The Many Flavors of Artificial Intelligence & Understanding Large Language Models

Tom Dietterich, Oregon State University

Outline

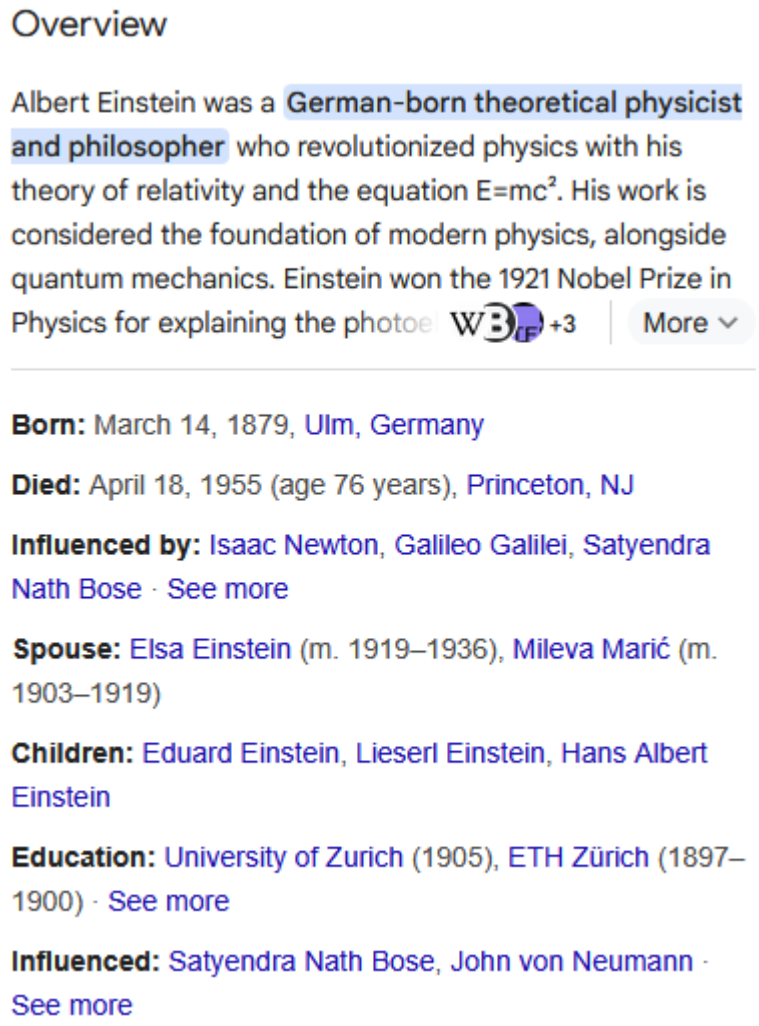
- Part 1: What is AI?
- Part 2: Understanding Large Language Models
- Part 3: Visions of the AI Future

Part 1: What is AI?

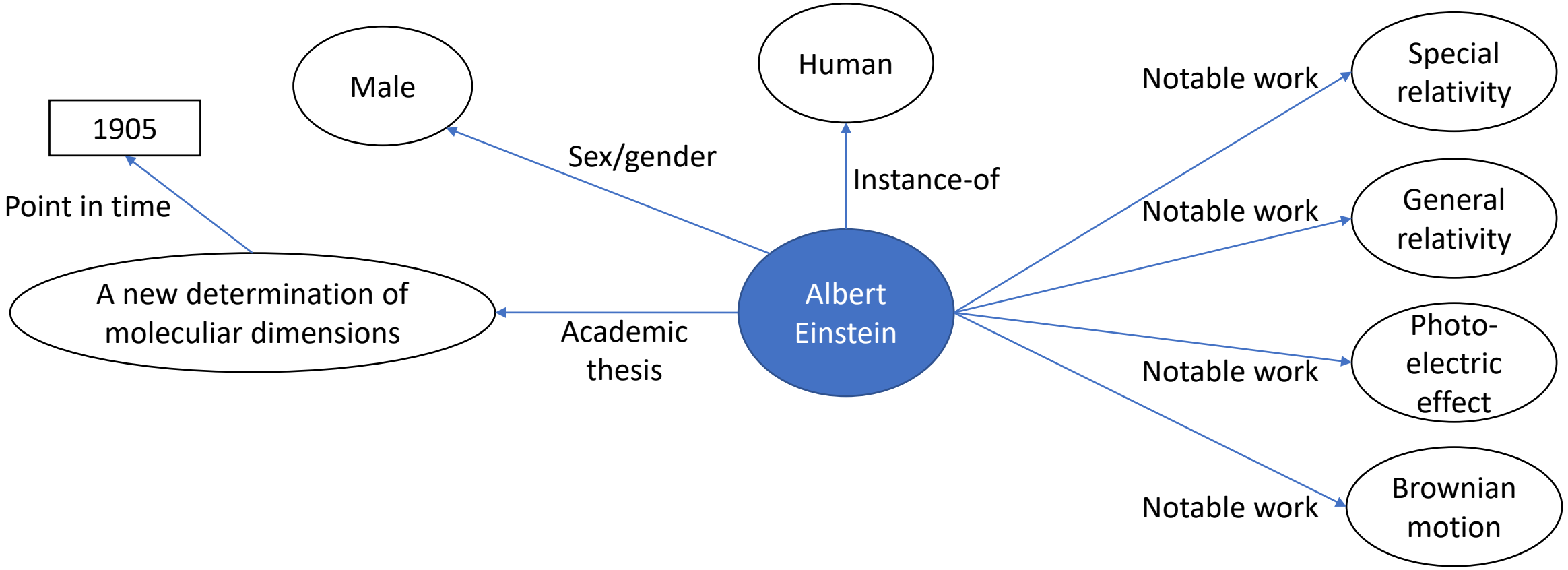
- Knowledge-Based Systems
- Optimizing Industrial Processes
- Natural language processing
- Computer vision and computer graphics
- Scientific applications
- Playing Games

Knowledge-Based Systems

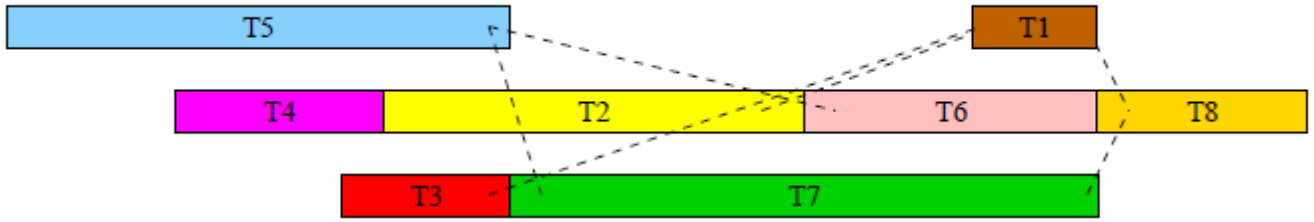
- Expert systems: Automated tax filing software
- Business workflows
- Information extraction: Google “snippets” and information boxes
- Google Knowledge Graph
- Recommendation systems



Knowledge Graph



Optimizing Industrial Processes

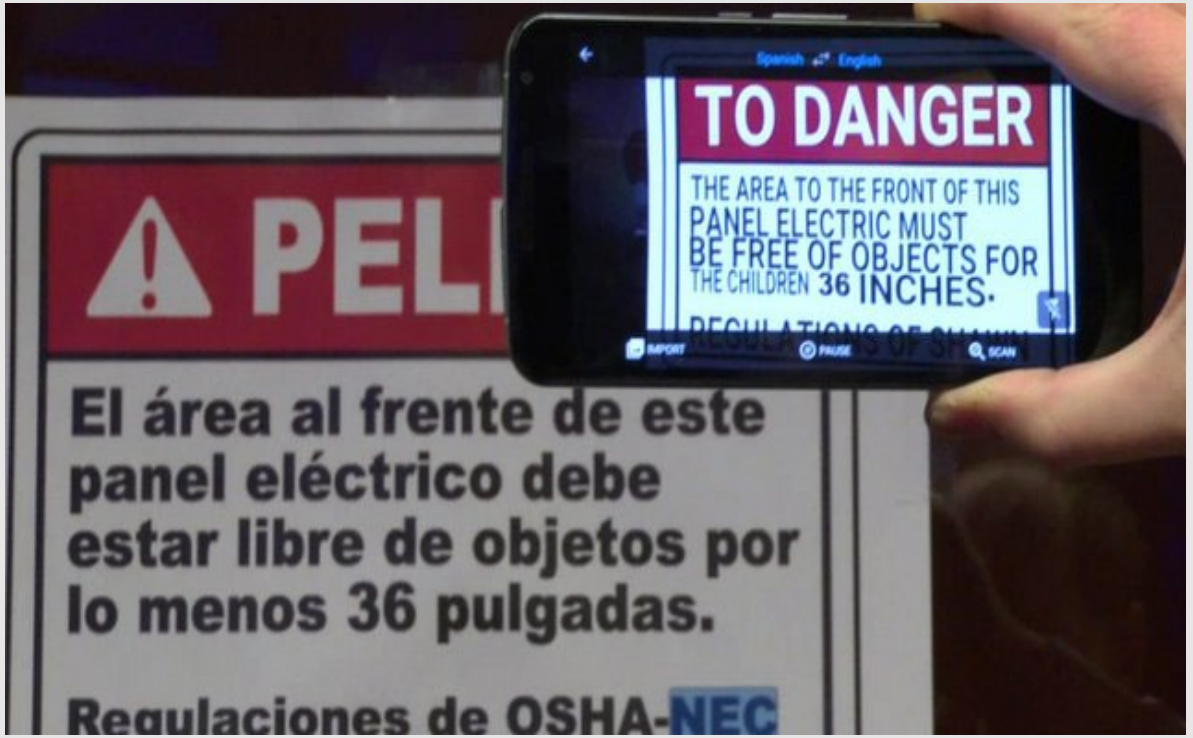


- Planning and scheduling software: shipping, NASA missions, airlines, UPS, Fedex, US Army, etc.
  - Set of tasks with prerequisites
  - Each task also requires “resources” (e.g., fork lift, testing rig, test engineer) that must be shared

Natural language processing

- Natural language translation (google translate)
- Large language models (ChatGPT)

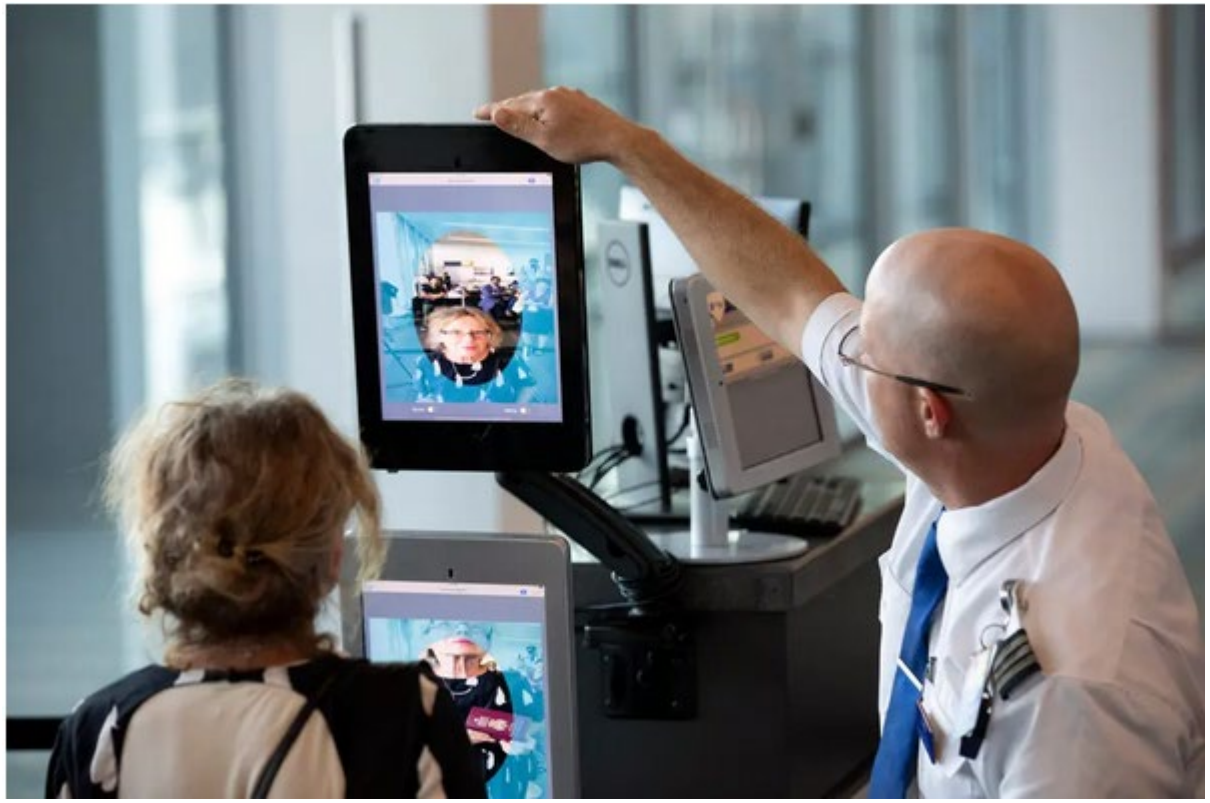
Google Translate from Images



Credit: www.bbc.com

Computer vision and computer graphics

- Face recognition
- Self-driving cars (and other robots)
- Image captioning (image-to-text)
- Image synthesis (text-to-image)



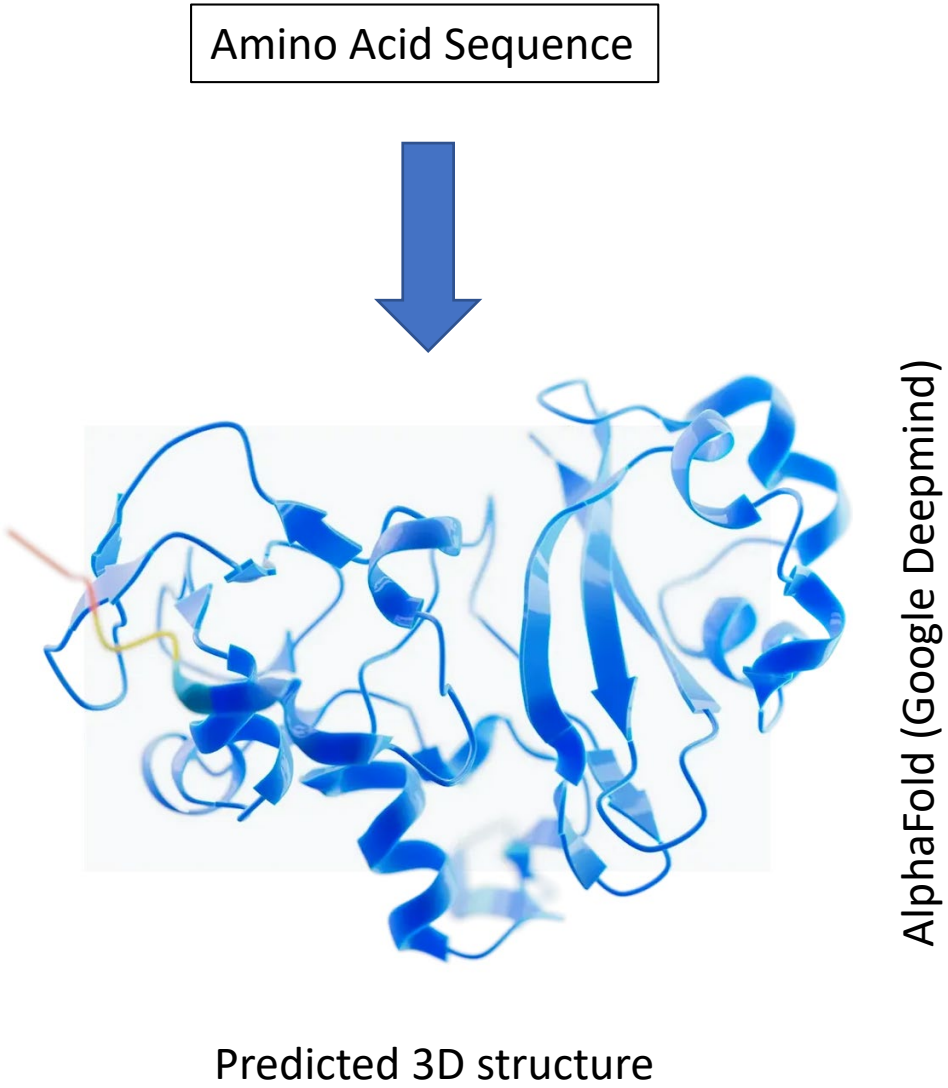
Ray Whitehouse for The New York Times

Self-driving cars (and other robots)

- Waymo robot cars
  - Cameras
  - LiDAR
  - RADAR
- Combines
  - Computer vision
  - Predicting future behavior of cars, pedestrians, animals
  - Planning a trajectory
  - Controlling the car to execute the trajectory
  - In milliseconds
- Full Autonomy

Scientific applications

- Protein folding
- Computational fluid dynamics
- Materials science
- Drug design
- Medical and scientific imaging



Games

- Playing Games
  - Chess, Go, Poker, Minecraft, etc.
- Designing Games
  - Dynamically designing new levels for video games

Part 2: Large Language Models

- A language model is a representation of the sequences of words that appear in a language
  - Unigram model: What is the frequency of each word?
  - Bigram model: What is the frequency of each pair of words
  - Trigram model: What is the frequency of each triple of words
  - Large Language Model: Examines several thousands of preceding words to predict the next word

Example: A Language Model for Sherlock Holmes

- The entire text of “The Adventures of Sherlock Holmes” and “The Casebook of Sherlock Holmes” is available online (e.g., at Project Gutenberg). I took this text and computed the statistics for unigram, bigram, and trigram language models

Generating “random” English from a unigram model

- Repeatedly select a word based on its frequency in Sherlock Holmes:

Unigram Statistics for Sherlock Holmes

Word	Frequency
.	14529
THE	9759
I	5258
AND	4941
A	4672
TO	4565
OF	4555
IT	3190
THAT	3108
IN	3054

Word	Frequency
YOU	2804
WAS	2768
HE	2760
IS	2113
HIS	2084
MY	1682
HAVE	1660
HAD	1535
WITH	1509
AS	1431

Word	Frequency
FOR	1322
AT	1313
BUT	1242
WHICH	1229
NOT	1187
ME	1115
BE	1090
WE	1008
HOLMES	979
THERE	944

Word	Frequency
THIS	938
HIM	879
SAID	849
FROM	827
HER	776
SO	773
NO	767
SHE	746
UPON	744
WHAT	742

Generating English from the unigram model

- Repeatedly choose a word at random in proportion to its frequency in the model

> Pipe-rack that. Of the to and can it in he wings that up stream own had old first is hat Watson anatomy been I sittingroom. Then as nothing my put as shut yourself. Forced. The passed have was getting agony ground of. Some station an. About us wanting was see be said question. In electric-blue might pointing the where daughter scent caused treat mirror shivers.

Random text from the bigram model

- To generate each word, only choose among words that follow the most recent word. Every pair of words occurred somewhere in Holmes:

And you from the necessary permits this fellow. Head of deduction that you make of a pretext. Said Miss Violet de Merville coldly. He stepped into the pastime brought in a situation Miss Mary was a day. I shall now. He looked at your permission I would suggest no suggestive than a long before you an end of Breckingridge of me Mr. I bathed in all wrong with a strange out-of-the-way place by a very soon devised seven years ago and tapestry and pity because you out he.

Random text from the trigram model

- Now the next word depends on the two preceding words.

Possible I made up my spirits and feeling that I would call here tomorrow evening. I know you well. She walked into the matter. There was some little time after listening to the hour when a door which led into the stable lane. So I think that her dowry will run to considerably over the telephone directory. Honest business men don’t conceal their place of the ancient hostel of which he explained in answer to this floor. About five feet deep. It is a fait accompli. I asked.

LLMs and the Transformer model

- The breakthrough that created LLMs is a mechanism for examining the preceding words and deciding how important each word is for predicting the next word. This is known as “attention”, because the LLM has learned what words to “pay attention to” in the context

So	I	think	that	her	dowry	will	run	to	considerably

more than her father lets on—perhaps five thousand pounds, if the rumours are to be credited. It is a tidy sum, sufficient to attract attention among the county families, and not without its influence on the calculations of certain gentlemen who have lately taken an interest in her walks and conversation

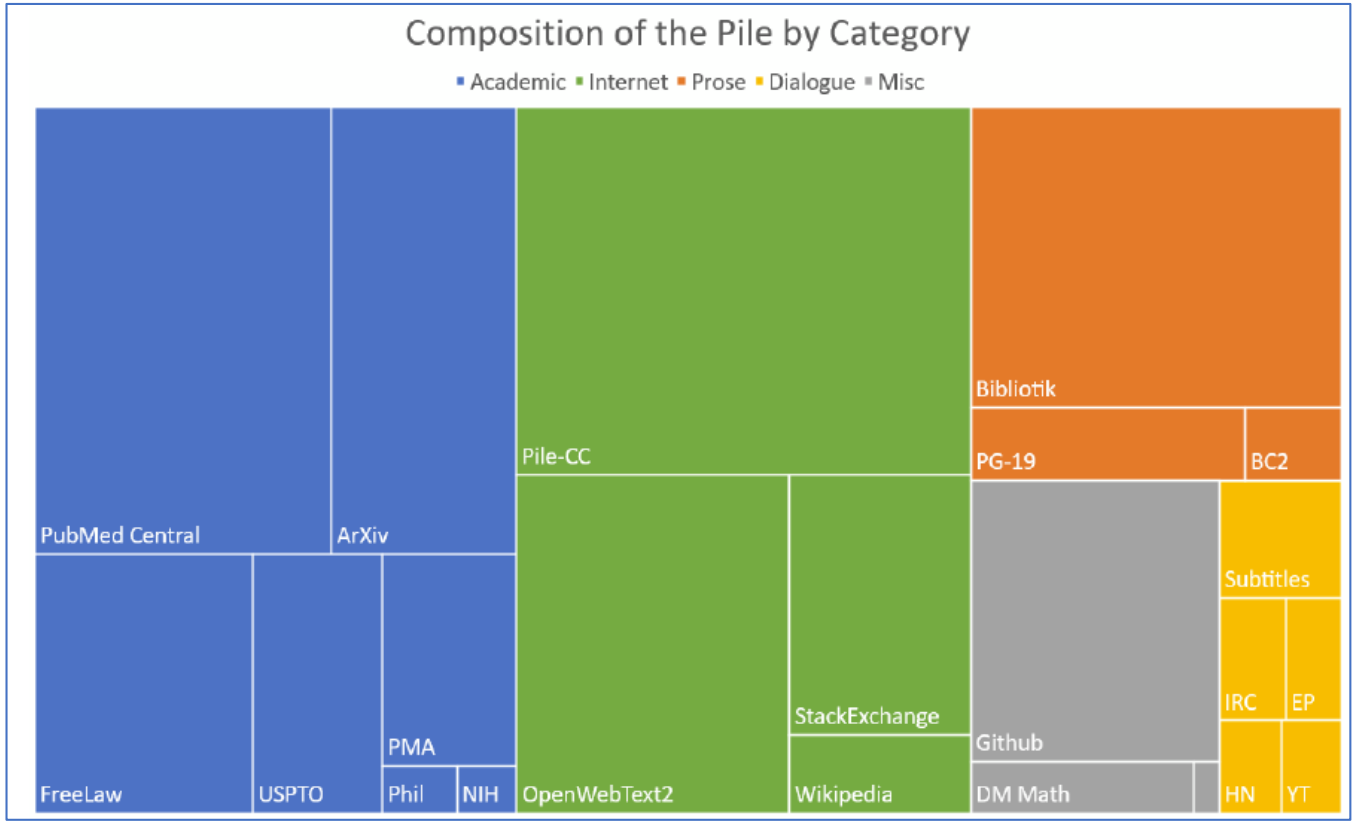
Microsoft Copilot

From Holmes:

Miss Doran whose graceful figure and striking face attracted much attention at the Westbury House festivities is an only child and it is currently reported **that her dowry will run to considerably over the six figures with expectancies for the future**. As it is an open secret that the Duke of Balmoral has been compelled to sell his pictures within the last few years and as Lord St. Simon has no property of his own save the small estate of Birchmoor it is obvious that the Californian heiress is not the only gainer by an alliance which will enable her to make the easy and common transition from a Republican lady to a British peeress.

LLM Training

- Scientific literature
- Wikipedia
- Patents
- Law cases
- Software
- Movie subtitles
- Most published books



The Pile: An 800GB Dataset of Diverse Text for Language Modeling

LLMs, knowledge, and mimicry

- LLMs are trained to mimic human language and programming languages
- Human language about history, philosophy, medicine, law
- Human language about how to do things: recipes, auto maintenance, plumbing, setting up your TV, investing in the stock market

- It can generate plausible answers to questions about all of these things, but these answers aren't necessarily correct

Shortcomings of LLMs

- Hallucination: The LLM can always generate an answer. It knows how court cases, medicines, and scientific papers look, and will generate fake ones easily

Decrypt

ChatGPT Wrongly Accuses Law Professor of Sexual Assault

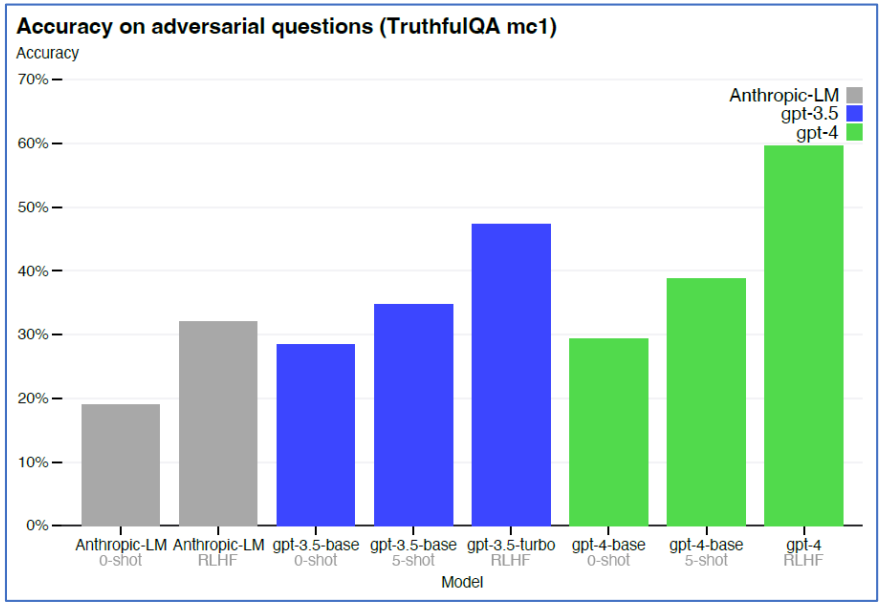
The chatbot says a prominent law professor committed sexual assault during a trip he never took.

By Jason Nelson

Apr 7, 2023

3 min read

GPT-4 Hallucination Rate is 40% on adversarial questions



GPT-4 Technical Report

Shortcomings of LLMs: Attribution

- Standard LLMs have no way of linking their output to the training documents that most influenced it
- Recent advances allow Microsoft and Google to sometimes give links to supporting documents

Shortcomings of LLMs: Popularity Bias

- In 2024, I asked ChatGPT about the pre-Columbian population of grazing animals in South America. It answered with the estimated human population
  - I suspect that is because asking about pre-Columbian human populations is much more common
- A common error by GPT-3 and GPT-4 was to answer the question “Which is larger, 9.5 or 9.11?” with the answer “9.11”.
  - It seems that because of popularity, it interpreted “9.11” as the date September 11, 2001, and this caused confusion

Key Point: Never Trust an LLM Output

- For anything important, always check the web pages or other links that Google and Bing provide

AI Systems Related to LLMs

https://this-person-does-not-exist.com

- Image and Video Synthesis
  - Generating fake images
  - Generating synthetic images of real people
    - Deep Fake Porn
    - Deep Fake Videos of politicians
- Speech synthesis
  - Synthetic voices language translation
  - Voice cloning

Problems Created by LLMs: Automated Mimicry

- Fake people
  - Fake customers clicking on ads, calling customer support, etc.
  - Phishing attacks through email, text messages, social media, phone calls
  - Voice cloning: fake kidnapping attacker allows you to talk with a cloned voice of your child
- Fake scientific papers
- Fake news stories
- Fake legal briefs
- Fake customer reviews
- Fake credit card receipts / travel expenses
- Fake homework essays

If we lower the cost of creating something, we lower the value of that thing

- “Sincere” email and text messages
- Music
- Video
- Personal physical appearance (e.g., on Instagram; dating sites)

Low Quality Software

- LLMs can write computer code
- This can be very helpful as a form of “auto-complete” or “auto-correct”
  - But the programmer needs to check it (of course)
- Studies show that LLMs often write buggy or insecure code
  - They have been trained on buggy and insecure code, so that is what they have learned to do

De-Skilling

- Students using LLMs to write their homework essays do not learn how to write or how to think carefully when forming an argument
- Scientists trusting an LLM to perform data analysis do not catch errors in their data and drawn incorrect conclusions
- Programmers trusting LLMs to write code do not learn to do it well
- What skills do we need to have as humans vs. what skills can we safely out-source to AI?
  - Only a few people need to know how to raise and train horses and use them for labor or transportation
  - Is software coding like horse raising?

Part 3: Visions of the AI future

- Tool AI
  - Help people make better decisions and create better designs, products, etc.
- Culture: knowledge, know-how
  - ChatGPT as an approximate knowledge base, as a replacement for Wikipedia?
- Communications medium
  - Language translation
- Teammate/Personal assistant
  - Human-AI teaming. Anthropomorphized tools
- Autonomous robots
  - Starship food delivery robots
  - Autonomous ships, cars, planes, weapons systems

Tool AI

- Many existing and anticipated applications of AI provide super-human tools
  - Knowledge-based Systems
  - Optimizing industrial processes (scheduling, logistics)
  - Highlighting suspicious cells in a cancer biopsy
  - Protein folding
  - Materials design

- If the shortcomings of LLMs can be eliminated, they may enable many more such tools

Culture: Knowledge and Know-How

- How to do anything
  - Cooking
  - Repair
  - Assembly
  - Maintenance
  - Leadership
  - HR functions
  - Raise horses
- Replacement/extension of Wikipedia, Wikihow, StackExchange
- Will LLMs replace the search engine + web as our primary way of interacting with the digital world?

AI as the Universal Communications Medium

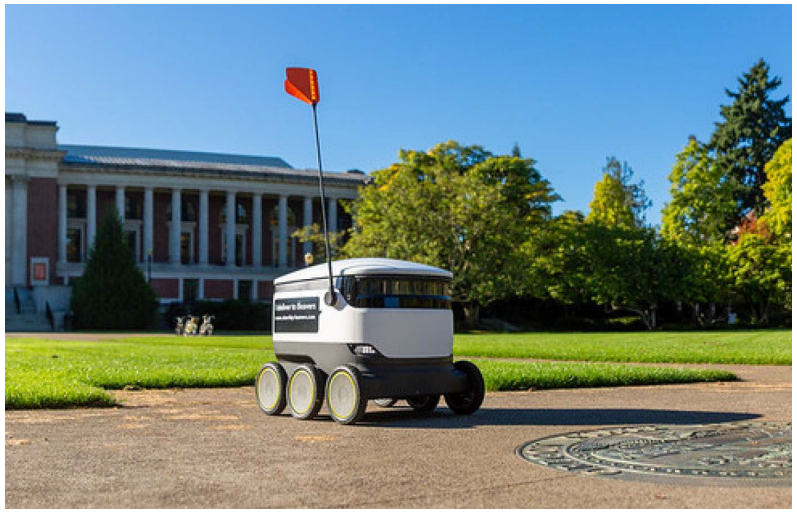
- Real-time conversational speech translation
  - “Babblefish”
- Language translation for text, movies, songs, etc.

AI as a personal assistant or teammate

- Personal assistant
  - Automatically pay your bills
  - Automatically re-order your prescriptions
  - Automatically reconcile and pay your medical bills
  - Automatically maintain your shopping list
  - Automatically shop for you, make restaurant reservations, schedule doctor's appointments
  - Help you plan trips
- Teammate
  - Autonomously carry out certain job functions
  - Advertising for positions
  - Screening job candidates
  - Annual performance reviews
  - Ordering office supplies
  - Scheduling meetings
  - Finding office space
  - Negotiating leases and insurance
  - How will companies be organized when there are AI teammates?

Autonomous Robots

- Starship food delivery robots
- Warehouse robotics
- Self-driving cars
- Self-flying planes
- Self-piloted ships
- Autonomous weapons



OSU

Agility Robotics

Autonomous Drones in Ukraine-Russia War

- Navigate to target and destroy
  - Kamikaze
  - Drop explosive
- Autonomous recognition of target
- Operate without radio contact
- We will soon live in a world where these are widely accessible to criminals and police as well as to militaries

Economic Considerations

- Effect of new tools/automation on employment
- Many job functions will be automated, but few total jobs will be automated
- What will be the new jobs? The new job skills?

LLMs and copyright

- Is training an AI model “fair use” under copyright law?
  - Argument in favor:
    - We are extracting general knowledge, just as a human would be reading the book
    - Copyright only protects the form and appearance of a work, not the knowledge contained in it
  - Argument against:
    - The systems can generate new works in the same style as the original author
    - Seems very close to forgery if such works are sold

Dangers of Anthropomorphizing

- Treating an AI system as human is unwise
  - Sets expectations for ethical behavior, social intelligence, self management
  - Blurs the line between humans and non-humans
- We love to anthropomorphize systems
  - “Alexa”, “Siri” are fake women
  - People claim psychological or emotional benefits from talking with chatbots
  - AI companions? AI therapists? AI boyfriends/girlfriends
  - Simulated AI doctors are “more empathetic”
  - AI-assisted suicide

Ethical Considerations

- Biases in decision making systems
  - AI based on machine learning picks up the statistics of the world
    - Doctors are men, Nurses are women, Rich people are white, etc.
    - Systems make more mistakes when trained on less data (minorities)
    - Predictive policing, insurance, loan approvals, etc.
- Moral responsibility for computer errors
  - Self-driving car
    - Occupant
    - Owner
    - Manufacturer
    - Software authors
    - The car itself as a legal entity?

Human dignity

- We need to draw a bright line between these simulated systems and human beings
  - Only people can be moral agents
  - Should only people be doing emotional work, psychotherapy, etc?

- Many arguments about the origin of moral responsibility
  - Systems that consider alternatives before taking action
  - “Consciousness”
  - “Free Will”
  - Capable of being punished; must feel pain
  - Current consensus is that no AI system is capable of being a moral agent

AI Bill of Rights

- Safe and effective systems
  - Requires testing, risk analysis, independent evaluation
- Protection against algorithmic discrimination
  - Pre-deployment evaluation of equity; oversight
- Privacy
  - Data should be collected only as needed for specific purposes
  - No continuous surveillance in education, work, housing, etc.
- Notice
  - You should know when and how automated systems are being used
- Right to a human alternative when automation fails
  - Burden for software failures should not fall on the public

Summary

- “AI” includes a wide variety of technologies and applications
  - Managing knowledge and information
  - Optimizing industrial processes
  - Sensing and communications
    - Language processing
    - Computer vision and image synthesis
    - Speech recognition and generation
  - Scientific applications
  - Playing games
  - Autonomous robots
- Large Language Models
  - Giant improvement on earlier language models
  - Mimic human language: hallucination, attribution, bias
- Causing New Problems
  - Fake people, papers, homework
  - Low quality software
  - De-skilling

Visions for the Future

- AI Roles
  - Tool AI
  - Library of all human knowledge
  - Universal communications medium
  - Personal assistant / Teammate
  - Autonomous robot
- AI Challenges
  - AI systems learn stereotypical biases
  - Effects on employment
  - Effects on human dignity
  - Only people can be moral agents
  - Need for an AI Bill of Rights for people

Disclaimer

- No AI was used to write this talk except
  - Generation of Sherlock Holmes example
  - Generation of fake person
  - Web searches from Bing Copilot and Google