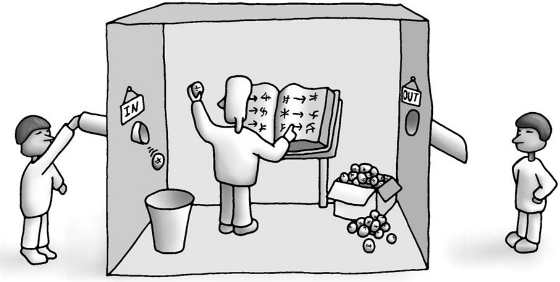
From http://america.pink/images/9/6/3/2/5/4/en/2-chinese-room.jpg

- a clerk can follow instructions for communicating in Chinese without understanding Chinese

mindless obedience $\neq$ mind from biological processes

Strong vs Weak AI

Chinese room argument (Searle's thought experiment)



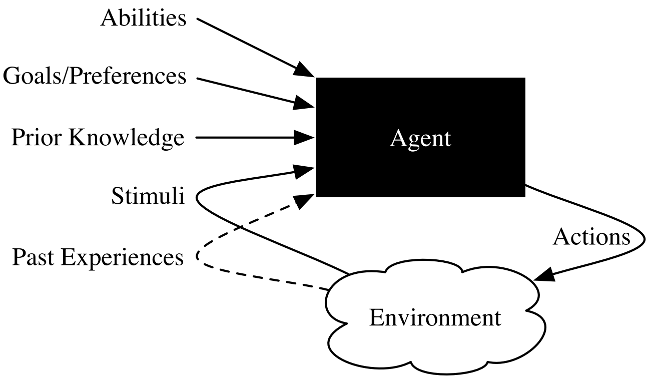
From http://america.pink/images/9/6/3/2/5/4/en/2-chinese-room.jpg

- a clerk can follow instructions for communicating in Chinese without understanding Chinese

mindless obedience $\neq$ mind from biological processes

Just do it $\rightsquigarrow$ Black Box, judged by its actions

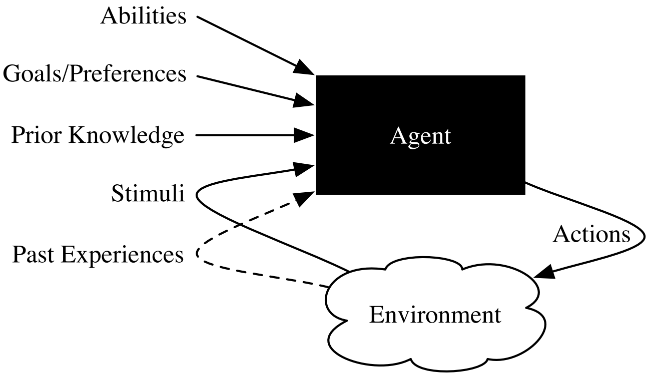
Locating intelligence (black box)



Poole & Mackworth

Intelligence: (abilities, goals, ..., experience) $\mapsto$ action

Locating intelligence (black box)



Poole & Mackworth

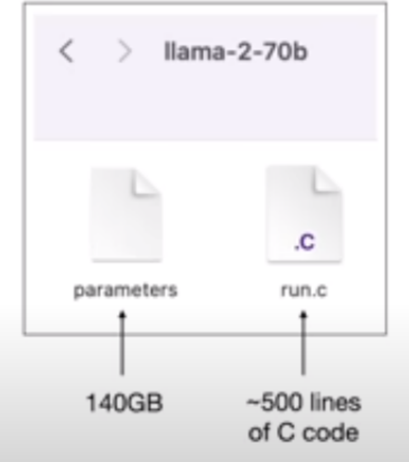
Intelligence: (abilities, goals, ..., experience) $\mapsto$ action

Turing test: what to say $\rightsquigarrow$ what to do

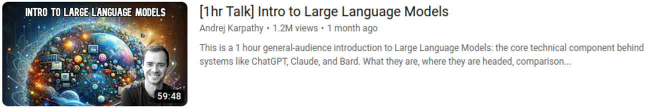
Environment vs agent

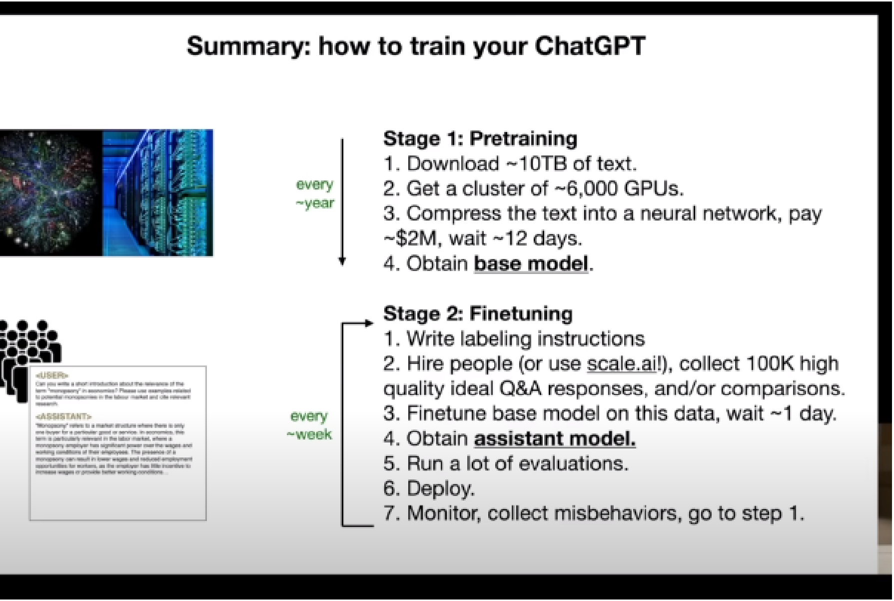
environment	agent
data	program
BIG DATA	Cognitive Revolution
experience	hard-wired
empiricist	rationalist
nurture	nature
behaviorist	nativist
tabula rasa	innate

Environment vs agent

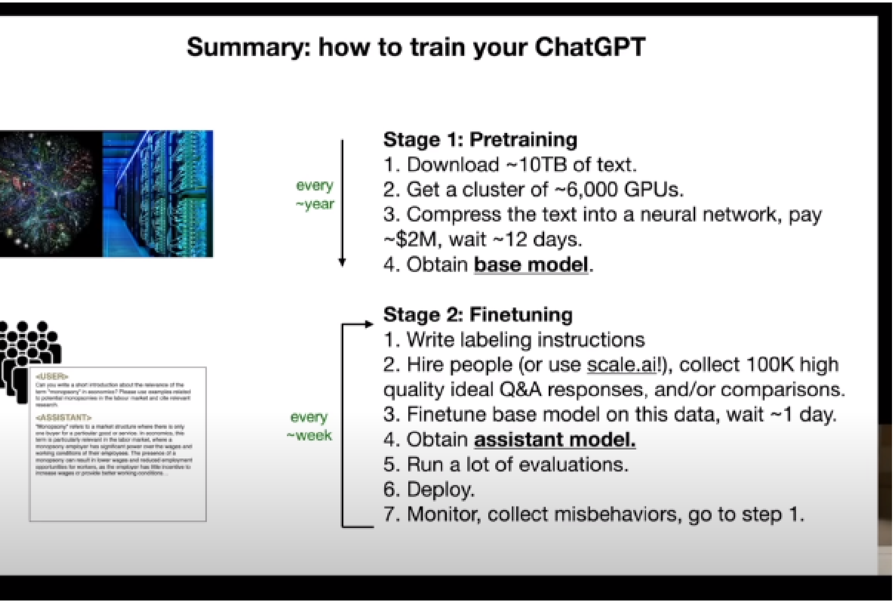


From AK:
A.Karpathy



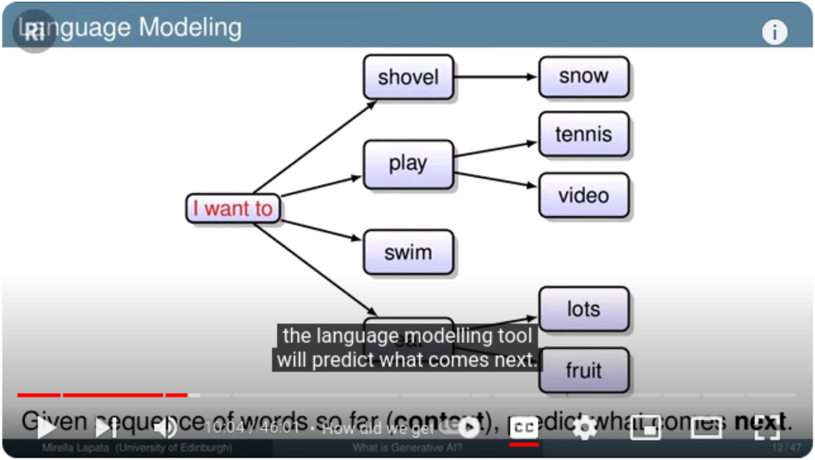

Summary: how to train your ChatGPT
every
~year
Stage 1: Pretraining
1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~$2M, wait ~12 days.
4. Obtain base model.
<USER>
<ASSISTANT>
every
~week
Stage 2: Finetuning
1. Write labeling instructions
2. Hire people (or use scale.ai!), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain assistant model.
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

From AK



From AK

Pretrain	quantity	predict next word	base model
Finetune	quality	learn QA pairs	assistant model

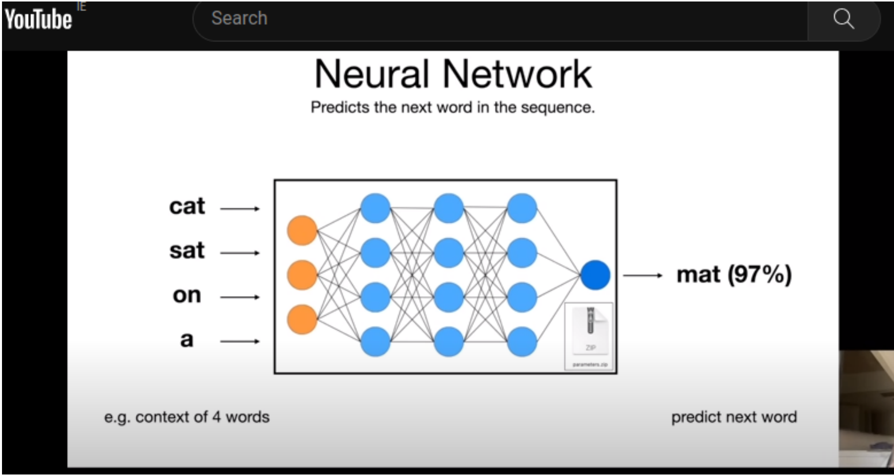


From ML: M.Lapata

YouTube · The Royal Institution
325K+ views · 3 months ago

What is generative AI and how does it work? - YouTube

How are technologies like ChatGPT created? And what does the future hold for **AI** language models? This talk was filmed at the Royal ...


YouTube IE
Search
Neural Network
Predicts the next word in the sequence.
cat
sat
on
a
mat (97%)
ZIP
e.g. context of 4 words
predict next word

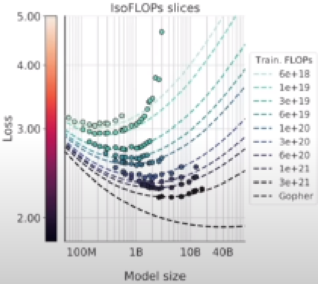
From AK

LLM Scaling Laws

Performance of LLMs is a smooth, well-behaved, predictable function of:
- **N**, the number of parameters in the network
- **D**, the amount of text we train on

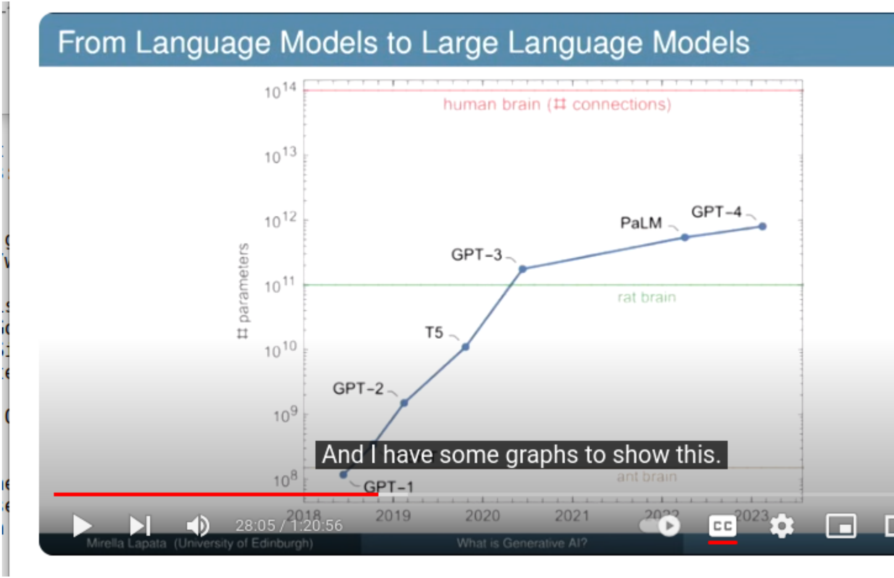
And the trends do not show signs of "topping out"

=> We can expect more intelligence "for free" by scaling

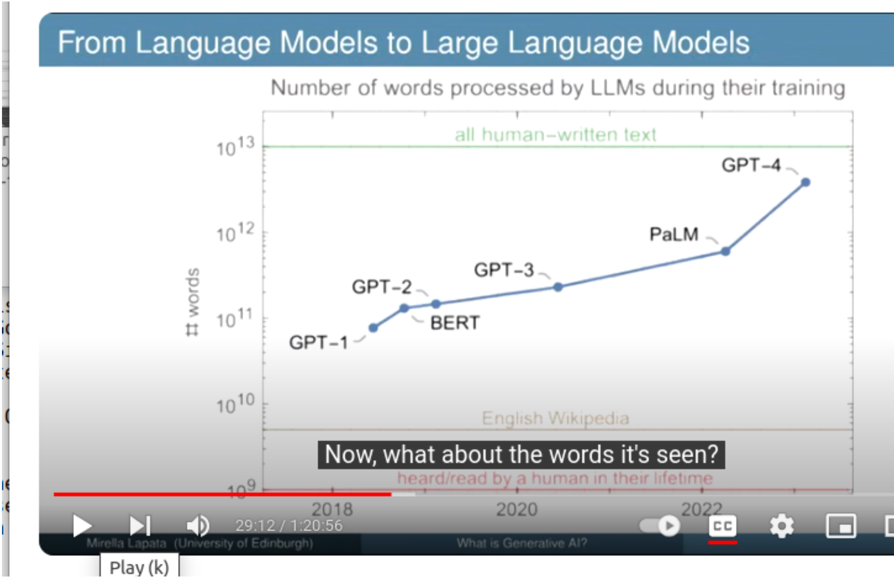


[Training Compute-Optimal Large Language

From AK

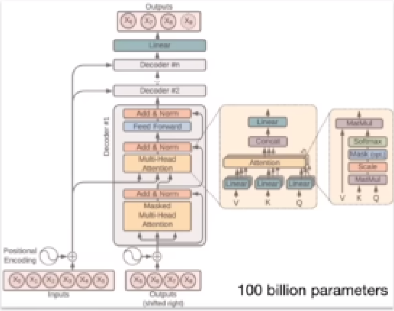

From Language Models to Large Language Models
human brain (# connections)
rat brain
ant brain
parameters
GPT-1
GPT-2
T5
GPT-3
PaLM
GPT-4
2018
2019
2020
2021
2023
And I have some graphs to show this.
28:05 / 1:20:56
Mirella Lapata (University of Edinburgh)
What is Generative AI?

From ML


From Language Models to Large Language Models
Number of words processed by LLMs during their training
all human-written text
English Wikipedia
heard/read by a human in their lifetime
GPT-1
GPT-2
BERT
GPT-3
PaLM
GPT-4
words
10^13
10^12
10^11
10^10
10^9
2018
2020
2022
Now, what about the words it's seen?
29:12 / 1:20:56
Mirella Lapata (University of Edinburgh)
What is Generative AI?
Play (k)

From ML

How does it work?



Little is known in full detail...

- Billions of parameters are dispersed through the network
- We know how to iteratively adjust them to make it better at prediction.
- We can measure that this works, but we don't really know how the billions of parameters collaborate to do it.

They build and maintain some kind of knowledge database, but it is a bit strange and imperfect:

Recent viral example: "reversal curse"

Q: "Who is Tom Cruise's mother"?
A: Mary Lee Pfeiffer ✅

Q: "Who is Mary Lee Pfeiffer's son?"
A: I don't know ❌ 🤔

=> think of LLMs as mostly inscrutable artifacts, develop correspondingly sophisticated evaluations.

From AK

Search: computation

```
search(Node) :- goal(Node).

search(Node) :- move(Node,Next), search(Next).
```

Search: computation & learning

```
search(Node) :- goal(Node).

search(Node) :- move(Node,Next), search(Next).
```

From Slide 21 above

Pretrain	quantity	predict next word	base model
Finetune	quality	learn QA pairs	assistant model

Search: computation & learning

```
search(Node) :- goal(Node).

search(Node) :- move(Node,Next), search(Next).
```

From Slide 21 above

Pretrain	quantity	predict next word	base model	know
Finetune	quality	learn QA pairs	assistant model	align

Search: computation & learning

```
search(Node) :- goal(Node).

search(Node) :- move(Node,Next), search(Next).
```

From Slide 21 above

Pretrain	quantity	predict next word	base model	know
Finetune	quality	learn QA pairs	assistant model	align

In July, OpenAI announced a new research program on "superalignment." The program has the ambitious goal of solving the hardest problem in the field, known as AI alignment, by 2027, an effort to which OpenAI is dedicating 20 percent of its total computing power.

What is the AI alignment problem? It's the idea that AI systems' goals may not align with those of humans, a problem that would be heightened if superintelligent AI systems are developed.

E. Strickland, IEEE Spectrum, Aug 2023

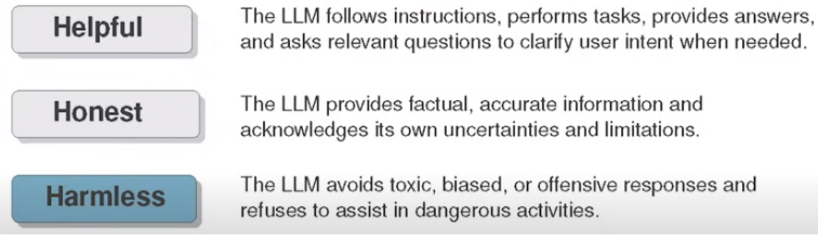
Search: computation & learning

```
search(Node) :- goal(Node).

search(Node) :- move(Node,Next), search(Next).
```

From Slide 21 above

Pretrain	quantity	predict next word	base model	know
Finetune	quality	learn QA pairs	assistant model	align



Askell et al. 2021, from ML