



Kunstig intelligens, maskinlæring, ChatGPT: Hvordan fungerer språkmaskineriet?

Stephan Oepen (Institutt for informatikk)

Real Undervisning – January 11, 2024



For å introdusere meg selv ...



Default (GPT-3.5)

OE

What does Stephan Oepen at the University of Oslo work with?



As of my last knowledge update in September 2021, Stephan Oepen was a computer scientist and researcher at the University of Oslo. He is known for his work in the field of natural language processing (NLP) and computational linguistics. Specifically, he has been involved in research related to deep learning, parsing, and machine translation. However, please note that his research interests and projects may have evolved or changed since then. To get the most up-to-date information on Stephan Oepen's current research focus and projects, I recommend visiting the official website of the University of Oslo's Department of Informatics or checking his personal webpage and recent publications.



For å introdusere meg selv ...



Default (GPT-3.5)

OE

Hva er det som Stephan Oepen ved UiO jobber med?



Per min siste oppdatering i september 2021, var Stephan Oepen en forsker ved Institutt for lingvistiske og nordiske studier ved Universitetet i Oslo (UiO). Han har en doktorgrad innen datalingvistik og har vært involvert i flere forskningsprosjekter innenfor dette feltet.



Stephan Oepen har spesialisert seg på flere områder innenfor datalingvistik, inkludert formell grammatikk, syntaktisk parsing, dyp læring og natursspråksbehandling. Han har vært med på å utvikle verktøy og teknikker for å analysere og forstå syntaktisk struktur i naturlige språk, noe som har applikasjoner innen maskinoversettelse, tekstgenerering, og annen NLP-relatert forskning.

Det er mulig at det har skjedd endringer i hans forskningsfokus eller stilling etter september 2021. Jeg vil anbefale deg å besøke universitetets nettsider eller kontakte dem direkte for å få den mest oppdaterte informasjonen om Stephan Oepens nåværende forskning og aktiviteter ved UiO.

ChatGPT: Hva menes med “hallusinering”?



OE

Does he have any scientific publications of relevance?



Stephan Oepen has numerous scientific publications in the field of natural language processing (NLP) and computational linguistics. Some of his publications have been influential in the NLP research community. Since my knowledge is up to date only until September 2021, I cannot provide you with the most recent publications. However, you can find his publications on academic databases and websites such as Google Scholar, ResearchGate, or the University of Oslo's website. These sources will provide you with a comprehensive list of his relevant scientific publications, including research papers, articles, and conference presentations.



ChatGPT: Hva menes med "hallusinering"?



OE

Har han skrevet noen artikler?



Ja, Stephan Oepen har publisert flere forskningsartikler innenfor området datalingvistikk.

Hans arbeid har vært fokusert på formell grammatikk, syntaktisk parsing og natursspråksbehandling. Noen av hans tidligere artikler inkluderer:



1. **"Towards Incremental Deep Parsing" (2004)** - En artikkel som utforsker metoder for syntaktisk parsing av naturlig språk ved hjelp av dyp læringsteknikker.
2. **"LinGO Redwoods: A Rich and Dynamic Treebank for HPSG" (2005)** - Her beskriver Oepen et prosjekt kalt LinGO Redwoods, som er en rik syntaktisk trebank basert på Head-Driven Phrase Structure Grammar (HPSG).
3. **"Wide-Coverage Semantic Representations from a CCG Parser" (2006)** - Dette arbeidet involverer semantisk representasjon av naturlig språk basert på Combinatory Categorical Grammar (CCG) parsing.
4. **"The Scream Parser: An Open-Source Broad-Coverage Dependency Parser" (2009)** - En artikkel som presenterer Scream-parseren, en åpen kildekodeavhengighetsparser for naturlig språk.

ChatGPT: Hva menes med "hallusinering"?



OE

Har han skrevet noen artikler?



Ja, Stephan Oepen har publisert flere forskningsartikler innenfor området datalingvistikk. Hans arbeid har vært fokusert på formell grammatikk, syntaktisk parsing og natursspråksbehandling. Noen av hans tidligere artikler inkluderer:



1. **"Towards Incremental Deep Parsing" (2004)** - En artikkel som utforsker metoder for syntaktisk parsing av naturlig språk ved hjelp av dyp læringsteknikker.
2. **"LinGO Redwoods: A Rich and Dynamic Treebank for HPSG" (2005)** - Her beskriver Oepen et prosjekt kalt LinGO Redwoods, som er en rik syntaktisk trebank basert på Head-Driven Phrase Structure Grammar (HPSG).
3. **"Wide-Coverage Semantic Representations from a CCG Parser" (2006)** - Dette arbeidet involverer semantisk representasjon av naturlig språk basert på Combinatory Categorical Grammar (CCG) parsing.
4. **"The Scream Parser: An Open-Source Broad-Coverage Dependency Parser" (2009)** - En artikkel som presenterer Scream-parseren, en åpen kildekodeavhengighetsparser for naturlig språk.

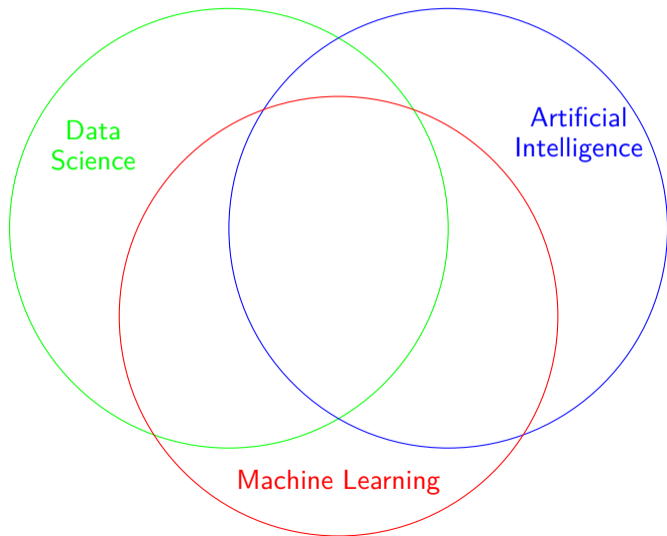
Safari 5G 06:55 98 %

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.



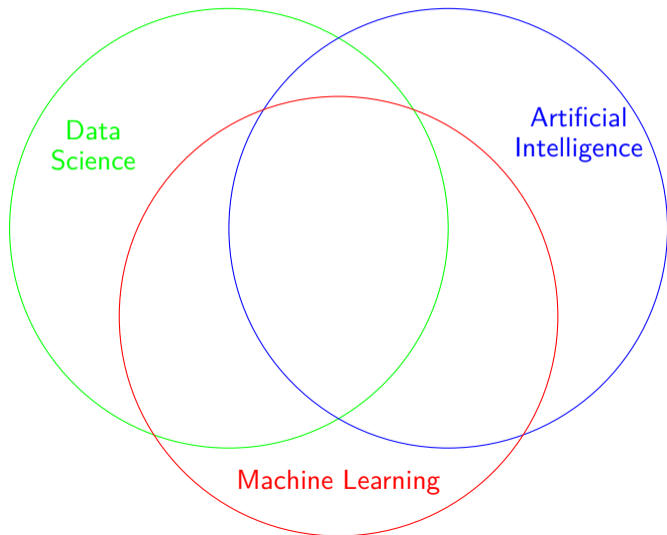
Én mulig tilnærming til begresavklaring



Én mulig tilnærming til begresavklaring

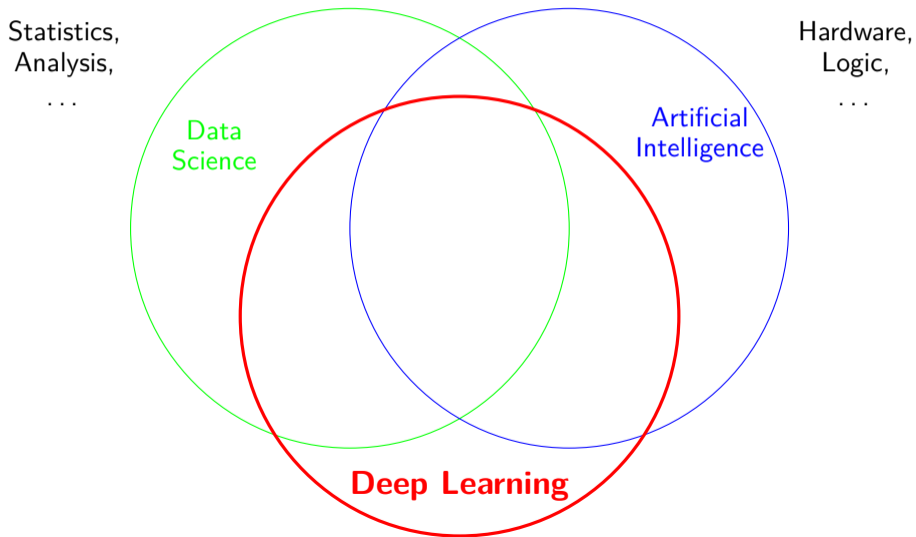


Statistics,
Analysis,
...



Hardware,
Logic,
...

Én mulig tilnærming til begresavklaring





Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai 'Ngaje Ngai,' the House of God.

Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called 'Ngaje Ngai' in Masai, the house of God.



Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai 'Ngaje Ngai,' the House of God.
(Ernest Hemingway)

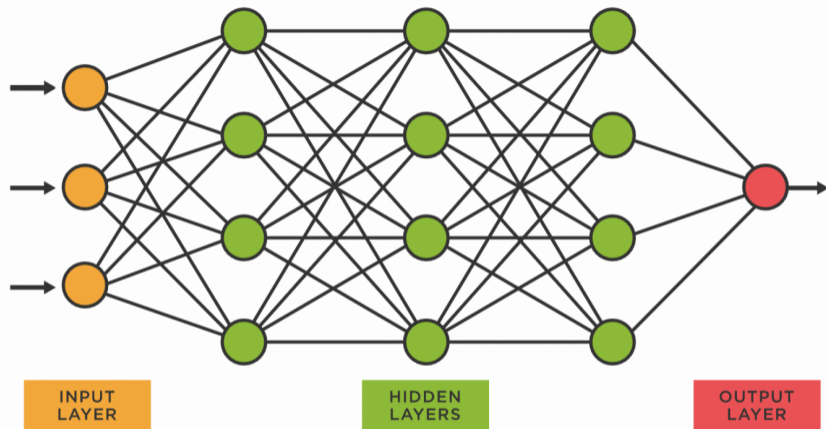
Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called 'Ngaje Ngai' in Masai, the house of God.
(Google Translate, November 2016)



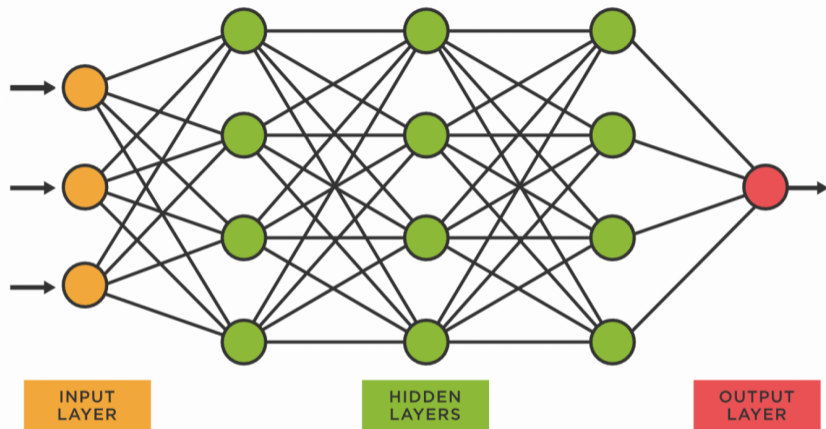
Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai 'Ngaje Ngai,' the House of God.
(Ernest Hemingway)

Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called 'Ngaje Ngai' in Masai, the house of God.
(Google Translate, November 2016)

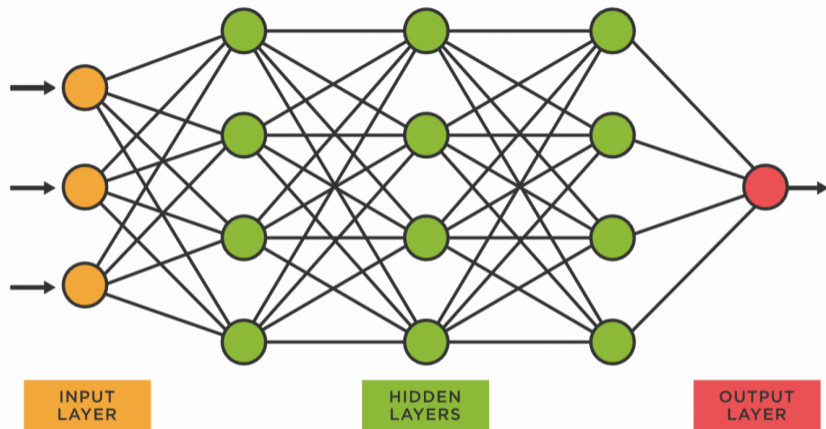
Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, 'Ngaje Ngai' in the Maasai language, has been referred to as the house of God.
(Google Translate, October 2016)



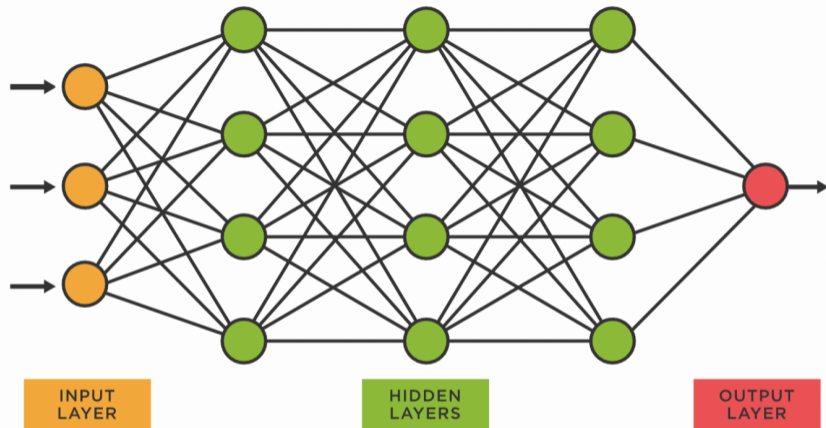
- ▶ Artificial (deep) **neural networks**: millions of units \rightarrow large-scale linear algebra;



- ▶ Artificial (deep) **neural networks**: millions of units \rightarrow large-scale **linear algebra**;

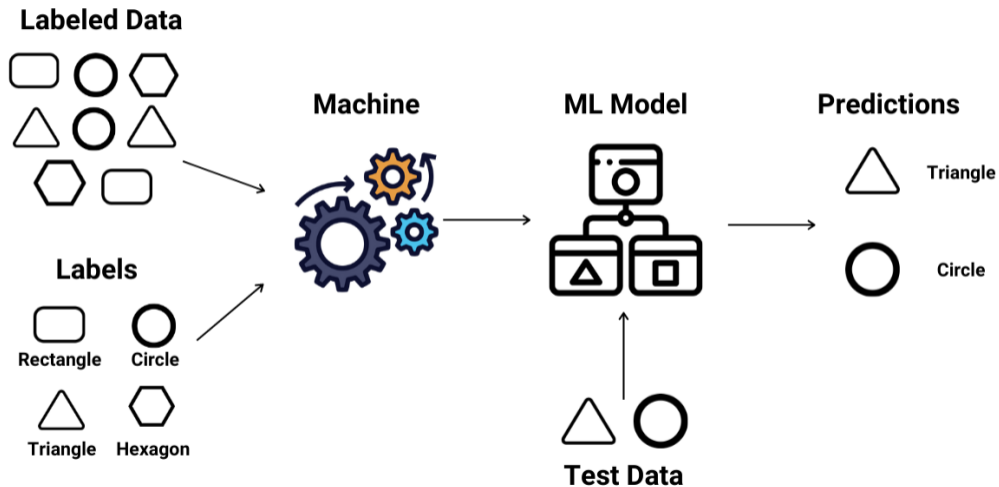


- ▶ Artificial (deep) **neural networks**: millions of units \rightarrow **large-scale** linear algebra;



- ▶ Artificial (deep) **neural networks**: millions of units \rightarrow **large-scale linear algebra**;
- ▶ abstract goal: learn **correspondence** relation between **inputs** x_1^n and **outputs** y_1^n .

Hvordan lærer en datamaskin? «Veiledet læring»



Hvordan lærer en datamaskin om språk?





Department of Informatics

← Research ← Research groups ← Language Technology Group (LTG)

Research seminar

Teaching and Supervision

Office Impressions

Master's seminar

NLP for Norwegian

Publications

Teaching and Supervision

The Language Technology Group is responsible for the study program in Informatikk: språkteknologi ("Informatics: language technology"). This is one of five bachelor programs in Computer Science, and it is taught in Norwegian. The group is also responsible for the English-taught master's program in Informatics: Language Technology. Furthermore, courses taught by LTG staff (on search, language technology, and machine learning) are part of other Computer Science programs, as well.

Group members supervise a variety of topics and kinds of MSc projects. A list of candidate projects is available on-line, but we will also be happy to discuss individual ideas and define new projects in the broad realm of theoretic and applied language technologies. Please make contact with us.

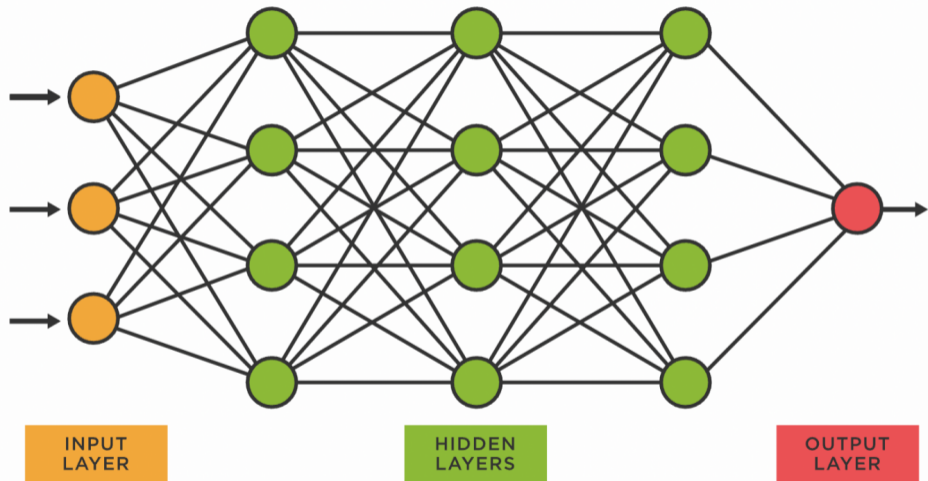


The Language Technology Group
is responsible for the
study in
Informatikk: språkteknologi.
This is one of five
bachelor programs in Computer Science,
and it is in Norwegian.

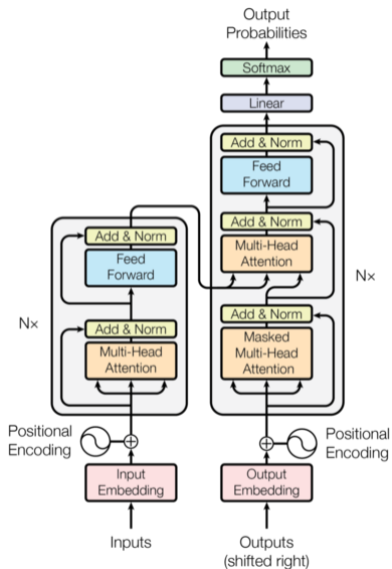


The Language Technology Group
is responsible for the
study **program** in
Informatikk: språkteknologi.
This is one of five
bachelor programs in Computer Science,
and it is **taught** in Norwegian.

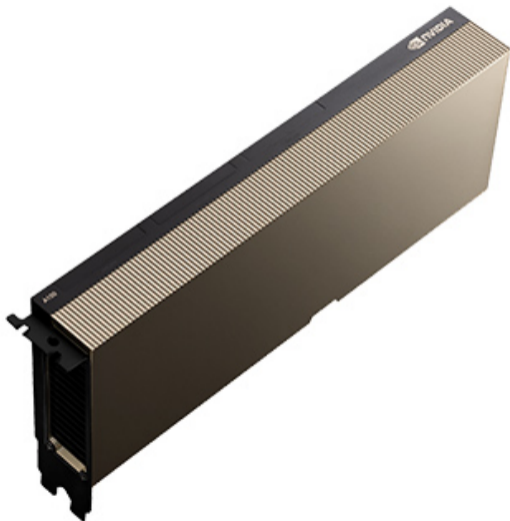
Trening (Optimisering): Finne gode parameterer (vektor)



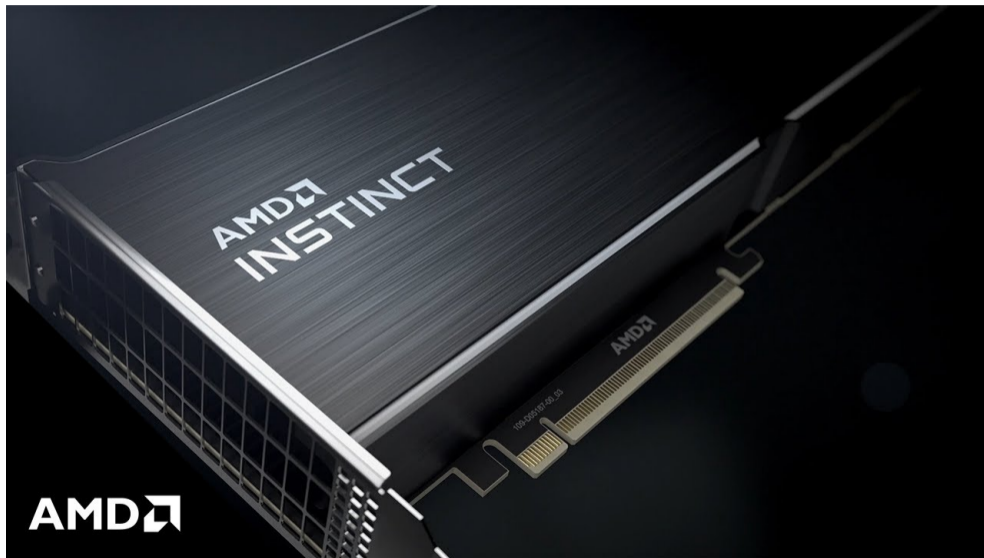
Hvordan er det virkelig? *Transformers og Attention*



Dyp læring er anvendt lineær algebra: kører på GPU-er



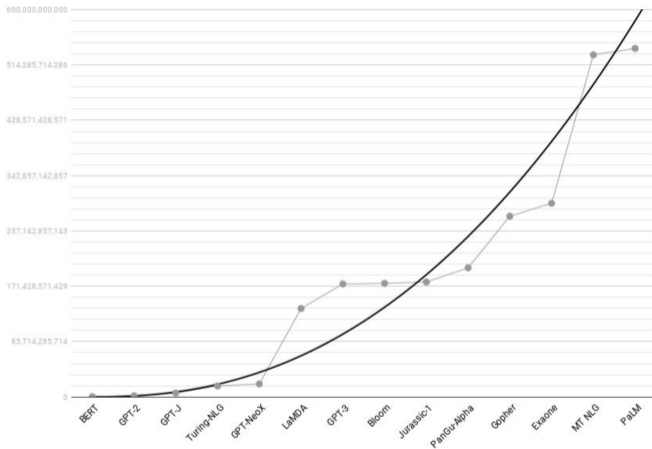
Dyp læring er anvendt lineær algebra: kører på GPU-er



Utviklingen den siste tiden er drevet av oppskalering



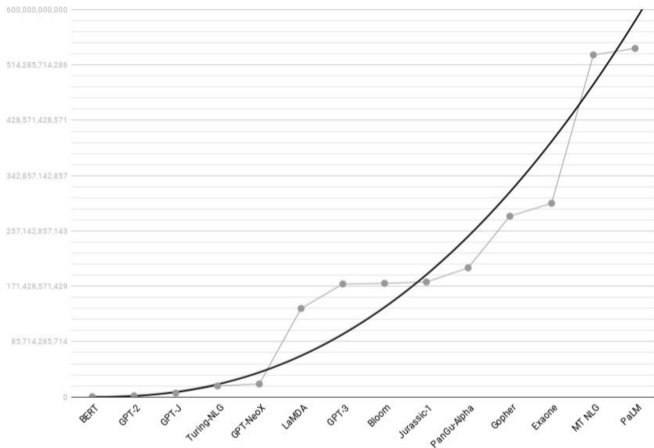
Transformer model parameters



Utviklingen den siste tiden er drevet av oppskalering



Transformer model parameters

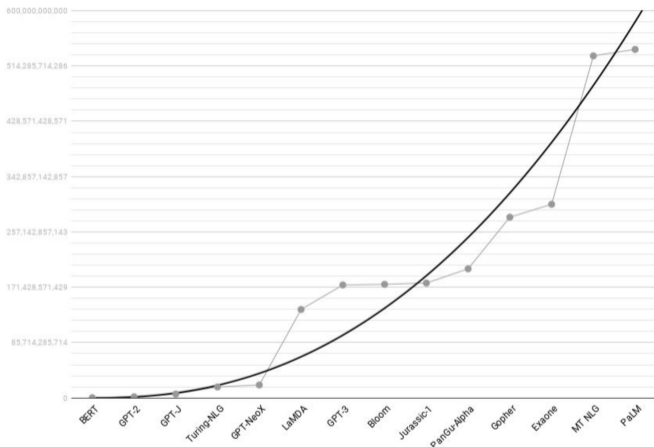


PETABYTES OF TRAINING DATA AND CENTURIES OF SPECIALIZED COMPUTE

Utviklingen den siste tiden er drevet av oppskalering



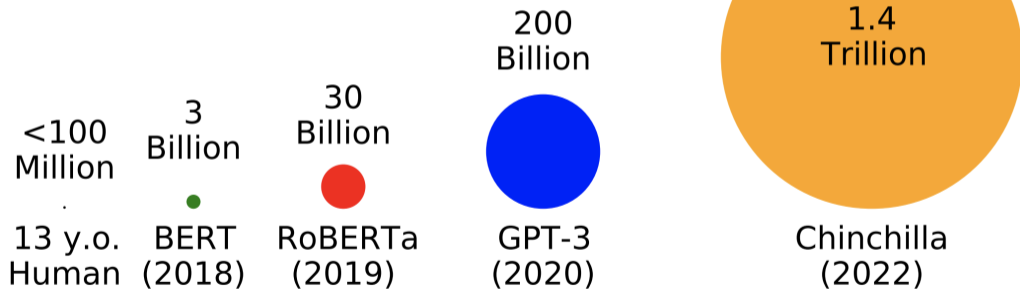
Transformer model parameters



PETABYTES OF TRAINING DATA AND CENTURIES OF SPECIALIZED COMPUTE

600 BILLION PARAMETERS \equiv 140 TIB OF MEMORY; AROUND 1750 GPUS

Oppskalering av treningsdata: antall "opplevde" ord







LUMI: The 3rd Fastest Supercomputer World-Wide in June 2023 (20,000 GPUs)

- ▶ Vast **storage** and **compute** available to Norwegian researchers (and collaborators);
- ▶ Fox (UiO) ... Betzy, Saga (Sigma2) ... LUMI: path to **scaling up incrementally**.

Hvordan bli en samtalepartner? *Instruction Fine-Tuning*



Instruction finetuning

Please answer the following question.
What is the boiling point of Nitrogen?

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

Language model

-320.4F

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$.

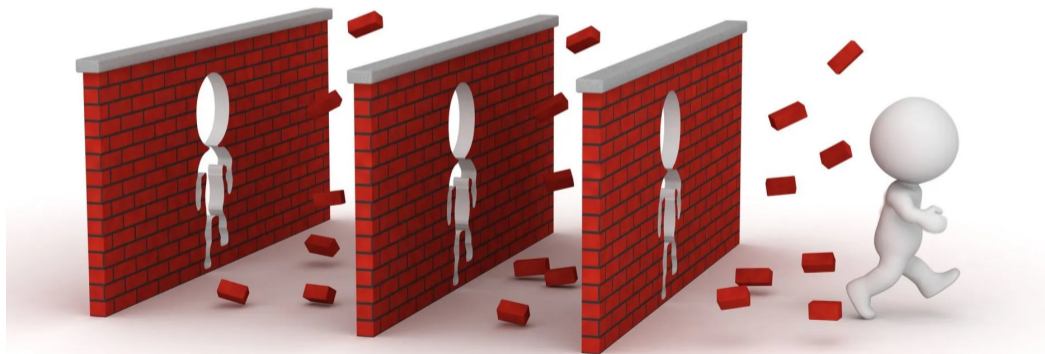
Multi-task instruction finetuning (1.8K tasks)

Inference: generalization to unseen tasks

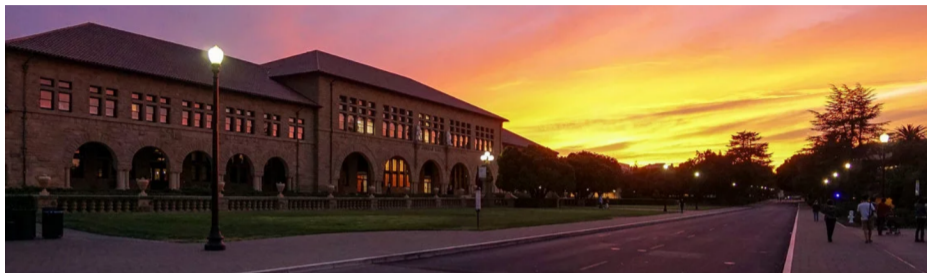
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Hva slags gjennombrudd er det egentlig?



Bruk av verktøy er et spørsmål om akademisk kultur



The Honor Code is an undertaking of the Stanford academic community, individually and collectively. Its purpose is to uphold a culture of academic honesty.

Students will support this culture of academic honesty by neither giving nor accepting unpermitted academic aid in any work that serves as a component of grading or evaluation, including assignments, examinations, and research.

Instructors will support this culture of academic honesty by providing clear guidance, both in their course syllabi and in response to student questions, on what constitutes permitted and unpermitted aid. Instructors will also not take unusual or unreasonable precautions to prevent academic dishonesty.



Stephan Oepen

Instituttleder

`oe@ifi.uio.no`