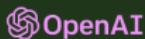


NLP and Text Analysis: Introduction

Daniel Hardt

CBS



ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

[TRY CHATGPT ↗](#)

ChatGPT 3.5 ▾



You

Explain the political system in Denmark



ChatGPT

Denmark has a parliamentary democracy with a constitutional monarchy.

The political system in Denmark is characterized by a separation of powers among the legislative, executive, and judicial branches. Here are key features of the political system:

1. **Constitutional Monarchy:** Denmark is a constitutional monarchy, meaning it has a monarch (currently Queen Margrethe II) as the head of state. However, the monarch's powers are largely symbolic, and executive authority is exercised by elected representatives.
2. **Parliament (Folketinget):** The Danish Parliament, known as the Folketinget, is the supreme legislative authority. It is a unicameral legislature with 179 members elected by proportional representation. Members of the Folketinget are elected for a four-year term.
3. **Executive Branch:** The executive branch is headed by the Prime Minister, who is the head of government. The Prime Minister is typically the leader of the political party that wins the most seats in the Folketinget. The Cabinet, appointed by the Prime Minister, consists of



You

Can you explain this in Danish, in the style of a rap song?



ChatGPT

Selvfølgelig! Lad mig give det en prøve på rapstil:

(Refrain)

I Danmark, hvor reglerne er klare,
En konstitutionel monarki uden fare.
Folketinget, vores parlament,
179 sæder, demokratisk engagement.

(Verse 1)

Vi har en dronning, symbolisk magt,
Men det er Folketinget, der fører os fremad.
Enkeltkammer-lov, 179 sæder stærk,
Valgt ved proportional representation, det' et værk.

(Chorus)

Parlamentarisk system, det' vores stil,
Socialdemokratiet, Venstre - spil for spill.
Koalitionsregering, det' ik' noget nyt,
Politisk samarbejde, det' vores liv.

(Verse 2)

Statsministeren, leder af nation,
Cabinet med ministre, stærk konnotation.
Retsystemet, uafhængigt og fri,
Tolker lovene, holder retfærdighed i live.

(Chorus)

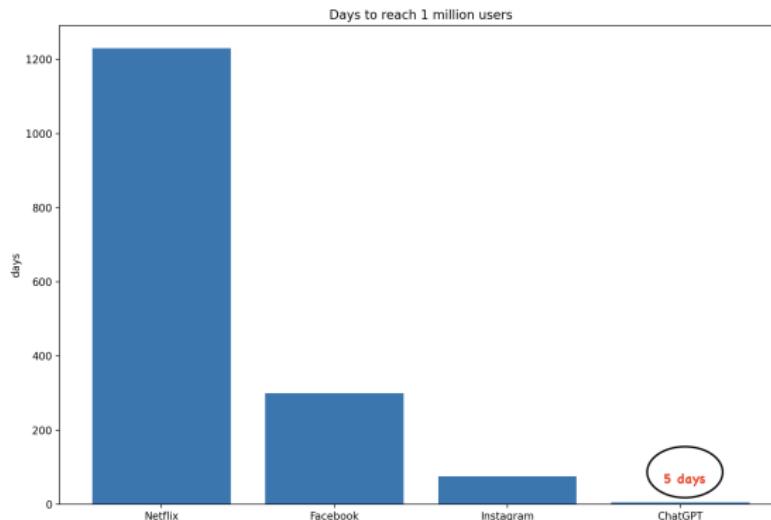
Multipartitsystem, vi har det her,
Enhedslisten, Radikale, det' klart.

ChatGPT: the Technology

- transformer model
- trained on missing word prediction
- Further training through Reinforcement Learning from Human Feedback

ChatGPT

November 2022



DeepSeek

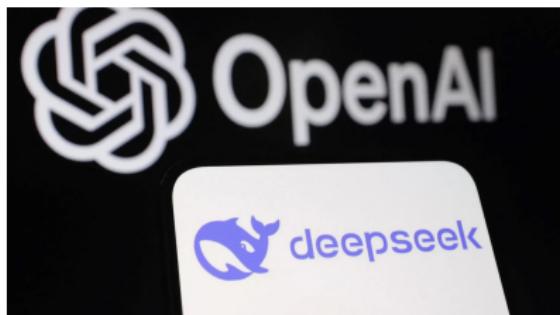
January 2025



TECH

China's DeepSeek AI dethrones ChatGPT on App Store: Here's what you should know

PUBLISHED MON, JAN 27 2025 10:13 AM EST | UPDATED MON, JAN 27 2025 4:05 PM EST



DeepSeek

DeepSeek-V3 Technical Report

DeepSeek-AI

research@deepseek.com

Abstract

We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens followed by Supervised Fine-Tuning and Reinforcement Learning stages to fully harness its capabilities. Comprehensive evaluations reveal that DeepSeek-V3 outperforms other open-source models and achieves performance comparable to leading closed-source models. Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for its full training. In addition, its training process is remarkably stable. Throughout the entire training process, we did not experience any irrecoverable loss spikes or perform any rollbacks. The model checkpoints are available at <https://github.com/deepseek-ai/DeepSeek-V3>.



Copenhagen
Business School
HANDELSHØJSKOLEN

DeepSeek

DeepSeek-V3 Technical Report

DeepSeek-AI

research@deepseek.com

Abstract

We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens, followed by Supervised Fine-Tuning and Reinforcement Learning stages to fully harness its capabilities. Comprehensive evaluations reveal that DeepSeek-V3 outperforms other open-source models and achieves performance comparable to leading closed-source models. Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for its full training. In addition, its training process is remarkably stable. Throughout the entire training process, we did not experience any irrecoverable loss spikes or perform any rollbacks. The model checkpoints are available at <https://github.com/deepseek-ai/DeepSeek-V3>.



Copenhagen
Business School
HANDELSHØJSKOLEN

DeepSeek

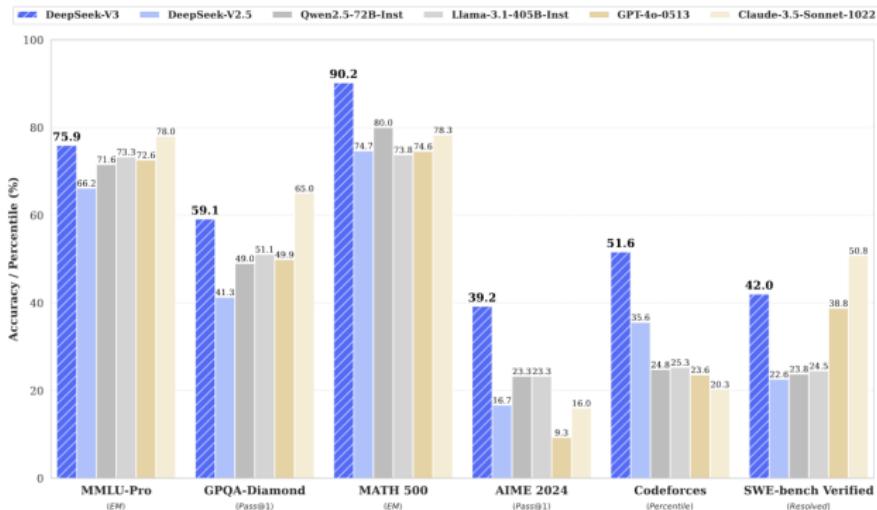


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

GPT4 Cost

Training Cost

The computational resources required to train GPT-4 are immense. The model's training involves approximately 2.15e25 floating-point operations, or FLOPs. A FLOP is a measure of computer performance, representing one floating-point operation per second. In the case of GPT-4, the model is performing an astronomical number of operations, highlighting the computational intensity of training such a large and complex AI model.

This training process was carried out on approximately 25,000 A100 GPUs over a period of 90 to 100 days. The A100 is a high-performance graphics processing unit (GPU) developed by NVIDIA, designed specifically for data centers and AI applications. It's worth noting that despite the power of these GPUs, the model was running at only about 32% to 36% of the maximum theoretical utilization, known as the maximum floating-point unit (MFU). This is likely due to the complexities of parallelizing the training process across such a large number of GPUs.

The financial cost of this training process is also significant. If we assume a cost of about \$1 per A100 GPU per hour in the cloud, the total cost for this training run would amount to approximately \$63 million. This figure underscores the substantial investment required to develop state-of-the-art AI models like GPT-4. It's a testament to the resources that organizations like OpenAI are willing to commit in their pursuit of advancing AI technology.

GPT4 Cost

Training Cost

The computational resources required to train GPT-4 are immense. The model's training involves approximately 2.15e25 floating-point operations, or FLOPs. A FLOP is a measure of computer performance, representing one floating-point operation per second. In the case of GPT-4, the model is performing an astronomical number of operations, highlighting the computational intensity of training such a large and complex AI model.

This training process was carried out on approximately 25,000 A100 GPUs over a period of 90 to 100 days. The A100 is a high-performance graphics processing unit (GPU) developed by NVIDIA, designed specifically for data centers and AI applications. It's worth noting that despite the power of these GPUs, the model was running at only about 32% to 36% of the maximum theoretical utilization, known as the maximum floating-point unit (MFU). This is likely due to the complexities of parallelizing the training process across such a large number of GPUs.

The financial cost of this training process is also significant. If we assume a cost of about \$1 per A100 GPU per hour in the cloud, the total cost for this training run would amount to approximately \$63 million. This figure underscores the substantial investment required to develop state-of-the-art AI models like GPT-4. It's a testament to the resources that organizations like OpenAI are willing to commit in their pursuit of advancing AI technology.

Language is Hard!

Current LLMs seem to have achieved human-level abilities to:

- process language
- reason using language

Is this really true?

- Many thinkers have argued that this could never happen

Descartes and AI



Rene Descartes: Discourse on the Method, Part V (1637)

Descartes and AI

Could a machine imitate a human?
No – you would always be able to tell the difference



Rene Descartes: Discourse on the Method, Part V (1637)

Descartes: Machines can't Imitate Humans

“...they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others.”

“...we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it ... But it never happens that it arranges its speech in various ways, in order to **reply appropriately to everything that may be said in its presence**, as even the lowest type of man can do.



Descartes: Machines can't Imitate Humans

“ . . . although machines can perform certain things as well as or perhaps better than any of us . . . they infallibly fall short in others . . . For while reason is a universal instrument which can serve for all contingencies, these organs have need of some special adaptation for every particular action.”

“ . . . it follows that it is morally impossible that there should be sufficient diversity in any machine to allow it to **act in all the events of life in the same way** as our reason causes us to act.”



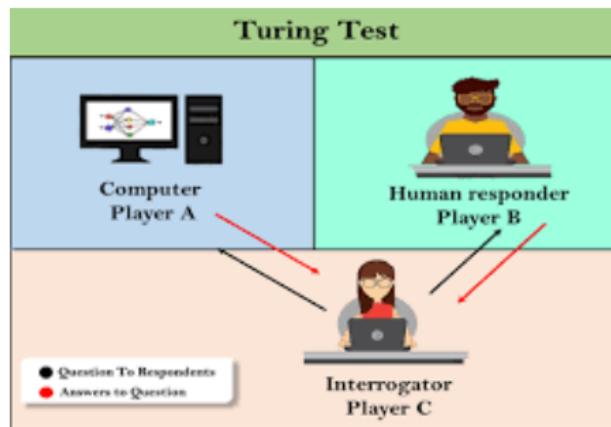
The Turing Test



(1950) Computing Machinery and Intelligence

- Test of a machine's ability to exhibit intelligent behaviour
- human judge engages in a natural language conversation with a human and a machine
- If the judge cannot reliably tell the machine from the human, the machine passes the test

The Turing Test



Why is Language Hard?

Why is Language Hard?

Language is infinite

Why is Language Hard?

Language is infinite

Most sentences you hear –

Why is Language Hard?

Language is infinite

Most sentences you hear –

- you have never heard them before,

Why is Language Hard?

Language is infinite

Most sentences you hear –

- you have never heard them before,
- and will never hear them again

Why is Language Hard?

Language is ambiguous

Why is Language Hard?

Language is ambiguous

Many words have multiple meanings

Why is Language Hard?

Language is ambiguous

Many words have multiple meanings

Lexical Ambiguity

Why is Language Hard?

Language is ambiguous

Many words have multiple meanings

Lexical Ambiguity

Phrases and sentence can have multiple meanings

Why is Language Hard?

Language is ambiguous

Many words have multiple meanings

Lexical Ambiguity

Phrases and sentence can have multiple meanings

Structural Ambiguity

Why is Language Hard?

Language is ambiguous

Many words have multiple meanings

Lexical Ambiguity

Phrases and sentence can have multiple meanings

Structural Ambiguity

A single sentence can have many different meanings

Why is Language Hard?

Language is ambiguous

Need Context to resolve ambiguities

Types of Data

- Numerical Data
- Categorical Data
- Text data is different
 - Content of an email
 - A Headline
 - Text of political speeches

Can text be treated as structured data?

Making Language into Structured Data

- The solution: Dummy values for **words**
 - One feature for each word
 - Value is 1 if word occurs in text, 0 otherwise
 - Alternative values: number of word occurrences, or TFIDF score
 - For any text, value of most features will be 0

Exploit Big Data for Text

- Use Supervised ML for text processing
- Can we get **labeled** text data?
- Build Classifiers
 - Spam Detection
 - Sentiment Analysis
 - Topic Detection
 - ...
- What does this have to do with AI? Real understanding?
Turing Test?



Treat text as a **Bag of Words**

Bag of Words Processing

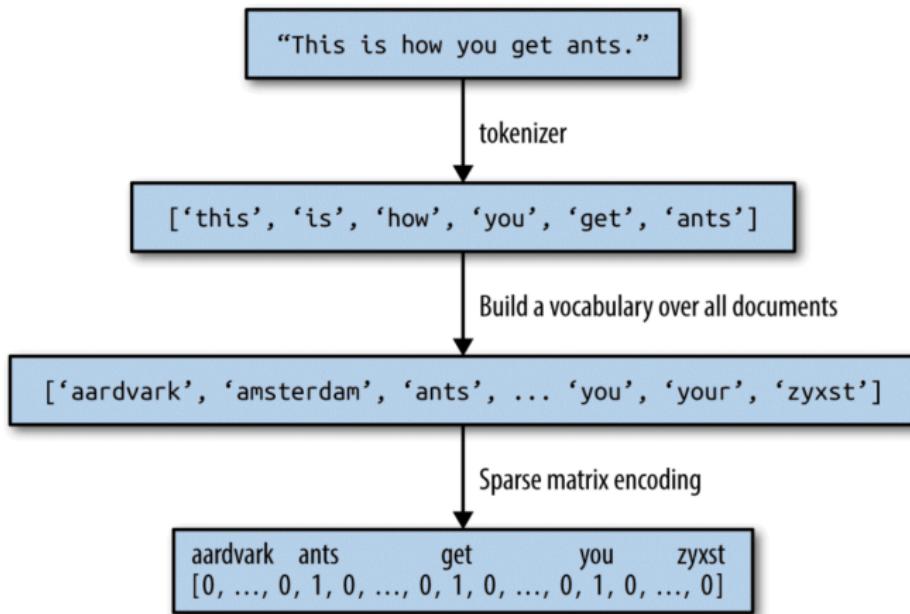


Figure 7-1. Bag-of-words processing

Markov Assumption

Simplifying assumption:

$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$



Andrei Markov

$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i \mid w_{i-k} \dots w_{i-1})$$

Approximate each component in the product

$$P(w_i \mid w_1 w_2 \dots w_{i-1}) \approx P(w_i \mid w_{i-k} \dots w_{i-1})$$

Language Modeling and Ngrams

Assign probability to a sequence of words

What is $p(\text{He went to the store})$?

Language Modeling and Ngrams

He went to the store

- 1-grams (unigrams): He, went, to, the, store (5)
- 2-grams (bigrams): He went, went to, to the, the store (4)
- 3-grams (trigrams): He went to, went to the, to the store (3)
- 4-grams: He went to the, went to the store (2)
- 5-grams: He went to the store (1)

Language Modeling and Ngrams

Bigram approximation:

$$p(\text{He went to the store}) = \\ p(\text{went}|\text{He}) * p(\text{to}|\text{went}) * p(\text{the}|\text{to}) * p(\text{store}|\text{the})$$

Trigram approximation:

$$p(\text{He went to the store}) = \\ p(\text{to}|\text{He went}) * p(\text{the}|\text{went to}) * p(\text{store}|\text{to the})$$

Language Models

- “A general language model (LM) should be able to compute the probability of (and also generate) any string.”
(Radford et al. 2019)

Guess the Next Word

As Descartes said, I think, therefore I ...
For all intents and ...
I learned how to drive a ...

Guess the Next Word

As Descartes said, I think, therefore I **am**
For all intents and ...
I learned how to drive a ...

Guess the Next Word

As Descartes said, I think, therefore I **am**
For all intents and **purposes**
I learned how to drive a ...

Guess the Next Word

As Descartes said, I think, therefore I **am**
For all intents and **purposes**
I learned how to drive a **car**

Guess the Next Word

Monica: Okay, everybody relax. This is not even a ...

Rachel: Oh God... well, it started about a half hour before the

...

Ross: (squatting and reading the instructions) I'm supposed to attach a brackety ...

Guess the Next Word

Monica: Okay, everybody relax. This is not even a **date**

Rachel: Oh God... well, it started about a half hour before the

...

Ross: (squatting and reading the instructions) I'm supposed to attach a brackety ...

Guess the Next Word

Monica: Okay, everybody relax. This is not even a **date**

Rachel: Oh God... well, it started about a half hour before the **wedding**

Ross: (squatting and reading the instructions) I'm supposed to attach a brackety ...

Guess the Next Word

Monica: Okay, everybody relax. This is not even a **date**

Rachel: Oh God... well, it started about a half hour before the **wedding**

Ross: (squatting and reading the instructions) I'm supposed to attach a brackety **thing**

The Training: Next Word Prediction

This is a pointless task

It's also impossible

But it's also genius

The Training: Next Word Prediction

- This is **Supervised Machine Learning** – each example is *labeled* with correct answer
- But no human labeling is needed – we just take text from the Internet, and hide a word
- Essentially unlimited amount of data

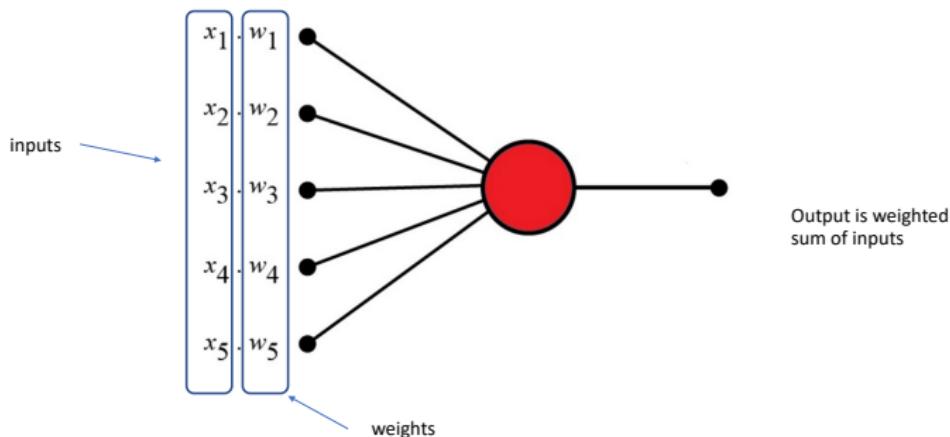
The Training: Next Word Prediction

System outputs a probability distribution for each word in vocabulary:

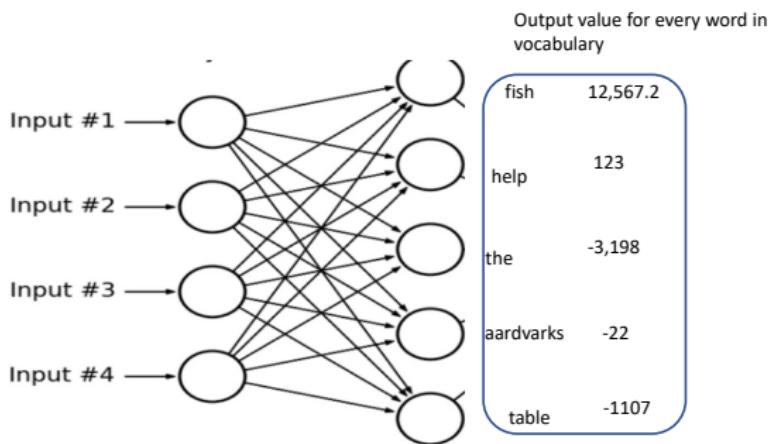
Thanks for all the [fish/.05, help/.04, support/.01, ...]

This talk is about [chatGPT/.02, AI/.019, business/.017, ...]

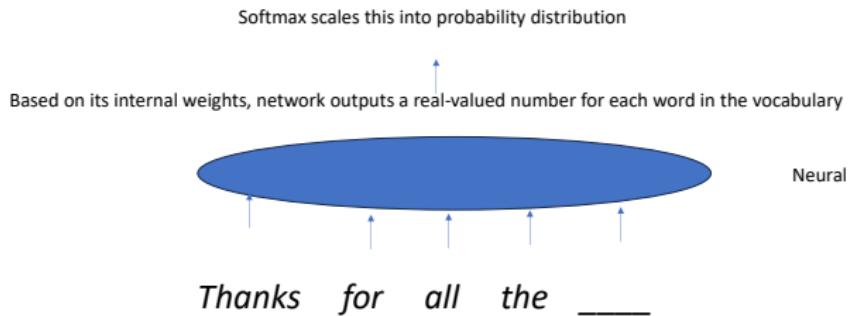
A Node of a Neural Network



A Neural Network for Word Prediction



The Training: Next Word Prediction



The Training: Next Word Prediction

- Let's say this is first sentence in training – so maybe all words get the same probability
- With, say 50,000 words in vocabulary, we would have output like this:

[fish/.00002, help/.00002, ..., the/.00002, ...aardvarks/.00002]

The Training: Next Word Prediction

- We compute a **loss function**
- Masked word was “fish”
- $\text{Loss} = 1 - \text{prob}(\text{"fish"})$
- We want to minimize loss
- So if model said $\text{prob}(\text{"fish"}) = 1$, loss would be 0
- If model said $\text{prob}(\text{"fish"}) = 0$, loss would be 1

Best possible

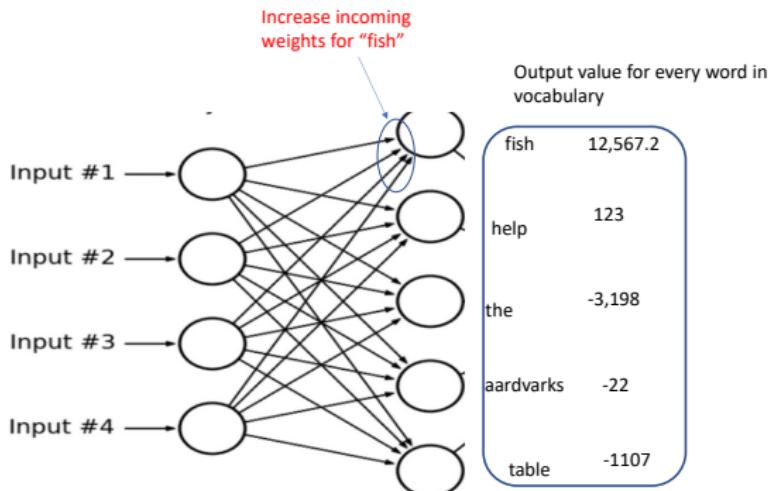
Worst possible

The Training: Next Word Prediction

Now, learning happens!

- Modify network weights to increase value computed for “fish”

A Neural Network for Word Prediction



Next Word Prediction: Learning

For this example:

- Decreases loss
- Increases probability for “fish”
- *Also, it will slightly reduce the probability for every other word in vocabulary*

The Training: Next Word Prediction

Each example produces a very small amount of useful information

- Make “fish” slightly more likely in the given context
- Also -- make every other word slightly less likely in this context

We repeat this an insane amount of times

Language Ability of GPT4

- LLMs (Large Language Models) today generate highly coherent, grammatical texts that can be indistinguishable from human output
- They exhibit at least some knowledge of hierarchical structure and abstract linguistic categories

(Mahowald et al. 2024)

Language Ability of GPT4

- These models are not perfect learners of abstract linguistic rules, but neither are humans
- LLMs are on track to acquiring formal linguistic competence
- LLMs have already overturned the claims about the fundamental impossibility of acquiring certain linguistic knowledge—including hierarchical structure and abstract categories—from the input alone

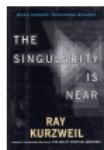
(Mahowald et al. 2024)

The Endgame

- Singularity
- AGI
- Powerful AI

The Singularity

- AI will pass the Turing Test by 2029
- The Singularity will be reached by 2045 – “human life will be irreversibly transformed . . . future machines will be human-like, even if they are not biological”



Ray Kurzweil (2005)

AGI

Artificial General Intelligence:

“matches or surpasses human cognitive capabilities across a wide range of cognitive tasks”

AGI: Tests

- **The Robot College Student Test (Goertzel):** A machine enrolls in a university, taking and passing the same classes that humans would, and obtaining a degree. LLMs can now pass university degree-level exams without even attending the classes.
- **The Employment Test (Nilsson):** A machine performs an economically important job at least as well as humans in the same job. AIs are now replacing humans in many roles as varied as fast food and marketing.



AGI: Tests

- **The Ikea test (Marcus):** Also known as the Flat Pack Furniture Test. An AI views the parts and instructions of an Ikea flat-pack product, then controls a robot to assemble the furniture correctly.
- **The Coffee Test (Wozniak):** A machine is required to enter an average American home and figure out how to make coffee: find the coffee machine, find the coffee, add water, find a mug, and brew the coffee by pushing the proper buttons. This has not yet been completed.
- **The Modern Turing Test (Suleyman):** An AI model is given \$100,000 and has to obtain \$1 million

Machines of Loving Grace

Dario Amodei



Machines of Loving Grace¹

How AI Could Transform the World for the Better

October 2024

I think and talk a lot about the risks of powerful AI. The company I'm the CEO of, Anthropic, does a lot of research on how to reduce these risks. Because of this, people sometimes draw the conclusion that I'm a pessimist or "doomer" who thinks AI will be mostly bad or dangerous. I don't think that at all. In fact, one of my main reasons for focusing on risks is that they're the only thing standing between us and what I see as a fundamentally positive future. I think that most people are underestimating just how radical the upside of AI could be, just as I think most people are underestimating how bad the risks could be.



Copenhagen
Business School
HANDELSHØJSKOLEN

Machines of Loving Grace

Dario Amodei



Machines of Loving Grace¹

How AI Could Transform the World for the Better

October 2024

I think and talk a lot about the risks of powerful AI. The company I'm the CEO of, Anthropic, does a lot of research on how to reduce these risks. Because of this, people sometimes draw the conclusion that I'm a pessimist or "doomer" who thinks AI will be mostly bad or dangerous. I don't think that at all. In fact, one of my main reasons for focusing on risks is that they're the only thing standing between us and what I see as a fundamentally positive future. I think that most people are underestimating just how radical the upside of AI could be, just as I think most people are underestimating how bad the risks could be.

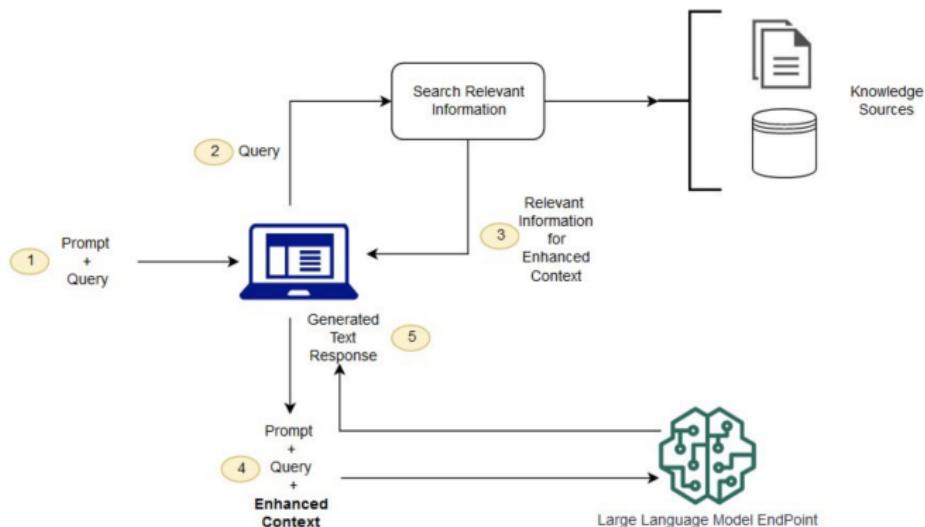
Machines of Loving Grace

- “Powerful AI”
- Could come by 2026
- Smarter than a Nobel Prize winner across multiple fields
- All the interfaces – text, audio, video, internet access
- ...
- “a country of geniuses in a datacenter”

LLM Techniques

- Prompt Engineering
- Few-shot learning
- RAG

RAG



Application Examples

- RAG interface for
 - customer support
 - onboarding
- CBS Research portal with RAG
- Zendesk – customer support
- BeMyAI

The Revolution in NLP
Language is Hard
NLP: The Basic Approach
AI – Where are We Heading?
Applications of LLMs – Project Ideas
Takeaways

BeMyAI

Now available

Be My AI™
AI-powered visual
assistance.

Be My AI™ integrates with Microsoft



Blind YouTube vlogger and TV presenter Lucy Edwards introduces Be My AI™.

Takeaways

- Language is Hard!
 - Language is infinite
 - The challenge of ambiguity – both lexical and structural

Takeaways

- The Revolution in NLP
- Current LLM's
 - Appear to have (close to) human-level language ability
 - Exciting research directions!
 - Building applications exploiting LLM's
 - Probing their abilities
 - Powerful AI (Machines of Loving Grace) is coming soon