

NorwAI

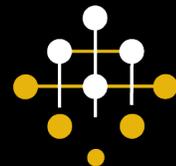
«Empowering News Summarization with Pre-trained Language Models»

Lemei Zhang & Peng Liu

May 2023

NorwAI

Norwegian Research Center
for AI Innovation



NTNU

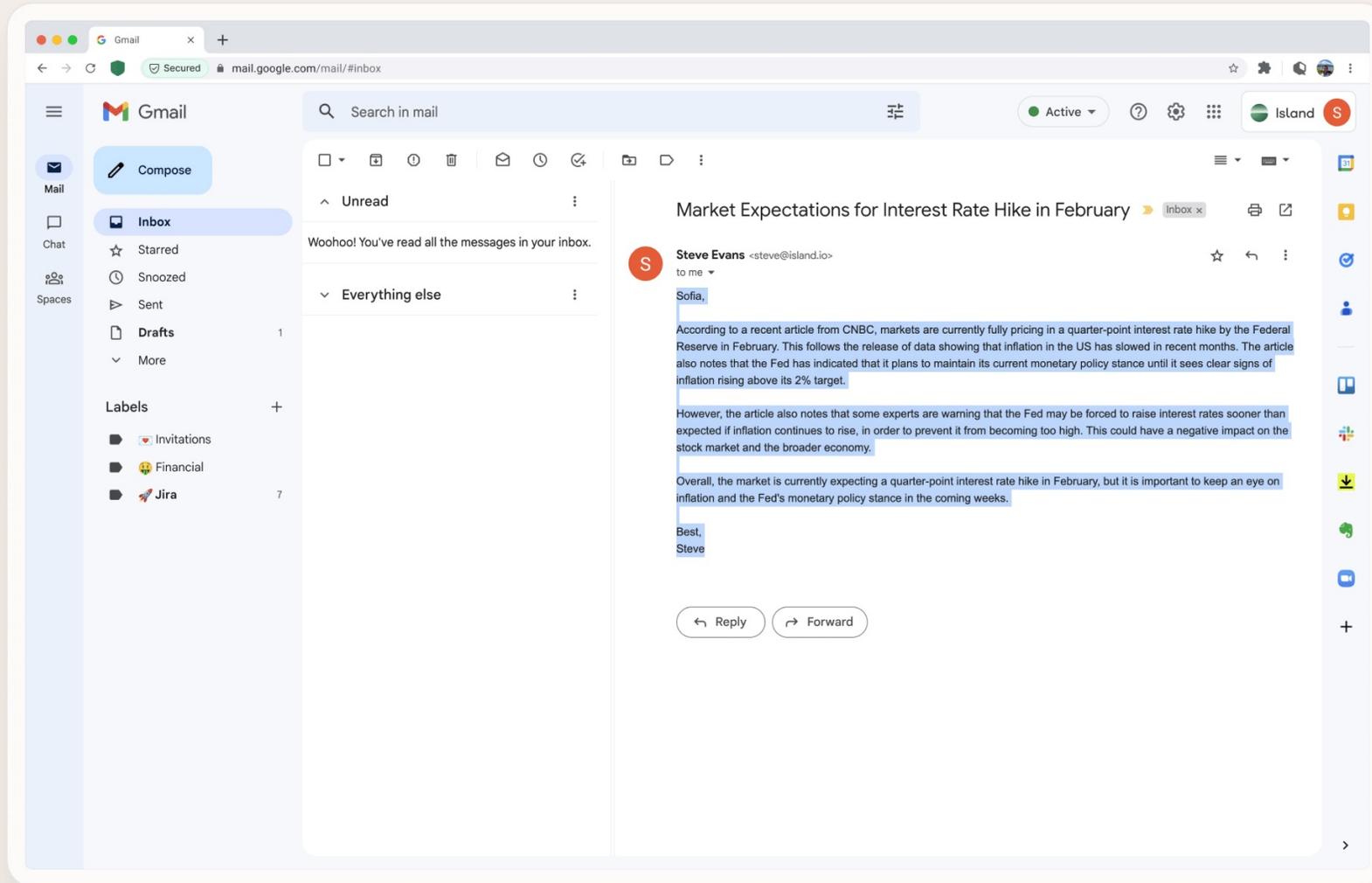
sfi Centre for
Research-based
Innovation

The Research Council of Norway

Outline

- 1 **Generative Language Models for Norwegian**
- 2 **News Summarization with Instruct Tuning**
- 3 **What's next?**







Welcome to the new Bing

Your AI-powered copilot for the web

🗨️ Ask complex questions

"What are some meals I can make for my picky toddler who only eats orange-colored food?"

🔍 Get better answers

"What are the pros and cons of the top 3 selling pet vacuums?"

🎨 Get creative inspiration

"Write a haiku about crocodiles in outer space in the voice of a pirate"

Let's learn together. Bing is powered by AI, so surprises and mistakes are possible. Make sure to check the facts, and [share feedback](#) so we can learn and improve!

[Terms of Use](#) | [Privacy Statement](#)

Choose a conversation style [Preview](#)

More
Creative

More
Balanced

More
Precise

No problem, I'm glad you enjoyed our previous conversation. Let's move on to a new topic. What do you want to know more about?



How do I set achievable goals?

Which is deepest ocean?

How many hours are in a year?



Ask me anything...

0/2000

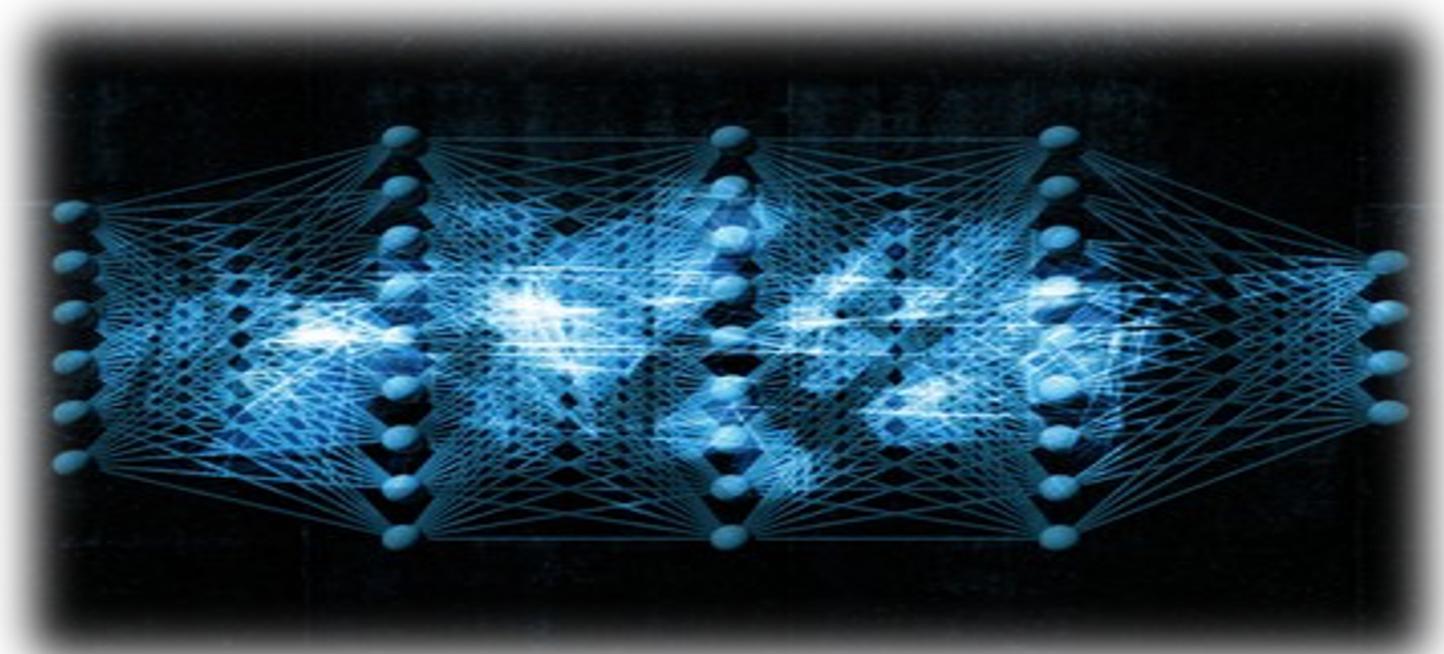


What is a Language Model?

Gigantic unlabelled training data sets

45 TB dataset

198 GB dataset



Billions of parameters in transformer model

175 billion parameters

23 billion parameters

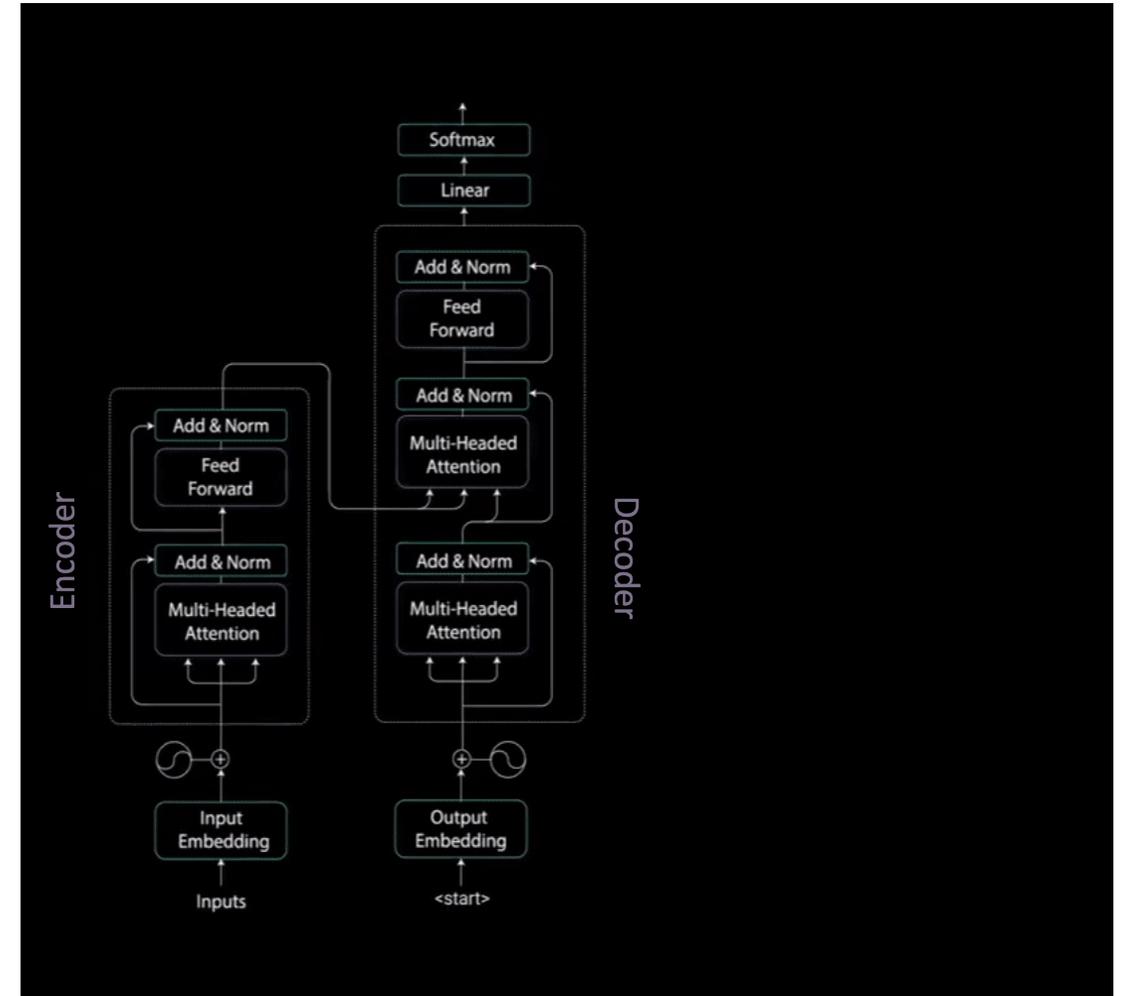
Contextualized understanding of words of a language

OpenAI GPT-3
NorwAI GPT-23B



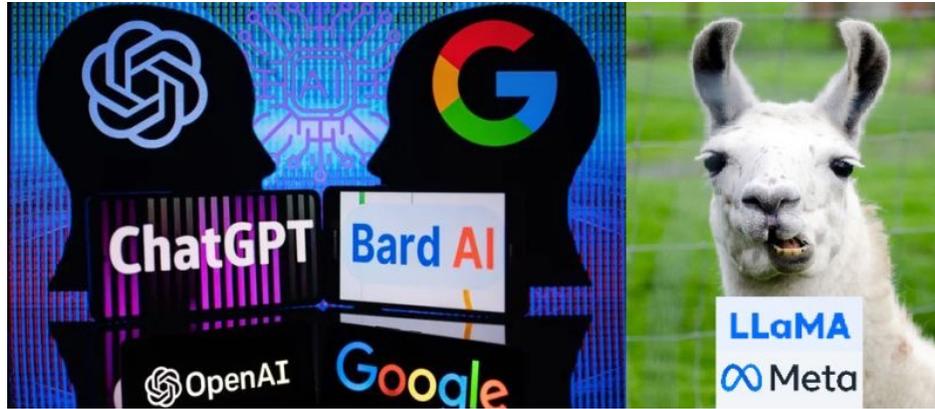
Transformer

- Encoder-Decoder model based on (self-) attention mechanism
- Regardless of sequence length, capture long-term dependencies between input and output and allow more parallelization.
- Basically targeting for translation tasks (WMT'14 EN-GE, EN-FR), it achieved state of the art BLEU score.



Vaswani, Ashish, et al.(2017). *Attention is all you need*. *Advances in neural information processing systems*, 30.

Motivation for a large Norwegian GPT Model



Large tech companies, typically based in the United States, are the main driving forces behind commercial LLMs.



Data Protection and Privacy

GPT-3 Training Dataset Statistics

language	number of words	percentage of total words
en	181014683608	92.64708%
fr	3553061536	1.81853%
de	2870869396	1.46937%
es	1510070974	0.77289%
it	1187784217	0.60793%
pt	1025413869	0.52483%
nl	669055061	0.34244%
ru	368157074	0.18843%
ro	308182352	0.15773%
pl	303812362	0.15550%
fi	221644679	0.11344%
da	221551540	0.11339%
sv	220920577	0.11307%
ja	217047918	0.11109%
no	212193299	0.10860%
zh	193517396	0.09905%

Limited support for low-to-medium resource languages



Generative Language Models for Norwegian

Challenges:

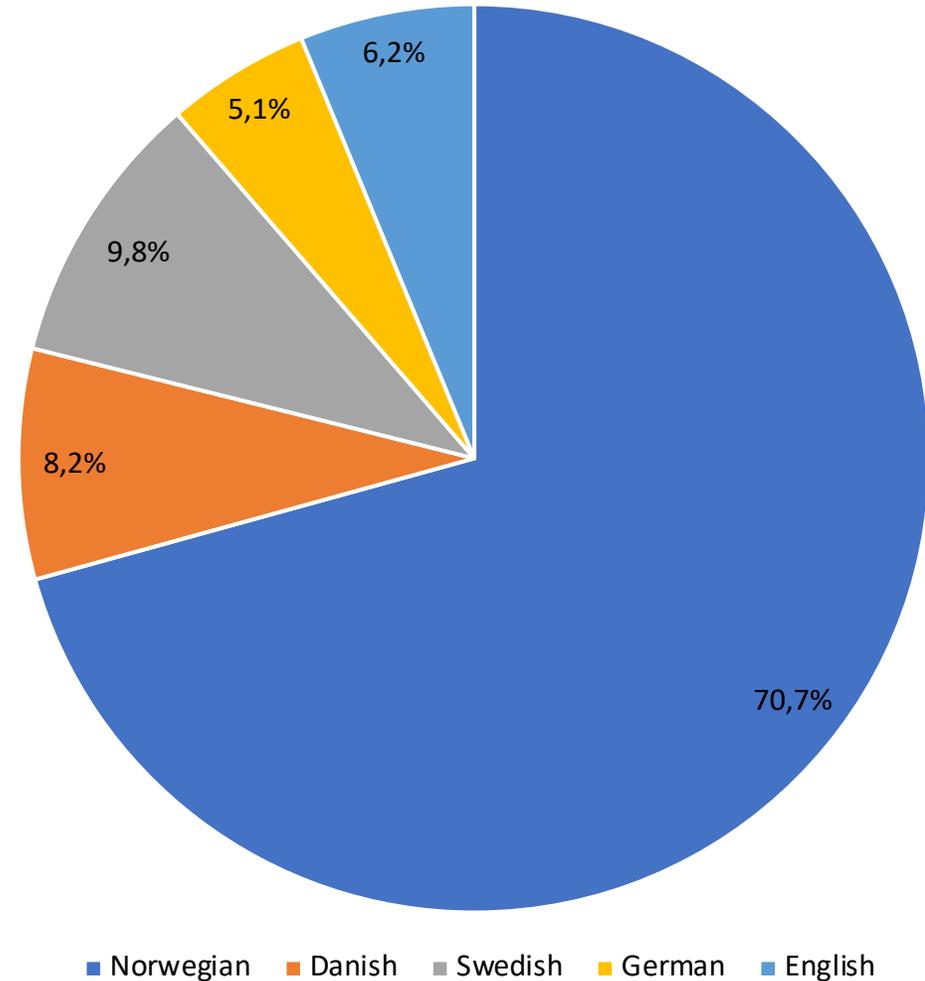
- Limited availability of Norwegian Text
- Finding sufficient computational resources (Idun)
- Implementing a software stack to make the model accessible

Goal: Develop a GPT model that can compete for Norwegian 🇳🇴



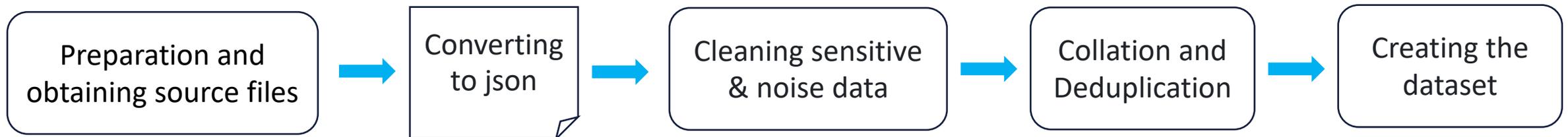
NorwAI GPT-23B Training Corpora

- Norwegian data (Dataset size: 139.82 GB)
 - Norwegian mC4 web-crawled corpus: 96.71 GB (9.5 billion tokens)
 - OSCAR web-crawled corpus: 4.96 GB (787 million tokens)
 - Schibsted: 6.9 GB (1.08 billion tokens)
 - National Library dataset: 29.45 GB (4.49 billion words)
 - Social media: Norwegian tweets (602 MB, 180.49 million tokens)
Norwegian redds (500 MB, 86.74 million tokens)
- Danish Common Crawler Dataset: 16.18 GB (2.5 billion tokens)
- Swedish Common Crawler Dataset: 19.43 GB (2.9 billion tokens)
- German Common Crawler Dataset: 10 GB (1.4 billion tokens)
- English Common Crawler Dataset: 12.37 GB (2 billion tokens)



NorwAI GPT-23B Training Corpora

- Data Pre-processing
 - Data clean pipeline
 - Removing sensitive information (username, email, url, etc.)
 - Removing emoji
 - Removing special characters & symbols
 - Deduplicating within & across the corpuses



NorwAI GPT Language Model Training

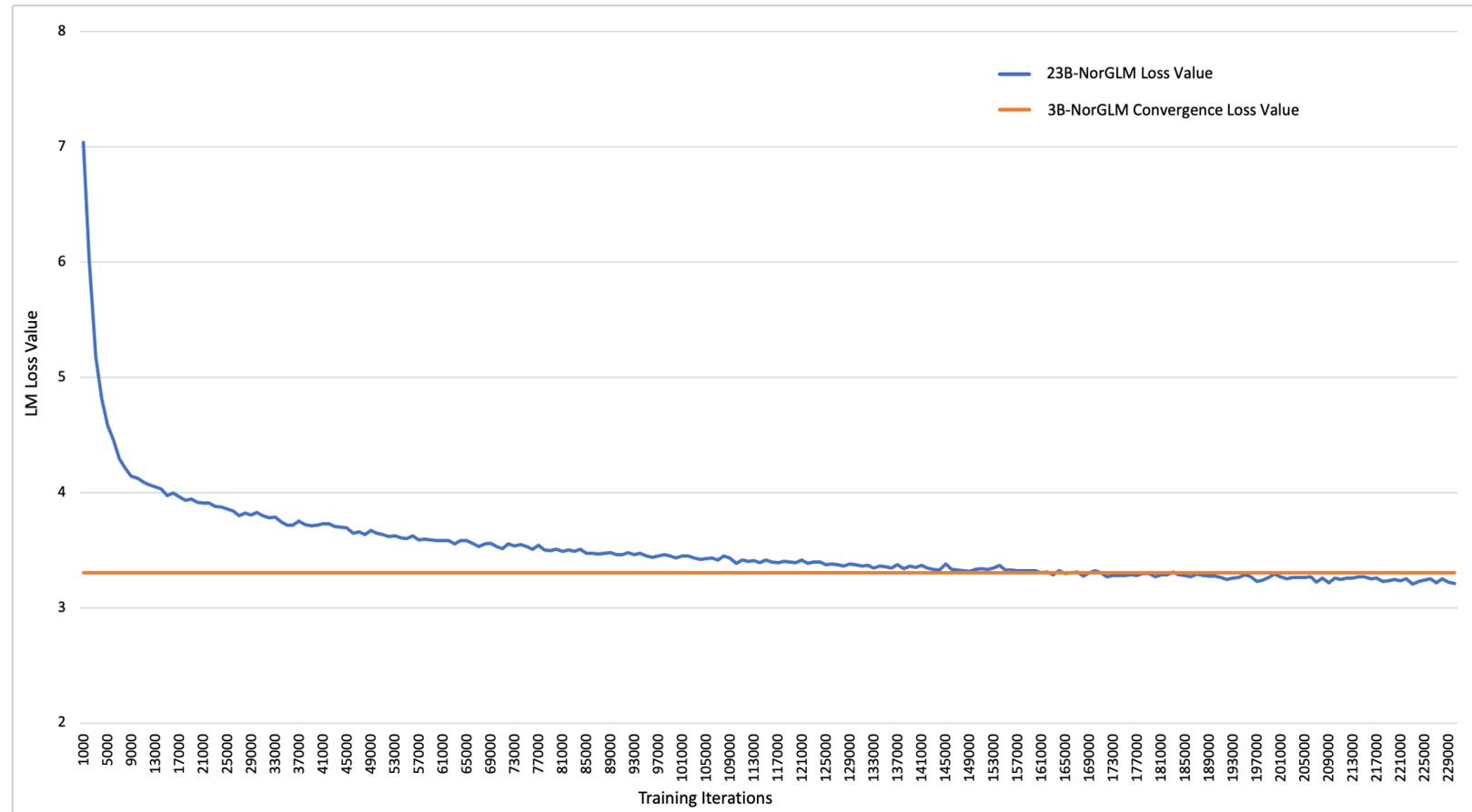
- NorwAI GPT Models

	NorGPT-318M	NorGPT-3B	NorGPT-23B	NorGPT-40B
#Parameters	318.6 Million	2.95 Billion	23.03 Billion	40 Billion
#Layers	24	32	49	?
#Attention heads	16	32	64	?
#Hidden unit size	1024	2688	6144	?
Context window	1024	2048	2048	?
Training data size	13 GB	196 GB	196 GB	> 196 GB
Vocabulary size	30,000	64,000	64,000	?
Convergence time	1 week	4 weeks	3-4 months	?



NorwAI GPT Language Model Training

- Training Setting:
 - 28 x NVIDIA A100 SXM 80GB GPUs
 - 20 TB Memory
 - GPT-2 implementation from Microsoft/Megatron-DeepSpeed
- Checkpoint of around (Huggingface version):
 - 1.4 GB (318M Model)
 - 6.2 GB (3B Model)
 - 93 GB (23B Model)



Benchmarking

- NorwAI is developing a comprehensive **benchmark set for Norwegian generative language models**
 - Translation from English supervised datasets
 - Collection of manually annotated datasets
 - Still work in progress

Conversational agent	Norwegian Persona-Chat dataset (26.3 MB)
Text summarization	Norwegian CNN/Dailymail (532 MB)
	Norwegian News Summaries (in progress)
Question answering	Norwegian Free-Style QA (in progress)

Title: Tror du europeisk kaffe er fra Europa? Da tar du feil

Newsroom: byas

Creation_date: 2022-06-01T09:48:02Z

Body: Det er mange myter og misforståelser om kaffe.Norge er i verdenstoppen i kaffedriking i forhold til folketal. Vi elsker kaffe. I tillegg til å gi en oppvakkende effekt, går også kaffedrikkere 1000 skritt mer, ifølge en studie Trd.by har omtalt.Samtidig sover de i snitt 18 minutter mindre, på den negative siden.Når kaffe er så populært og vanlig i det norske hjem, kan det fort også oppstå noen myter og misforståelser. Det amerikanske nettstedet Wired har tatt en gjennomgang av fem av dem:Nei, det er faktisk et frø. Til å begynne med er frøet pakket inn i en tynn rød frukt som skrelles under rengjøringsprosessen. Da er den en lys sølvgrønn farge inntil den er stekt. Siden bønnene vi maler opp og brygger, er stekte, er de heller ikke plantebare lenger. Det ville uansett tatt umåtelig lang tid, ifølge Wireds skribent Jaina Grey.Kaffebønner vokser nemlig ikke i Europa. De vokser i Afrika, den arabiske halvøy, Asia og Hawaii. Med andre ord er kaffe importert fra for eksempel Italia og Frankrike dårligere enn om den ble importert fra for eksempel Etiopia. Det er fordi den best smakende kaffen alltid brennes kort tid før den konsumeres. Smakene i brennt kaffe brytes nemlig fort ned. Derfor er rådet for best mulig kaffe å kjøpe lokalt brennt kaffe og male dem selv hjemme.Det er ikke nødvendigvis sant, det er mer som en godt stekt pizza. Den er bare mer brennt, og smaken er sterkere. Lysbrent kaffe blir ikke brennt like lenge, og av den grunn brytes ikke koffeinen nødvendigvis like mye ned under prosessen. Derfor kan lysere kaffe gjerne ha mer koffein.På norsk blir dette forvirrende. Til nå har vi skrevet «brent kaffe» som i «roasted coffee». Nå snakker vi om brennt kaffe som i «burned», og Starbucks-kaffen kan visstnok oppfattes slik.Det stemmer ikke. Kaffen er mørkbrent, altså «dark roasted», og det er fordi den som alle andre store fast food-sjapper, ønsker å ha et produkt som smaker akkurat likedan uansett hvor du kjøper den. Med mørkbrent kaffe er det enklere å få til. Ifølge Wireds skribent er koffeinfri kaffe urettferdig utskjelt. Du er liksom ikke ekte fan av kaffe om du forfekter den uten koffein. Men om den utbrente kaffebønna vaskes på en måte som gjør at koffeininnholdet trekkes ut før brenning, skal dette være en skånsom metode som bevarer smakene til kaffen. Rådet er derfor å se etter på pakken om den koffeinfrie kaffen er bearbeidet med vann. Da skal det ikke stå noe i veien på smaken.

Category: Aktuelt

Question-Answering

Question-1: Hvor vokser kaffebønner?

Answer-1: Kaffebønner vokser i Afrika, den arabiske halvøy, Asia og Hawaii.
Source-1: De vokser i Afrika, den arabiske halvøy, Asia og Hawaii.

Question-2: Hvorfor bør man kjøpe lokalt brennt kaffe?

Answer-2: Lokalt brennt kaffe bør kjøpes fordi den best smakende kaffen alltid brennes kort tid før den konsumeres. Smakene i brennt kaffe brytes fort ned, så for å få best mulig kaffe bør man kjøpe lokalt brennt kaffe og male den selv hjemme.
Source-2: Det er fordi den best smakende kaffen alltid brennes kort tid før den konsumeres. Smakene i brennt kaffe brytes nemlig fort ned. Derfor er rådet for best mulig kaffe å kjøpe lokalt brennt kaffe og male dem selv hjemme.

Question-3: Hva er en negativ konsekvens av å drikke kaffe?

Answer-3: Ifølge en studie sover kaffedrikkere i snitt 18 minutter mindre enn personer som ikke drikker kaffe.
Source-3: Samtidig sover de i snitt 18 minutter mindre, på den negative siden.

Question-4: Hva skjer under rengjøringsprosessen av kaffebønnen?

Answer-4: Under rengjøringsprosessen blir den tynne røde frukten som kaffefrøet er pakket inn i skrelt av.
Source-4: Til å begynne med er frøet pakket inn i en tynn rød frukt som skrelles under rengjøringsprosessen.

Question-5: Hva er en skånsom metode for å lage koffeinfri kaffe som bevarer smakene til kaffen?

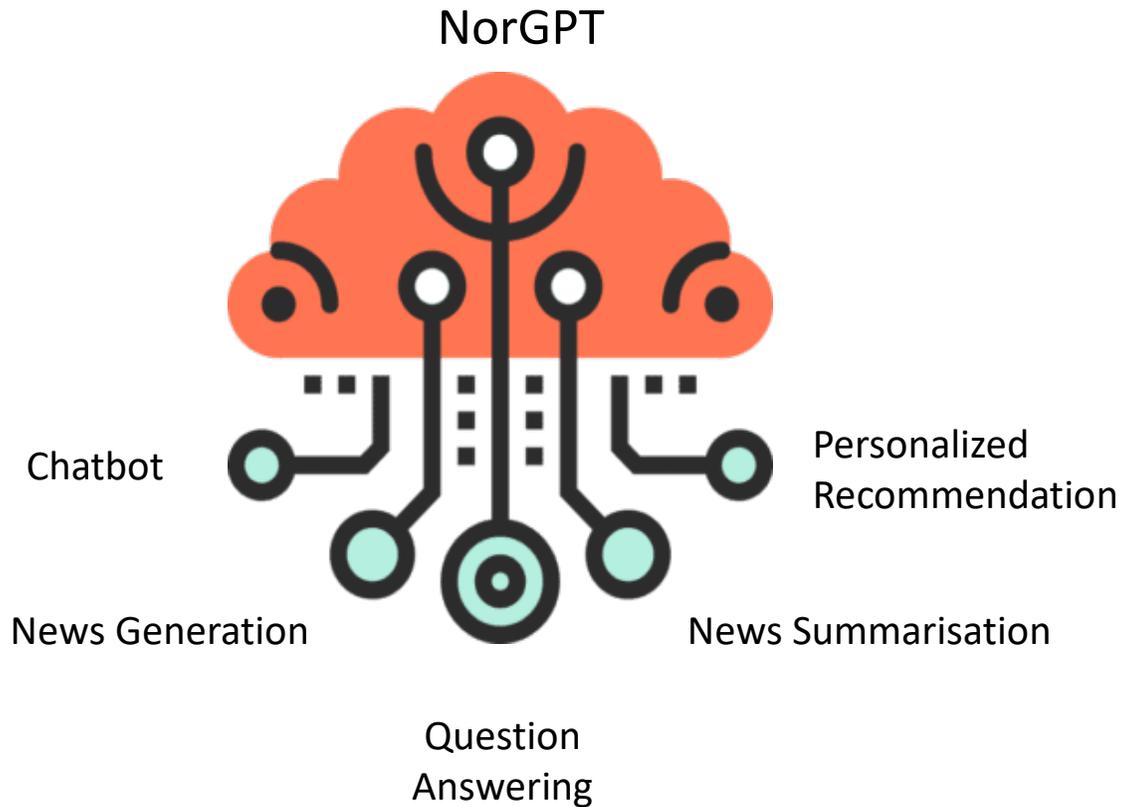
Answer-5: En skånsom metode for å lage koffeinfri kaffe er å vaske kaffebønnene på en måte som trekker ut koffeininnholdet før brenning.
Source-5: Men om den utbrente kaffebønna vaskes på en måte som gjør at koffeininnholdet trekkes ut før brenning, skal dette være en skånsom metode som bevarer smakene til kaffen.

News summarization

Summary: Artikkelen handler om myter og misforståelser om kaffe, blant annet om koffeinfri kaffe og prosessen for å få kaffebønner. I artikkelen gis det også råd for å få best mulig kaffe.
Sum-source: Det er mange myter og misforståelser om kaffe. Ifølge Wireds skribent er koffeinfri kaffe urettferdig utskjelt. Til å begynne med er frøet pakket inn i en tynn rød frukt som skrelles under rengjøringsprosessen. Derfor er rådet for best mulig kaffe å kjøpe lokalt brennt kaffe og male dem selv hjemme.



NorwAI GPT Language Model Applications



[1] Xing, Yujie and Peng Liu. "Prompt and Instruction-Based Tuning for Response Generation in **Conversational Question Answering**", to appear in NLDB 2023.

https://github.com/yujie-xing/Multi-Turn_QA_Prompt

[2] Liu, Peng, Lemei Zhang, and Jon Atle Gulla. "Pre-train, prompt and recommendation: A comprehensive survey of language modelling paradigm adaptations in **recommender systems**." arXiv preprint arXiv:2302.03735 (2023).

<https://github.com/liupeng9966/LMRS>

[3] **Social Robot** in NorAI.



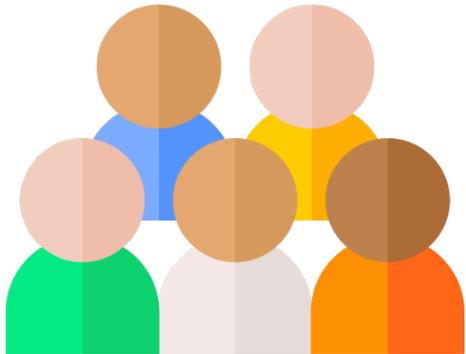
Something generated from NorGPT-3B...

Norwegian:

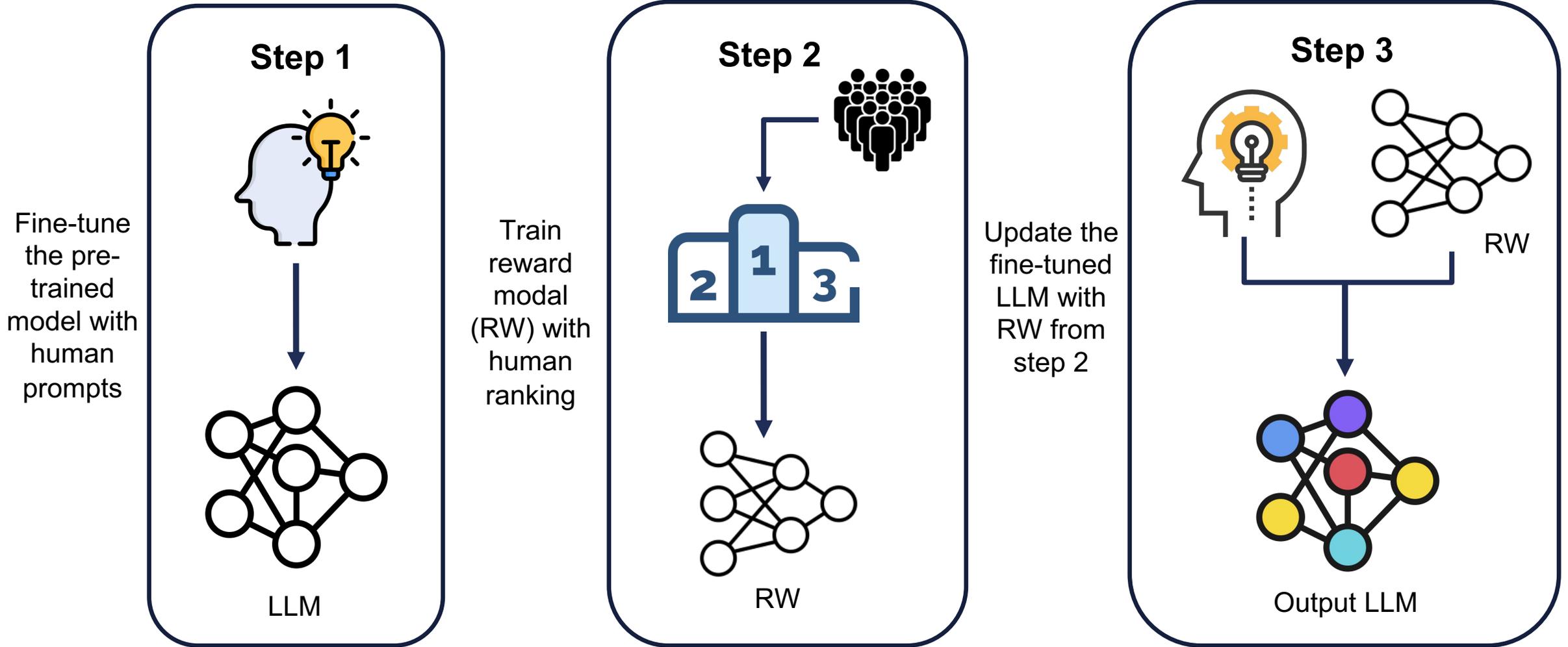
CNN.COM : CNN reporters and contributors to CNN in Hong Kong gather in front of the U.S. presidential election in South Korea in March, 2013.(Photo: KCNA/Jason K. Smith, AFP Photo)

English:

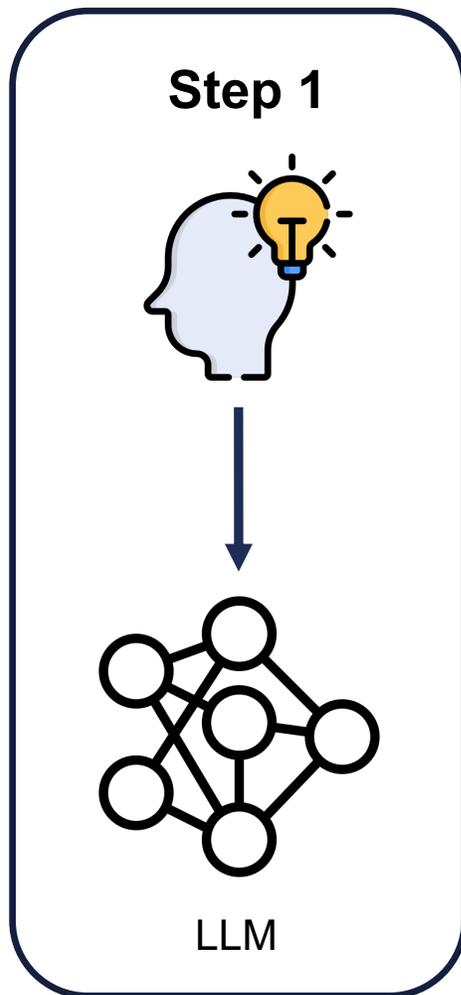
CNN.COM : CNN reporters and contributors to CNN in Hong Kong gather in front of the U.S. presidential election in South Korea in March, 2013.(Photo: KCNA/Jason K. Smith, AFP Photo)



NorGPT-3B Instruct Tuning

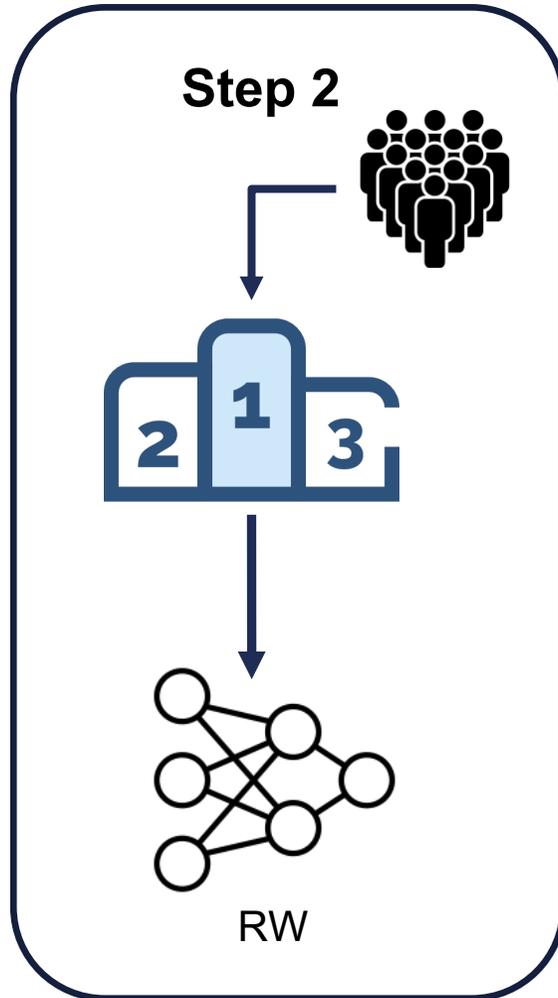


Step 1: Fine-tune with human feedback



- Norwegian supervised translation dataset:
 - News Summarization dataset from CNN/Dailymail
 - 1.18G, 128,354 prompts in total
 - 1 phase: 18,336 articles
 - 2 phase for ranking: 55,008 articles
 - 3 phase: 55,008 articles

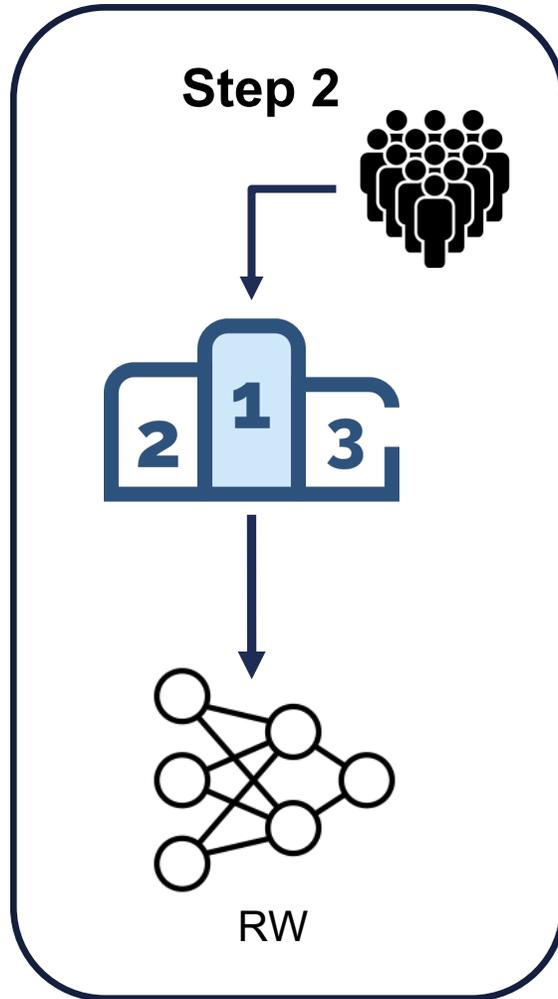
Human Feedback Collection



The image shows two overlapping website screenshots. The top one is for Prolific, with the text "Find your participants" and a link "Run a quick survey to recruit your sample." The bottom one is for Amazon Mechanical Turk, with the text "Amazon Mechanical Turk" and "Access a global, on-demand, 24x7 workforce".



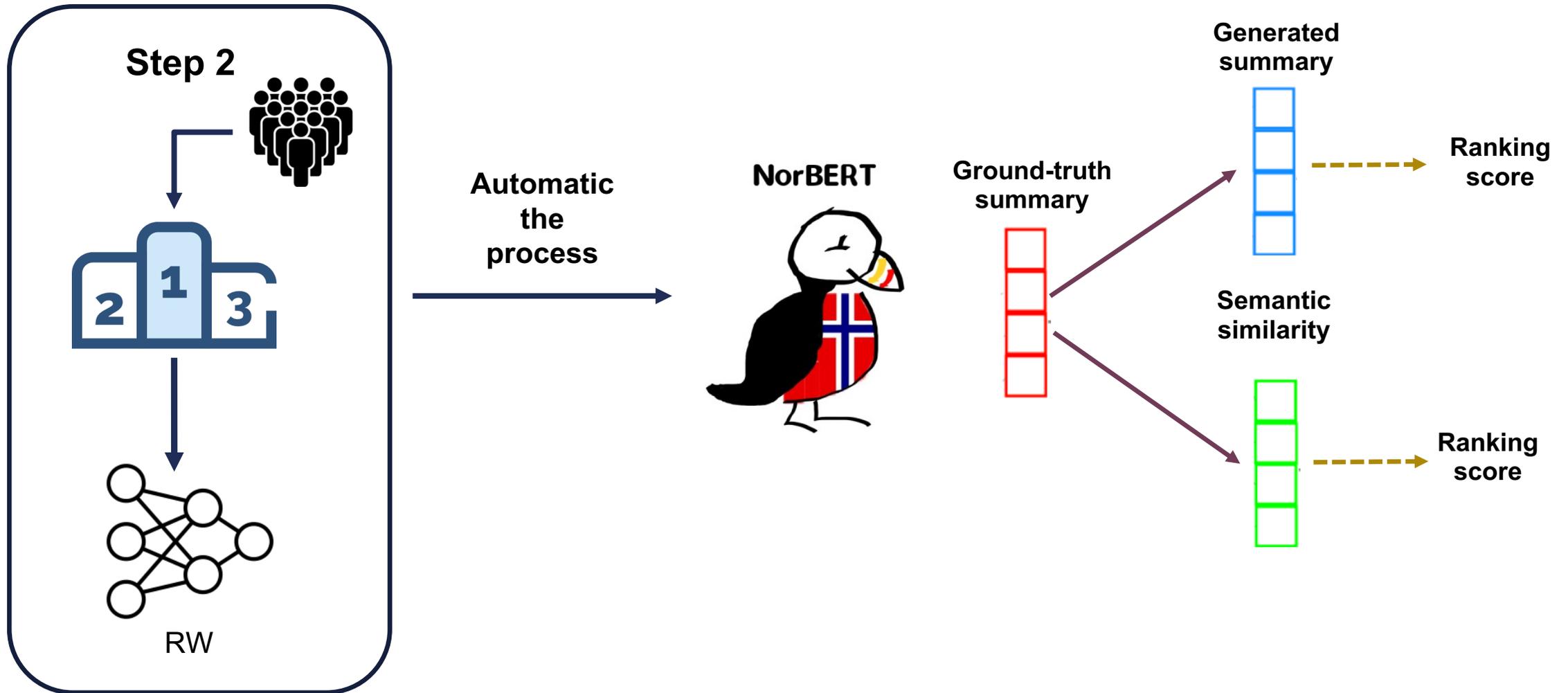
Human Feedback Collection



Recruit
participants
for small
samples to
get start



Human Feedback Collection



Case study

Norwegian

Baseball Hall of Famer Bob Feller døde onsdag kveld i Cleveland, Ohio, av akutt leukemi, ifølge en uttalelse fra Cleveland-indianerne. Han var 92. Feller ble diagnostisert med akutt myeloid leukemi i august og ble operert ved Cleveland Clinic for å implantere en pacemaker i oktober. I midten av november ble Feller innlagt på sykehus med lungebetennelse. Hovedligaen var et ikon for indianernes fans. "Vi i Cleveland har vært velsignet som har hatt ham med oss i disse mange årene," sa Indians eier Larry Dolan. "Å si at han vil bli savnet er en underdrivelse." Feller tilbrakte 18 år i en Cleveland Indians uniform. Han startet 484 kamper, og vant 266 av dem. Feller hadde en livstids ERA på 3,25, kastet tre no-hitters, inkludert den eneste Open Day no-hitteren i Major League Baseball-historien, den 16. april 1940. I 1946 spilte han 36 komplette kamper. "Siden 1936 har han vært med oss. I 75 år har han vært en medvirkende borger, en modell for alle idrettsutøvere og venn av tusenvis," sa Dolan. "Slik det skal være i større forstand, vil Bob alltid være med oss. Ikke på åpningsdagen, ikke på Fantasy Camp, ikke i presseboksen, men i våre hjerter." Indians-manager Manny Acta sa at Feller var en spesiell person. "Det har aldri vært en stor med en slik tilknytning til hans opprinnelige franchise," sa Acta. "Når du tenker på Cleveland-indianere, tenker du på Bob Feller og vice versa. Han var en ekte patriot og en stor Hall of Famer. Gutt, han elsket indianerne og vi elsket ham alle tilbake."

English

Baseball Hall of Famer Bob Feller died Wednesday night in Cleveland, Ohio, of acute leukemia, according to a statement from the Cleveland Indians. He was 92. Feller was diagnosed with acute myeloid leukemia in August and underwent surgery at the Cleveland Clinic to implant a pacemaker in October. In mid-November, Feller was admitted to hospital with pneumonia. The major leaguer was an icon for Indians fans. "We in Cleveland have been blessed to have him with us these many years." Indians owner Larry Dolan said. To say he will be missed is an understatement" Feller spent 18 years in a Cleveland Indians uniform. He started 484 games, winning 266 of them. Feller had a lifetime ERA of 3.25, threw three no-hitters, including the only Open Day no-hitter in Major League Baseball history, on April 16, 1940. In 1946, he pitched 36 complete games. "Since 1936 he has been with us. For 75 years he has been a contributing citizen, a model for all athletes and a friend to thousands," Dolan said. "As it should be in a larger sense, Bob will always be with us. Not on Opening Day, not at Fantasy Camp, not in the press box, but in our hearts." Indians manager Manny Acta said Feller was a special person. "There has never been a great with such a connection to his original franchise," Acta said. "When you think of the Cleveland Indians, you think of Bob Feller and vice versa. He was a true patriot and a great Hall of Famer. Boy, he loved the Indians and we all loved him back."



Case study

NorGPT generated summary (before fine-tune)

Norwegian

“Vi vil alltid savne Bob Feller” Faller døde torsdag kveld i Cleveland, Ohio, av akutt leukemi. Han var 92 år gammel. I sin biografi, " The Baseball Hall of Fame," skrev han at "han var den beste, den mest lojale, den beste fansen og en meget god leder".

English

“We will always miss Bob Feller.” Faller died Thursday night in Cleveland, Ohio, of acute leukemia. He was 92 years old. In his biography, "The Baseball Hall of Fame," he wrote that "he was the best, the most loyal, the best fan and a very good manager".

-0.18

Reward score

NorGPT generated summary (after fine-tune)

Norwegian

Baseball Hall of Famer Bob Feller dør på akutt leukemi i en alder av 92. Feller ble operert ved Cleveland Clinic for å implantere en pacemaker i oktober. Han ble innlagt på sykehus etter å ha utviklet lungebetennelse. Feller tilbrakte 18 år på uniform.

English

Baseball Hall of Famer Bob Feller dies of acute leukemia at age 92. Feller underwent surgery at the Cleveland Clinic to implant a pacemaker in October. He was hospitalized after developing pneumonia. Feller spent 18 years in uniform.

-0.15



What's next?

- Generative models such as LLaMA on the LUMI supercomputer
- Separate Bokmål and Nynorsk models
- Streamline the benchmarking workflow
- Evaluation metrics for news summarization



Project Team



Benjamin Kille



Peng Liu



Lemei Zhang

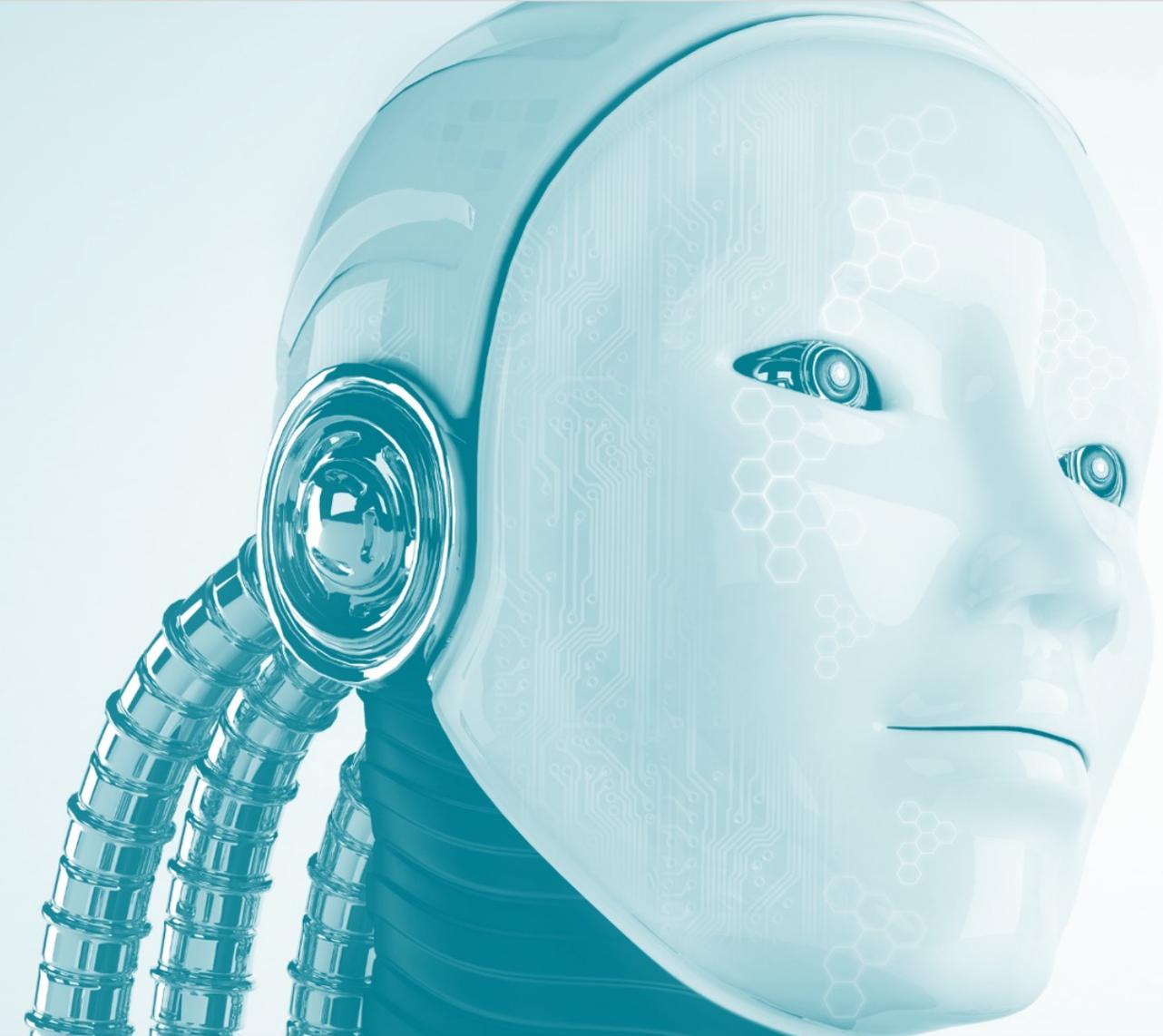


Jon Atle Gulla



Jon Espen Ingvaldsen





Thank you!

Q&A

Norw**AI**

Norwegian Research Center
for AI Innovation

