Nye norske språkmodellar 6 månader seinare: Evaluering og erfaringar

Jon Atle Gulla
Norsk forskingssenter for AI-innovasjon (NorwAI)
NTNU

NorwAI Norwegian Research Center for AI Innovation

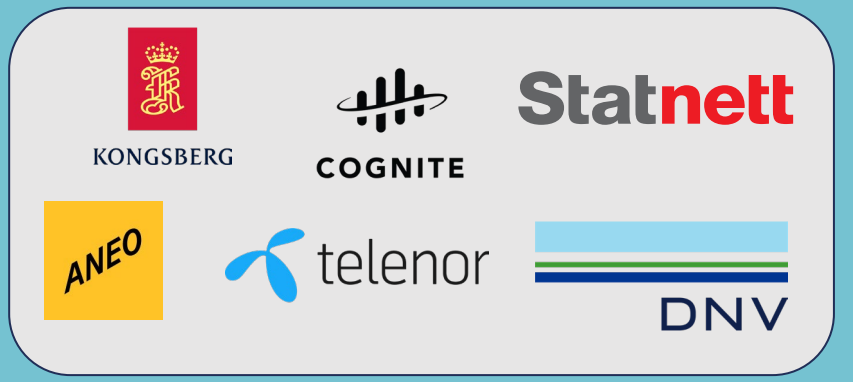
NTNU

sfi Centre for Research-based Innovation
The Research Council of Norway

Norsk forskingssenter for AI-innovasjon

Senter for forskningsdriven innovasjon (SFI):
Styrke teknologioverføring, internasjonalisering og forskartrening gjennom langsiktig forskingssamarbeid mellom forskingstunge selskap og prominente forskingsgrupper

- 286 MNOK over 8 år (2020 – 2028)
- Finans, media and Industry 4.0
- Forsking på anvend KI:
  - Nye produkt og tenester
  - Startups
  - 500+ masterkandidatar og 25+ PhD-kandidatar
  - Kurs, seminar, konferansar, workshops, etc.
  - Kompetansenettverk



NorwAIs språkmodellar

Hugging Face

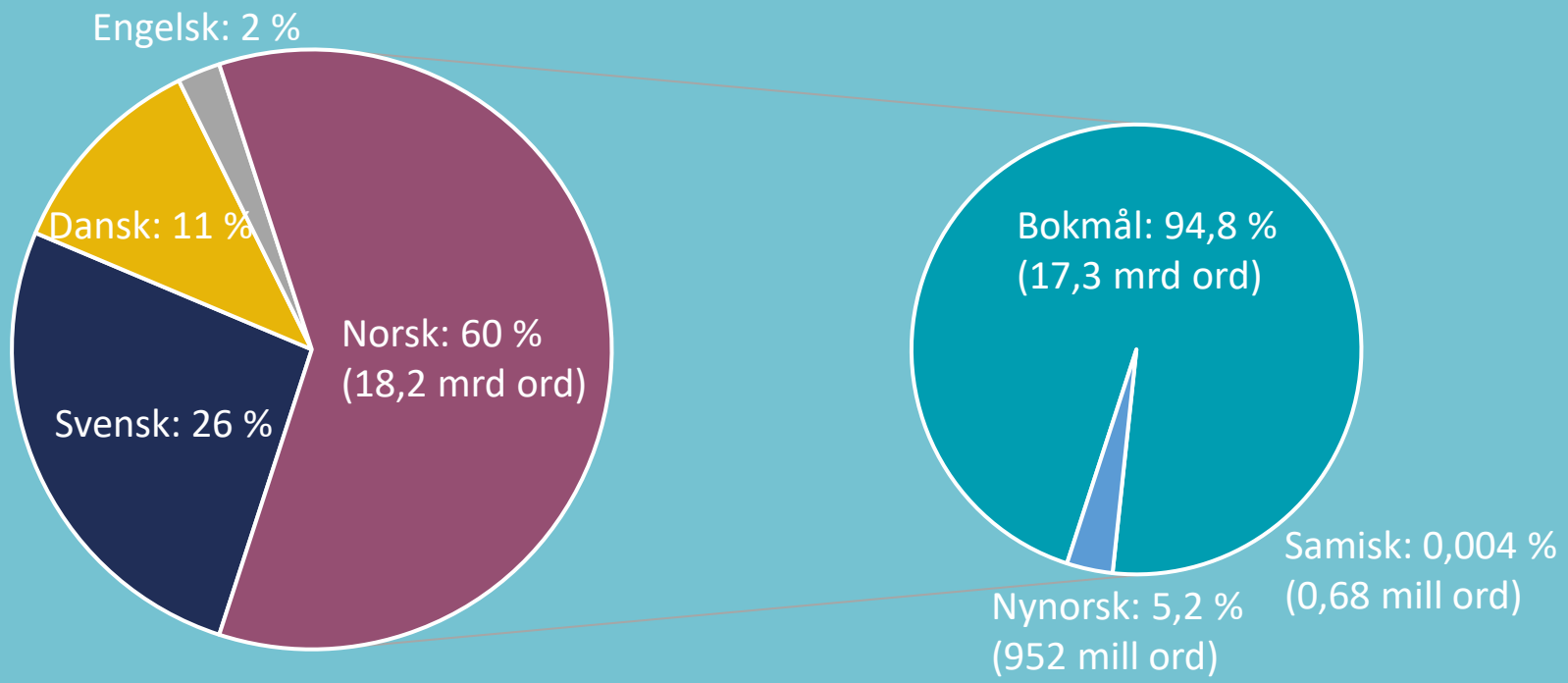
NorwAI-Mistral-7B-instruct:	578 (72)
NorwAI-Llama2-7B:	957 (134)
NorwAI-Mistral-7B:	2783 (136)
NorwAI-Mixtral-8x7B:	47 (5)
NorwAI-Mistral-7B-pretrain:	232 (50)
NorwAI-Mixtral-8x7B-instruct:	224 (39)

(Nedlastingar 1. juni – 13. september 2024)

Dagbladet,
Arkitektum AS,
Appfabrikken AS,
Horten kommune,
Telemark toppidrett ungdomsskole,
Telemark Fylkeskommune,
Telemarksforsking,
Even Vinge,
Lierne Kommune Oppvekst- Og Kultur,
Statped,
Bekk Consulting AS,
Kanoa AS,
Advania Norge AS,
Landbruksdirektoratet,
NRK,
Trondheim kommune,
PwC Norge,
DNB,
SINTEF Community,
Helse-Vest IKT,
Helse Midt-Norge IT,
BI,
Egde consulting,
Duplo Media AS,
Posten Bring AS,
Norconsult Digital AS,
Avarn Security AS,
Machina AS,
Å Energi,
Ruter,
HomeKey,
Bineric,
Lingit AS,
ByGuru AS,
Forsvarsdepartementet,
Norsk Helsenett,
arbeidstilsynet,
Nettskulen Vestland,
BoostSecurity.io,
Forsvarsmateriell,
Sunnaas HF,
Frontkom AS,
ideas2evidence,
TUM Media,
Trivselslaben,

Språkmodellar for norsk språk

30,3 milliardar ord / 51,2 milliardar tokens / 348,1 GB



Kategoriar:

Web: 35,5%
Off. dokument: 26,6%
Nyheiter: 6,3%
VG debatt: 1,9%
Bøker: 1,3%
Andre: 28,3%

Nasjonalt initiativ: MIMIR

Offentleg

Nasjonalbiblioteket

Akademia

Universitetet i Oslo

NTNU

Kulturdepartementet

- Januar – juli 2024
- Vurdere verdien av opphavsbeskytta innhald i norske språkmodellar

- Pretrening og evaluering av ei rekkje språkmodellar
  Treningsdata med og utan opphavsbeskytta innhald (NB)
  Trening av 7B-modellar
  Evaluering langs fleire dimensjonar

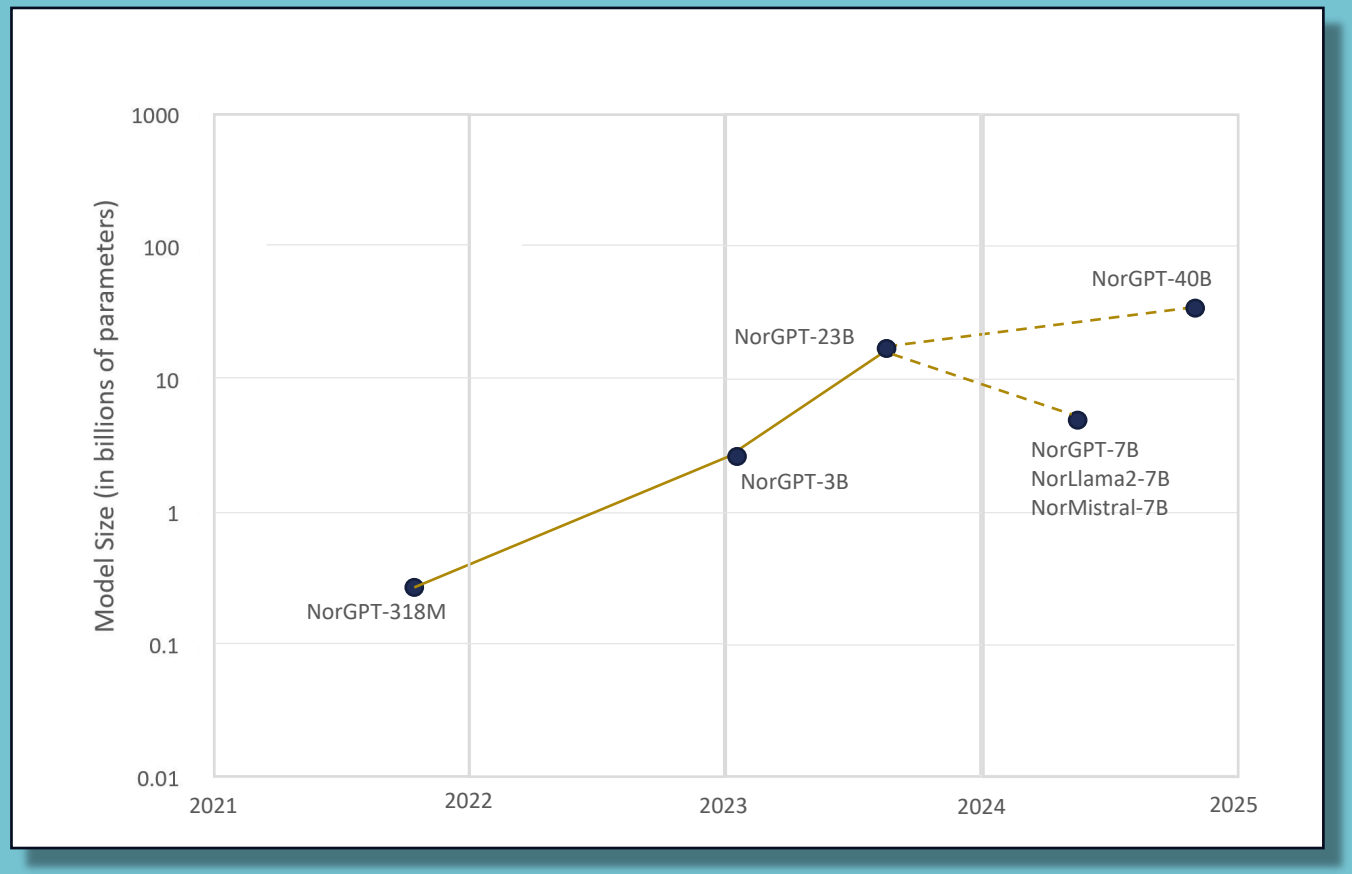
NorwAIs satsing over tre år

- Nasjonale generative språkmodellar for bokmål/nynorsk
- Grunnmodell på 40 milliardar parameter

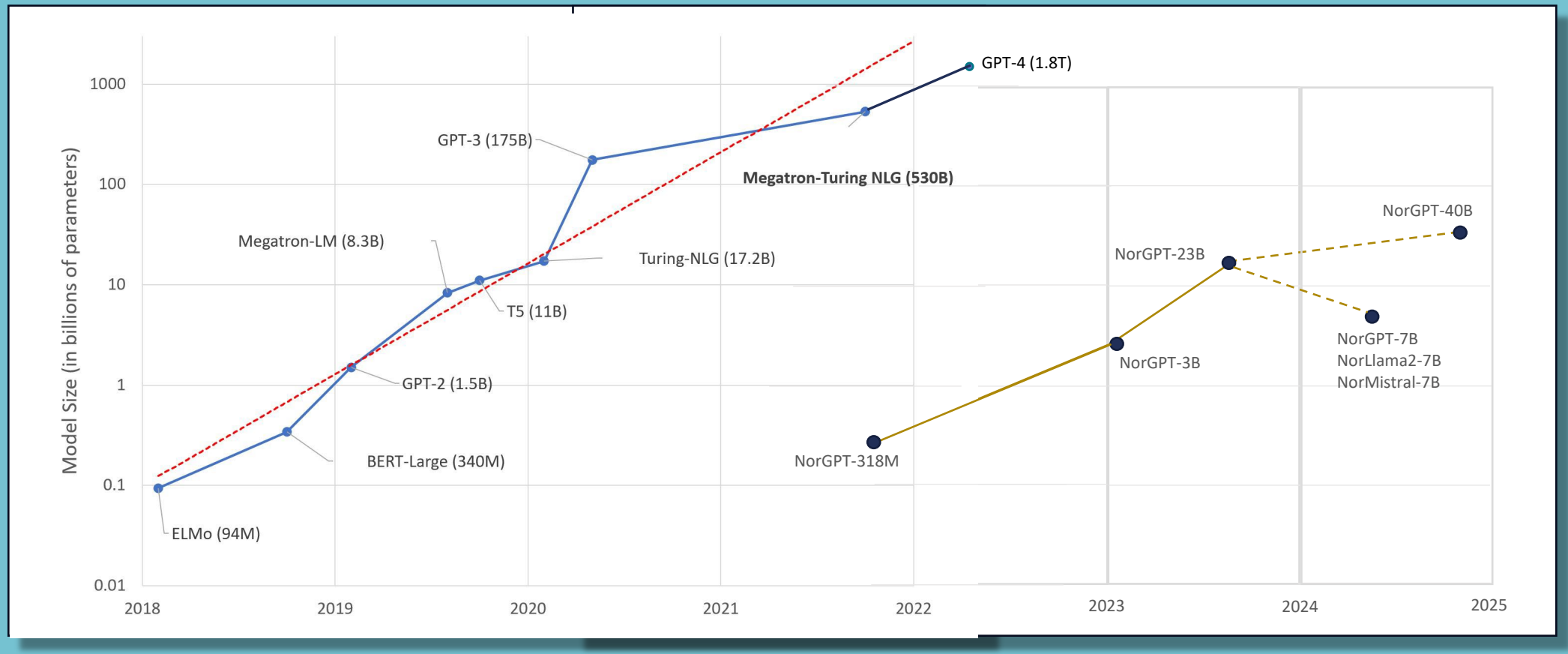
  Sentral drifting og finjustering
  Opphavsbeskytta innhald

- Modellar på 7 milliardar parameter

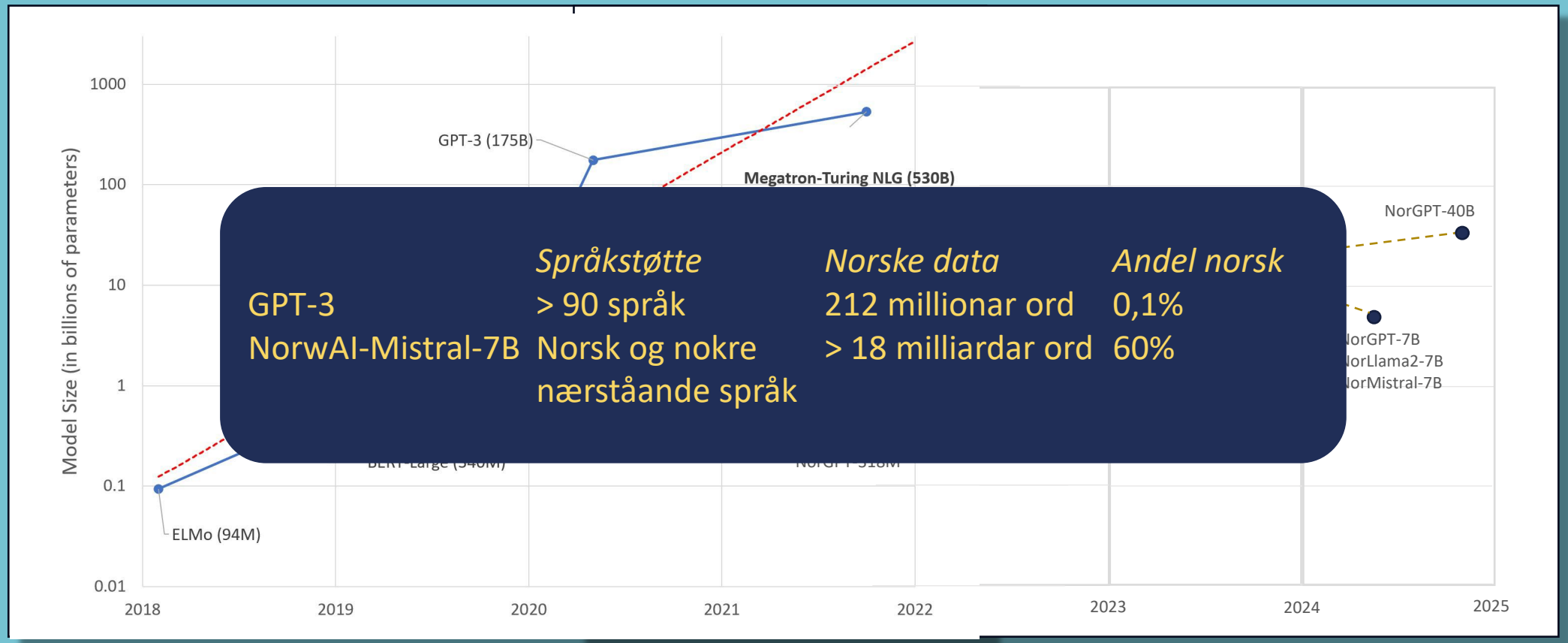
  Lokal drifting
  Frie treningsdata
  Lokal finjustering



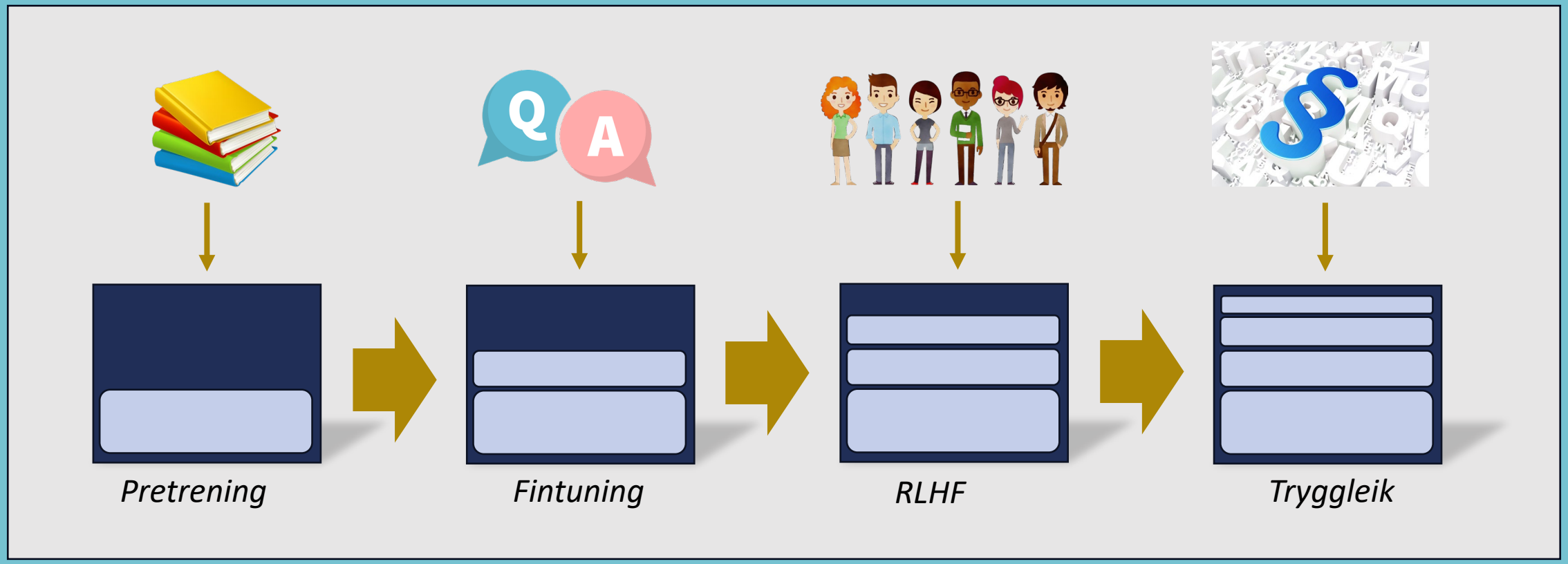
Internasjonale modellar



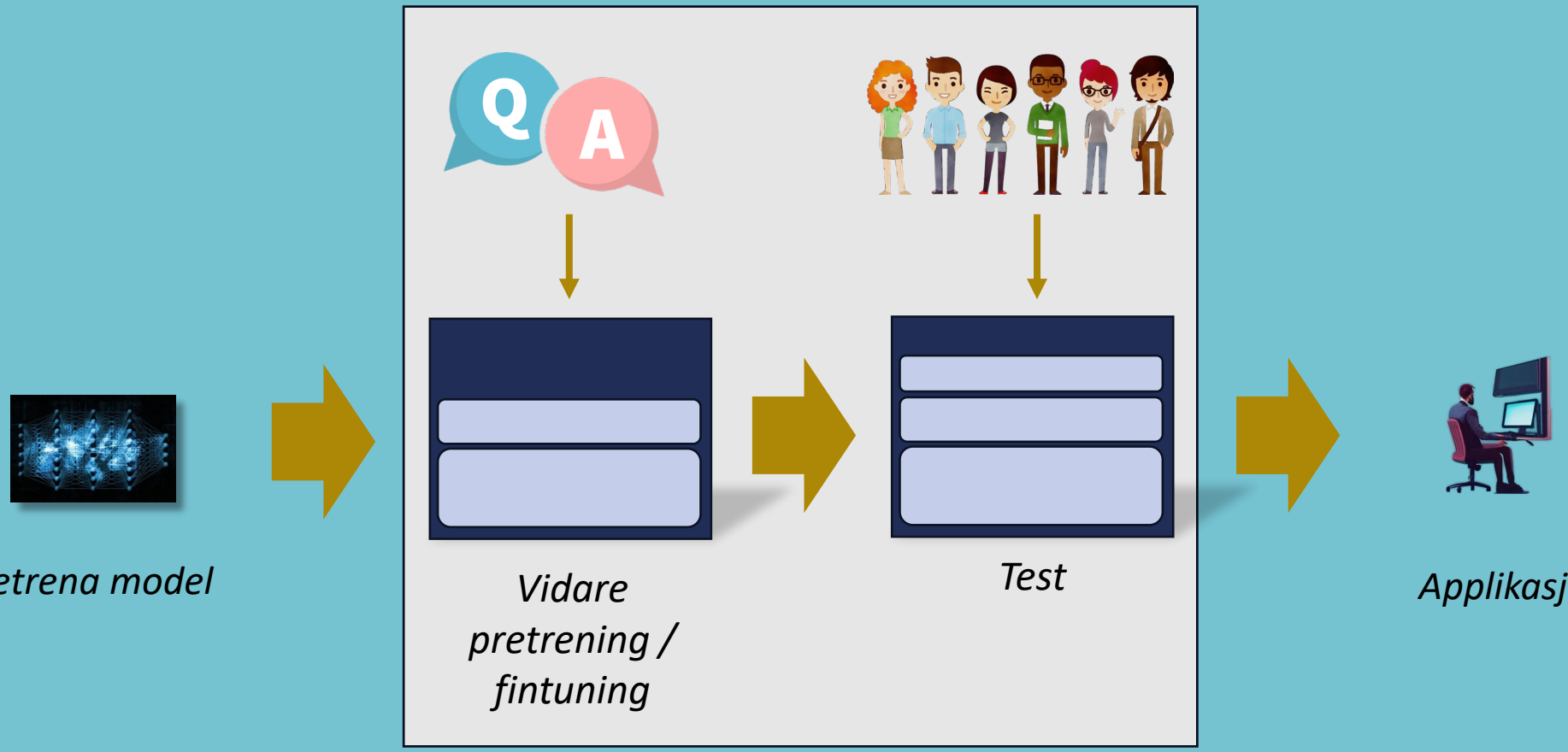
Internasjonale modeller



Trening av språkmodellar



Tilpass til eigne data - test på eigne oppgåver



Erfaringar så langt

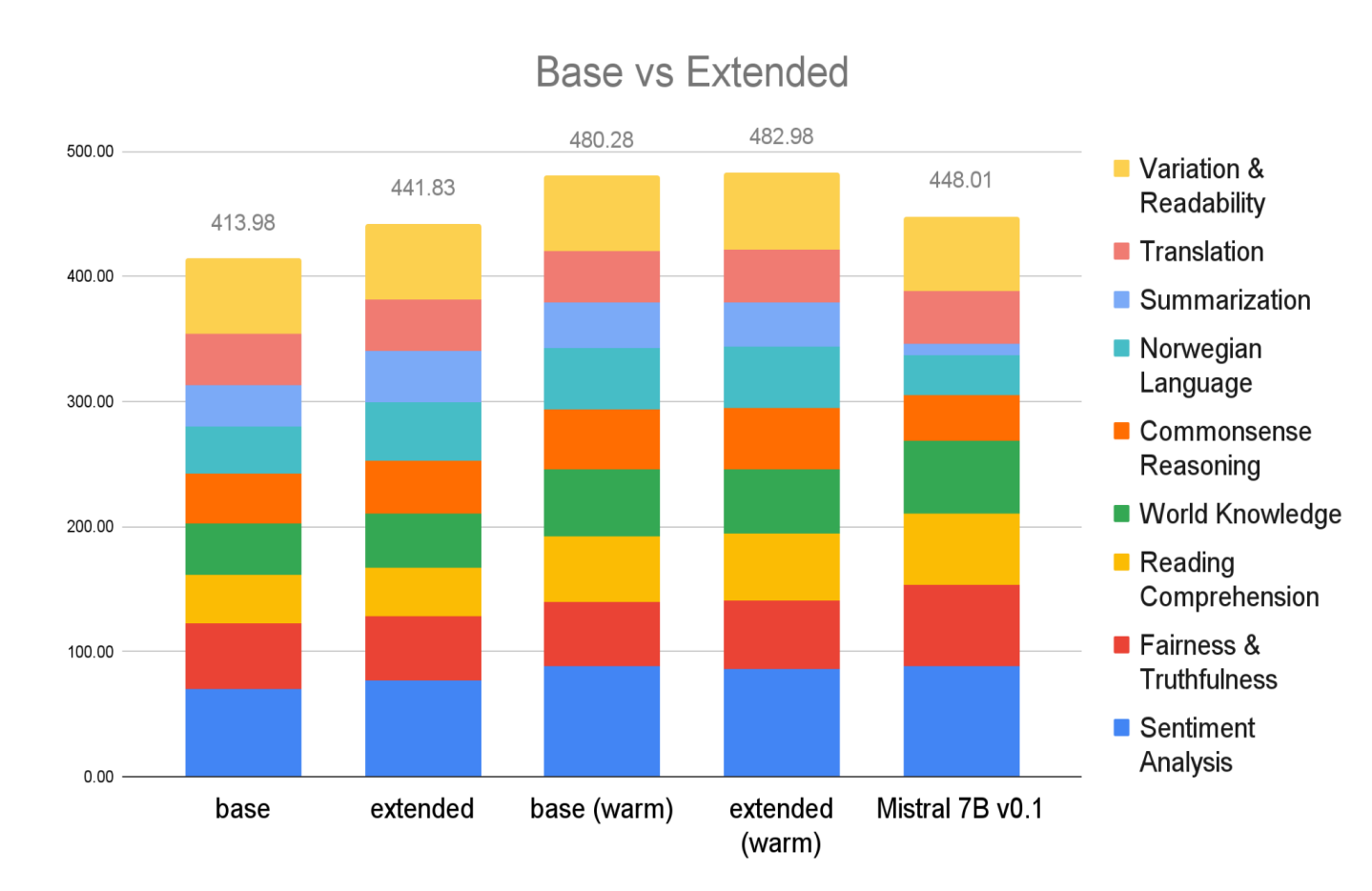
Evalueringar og bruk av
norske språkmodellar

NorwAI

Norwegian Research Center
for AI Innovation

NTNU

Evalueringen i MIMIR



Base: *treningsdata utan opphavsrettar frå NB (40 milliardar ord)*
Extended: *Treningsdata med alt publisert innhald frå NB (82 milliardar ord)*

Mistral 7B v0.1:
Open internasjonal modell
7B parameter
Brukt som basis for MIMIR

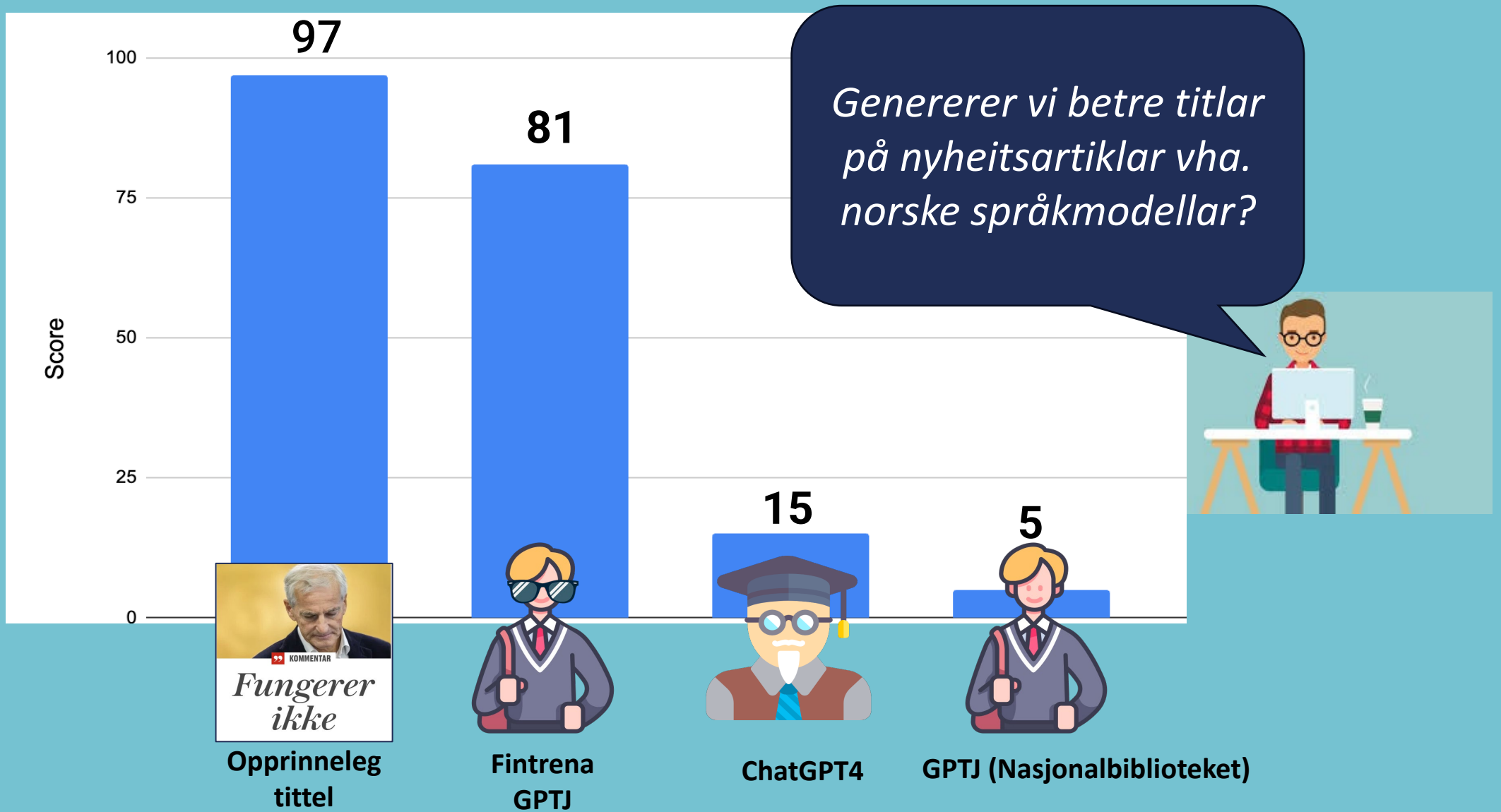
Warm: *pretrening på toppen av pretrena Mistral 7B v0.1*
Kald: *Pretrening frå scratch*

Norske modellar betre med norske data

- Vidare pretrening med norske data betrar internasjonale modellar på norsk
- Høgkvalitetsinnhald og/eller meir innhald gir betre modellar
- Datakvalitet viktig for modellkvalitet
  - Skjønnlitteratur -> betre lingvistisk kvalitet på modellar
  - Aviser -> betre for kunnskapsuthenting
- Instruksjonstuning betrar modellane betydeleg

- NB! Vidare pretrening gir betre modellar enn pretrening frå scratch (men er vi sikre på innhaldet i basemodellen?)

Tekstgenerering i VG (Schibsted Media)



Norske språkmodellar testa i VG

	NB-GPTJ 6B	NorwAI Mistral 7B	NorwAI Mixtral 7B
Tittel	☑		
Ingress	✗		
Skrivehjelp	✗		
Samandrag	✗		

NorwAI Norwegian Research Center for AI Innovation

NTNU

Norske språkmodellar testa i VG

	NB-GPTJ 6B	NorwAI Mistral 7B	NorwAI Mixtral 7B
Tittel	☑	☑	
Ingress	✗	☑	
Skrivehjelp	✗	⧗	
Samandrag	✗	✗	

Store helseeffekter: Slik trener du best med smartklokke

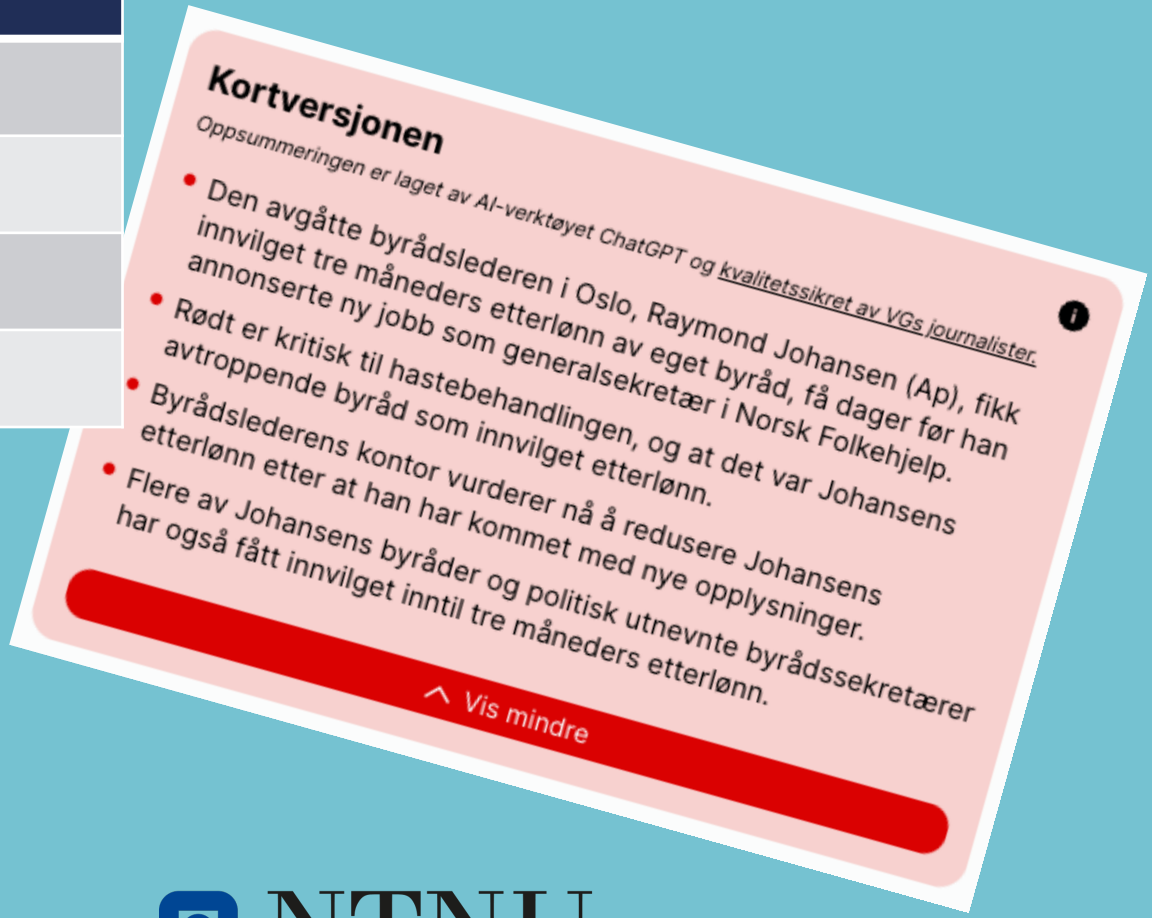
Hvordan bruke smartklokken smart, og få bedre pulsvariasjon og «body battery»? Treningsekspert Halvor Lauvstad (51) har klare råd, og trekker frem én treningsform som den mest effektive.

NorwAI Norwegian Research Center for AI Innovation

NTNU

Norske språkmodellar testa i VG

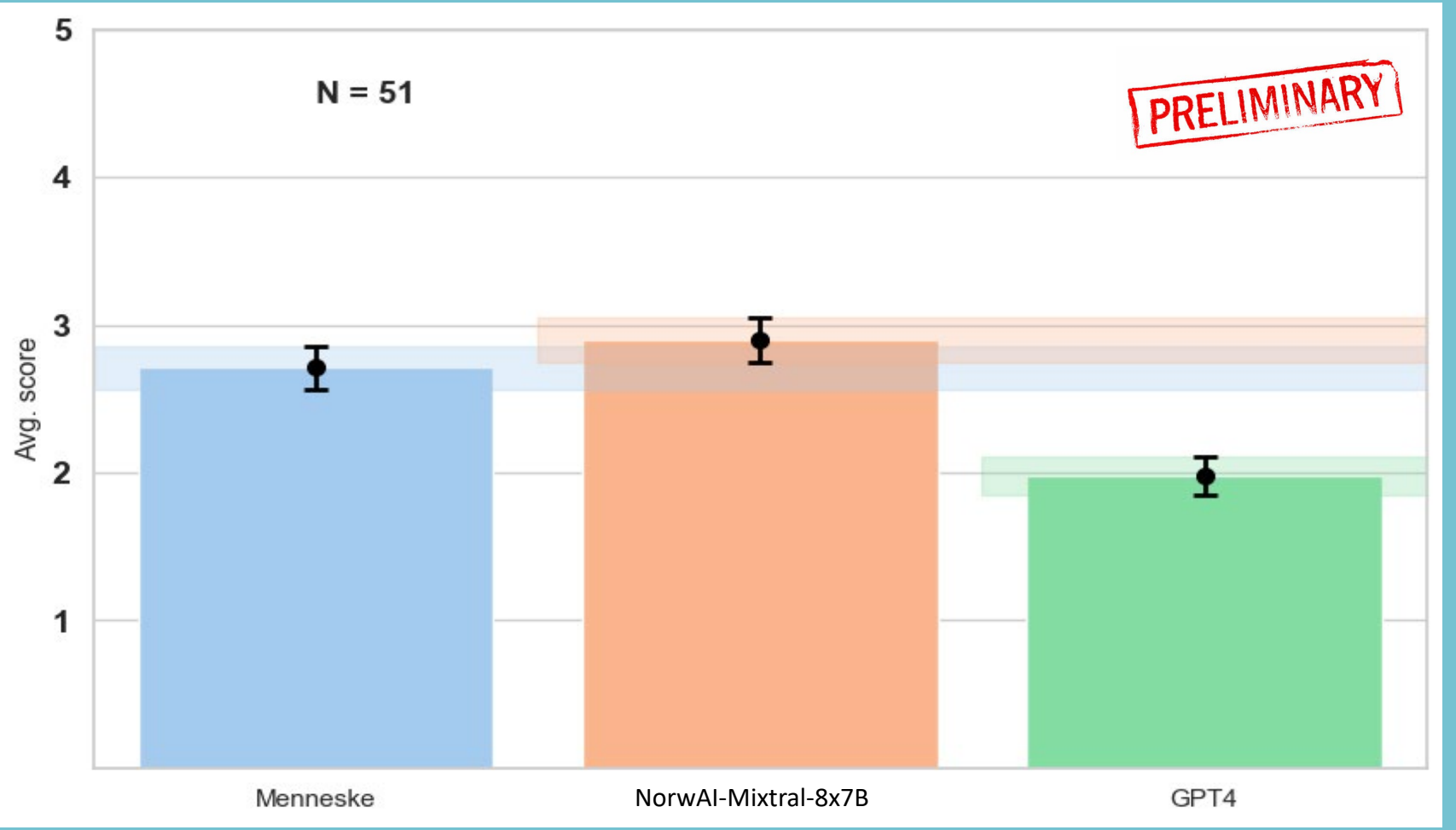
	NB-GPTJ 6B	NorwAI Mistral 7B	NorwAI Mixtral 7B
Tittel	☑	☑	
Ingress	✗	☑	
Skrivehjelp	✗	⌛	
Samandrag	✗	✗	☑



NorwAI Norwegian Research Center for AI Innovation

NTNU

Kortversjonen (samandrag)



NorwAI Norwegian Research Center for AI Innovation

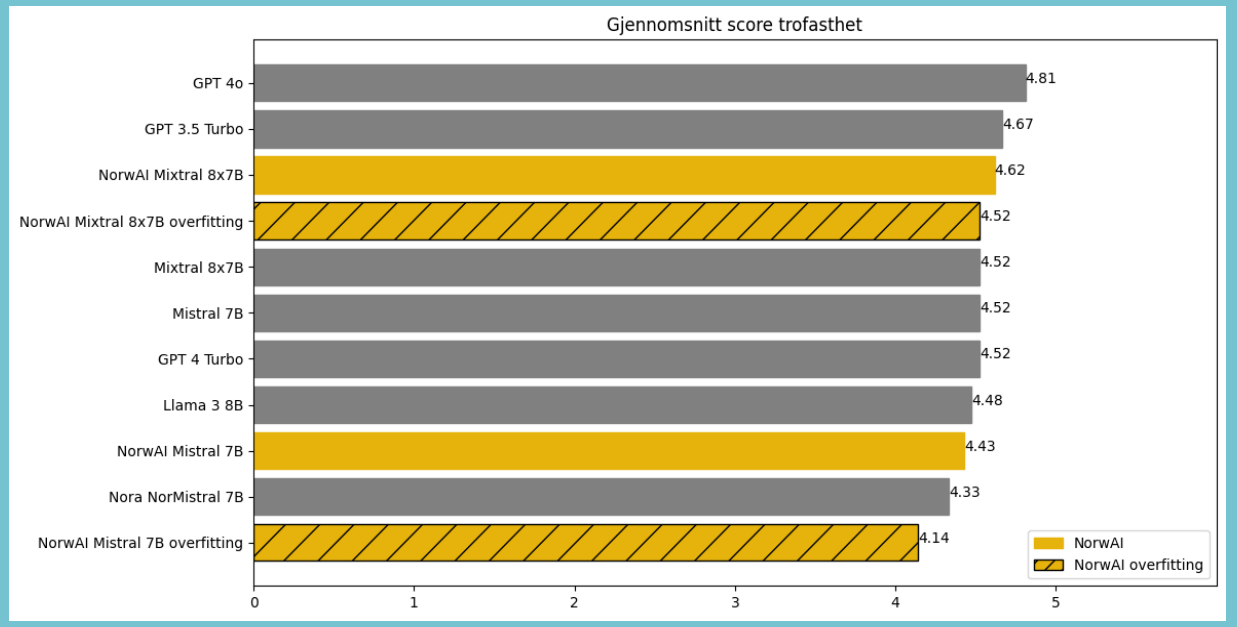
NTNU

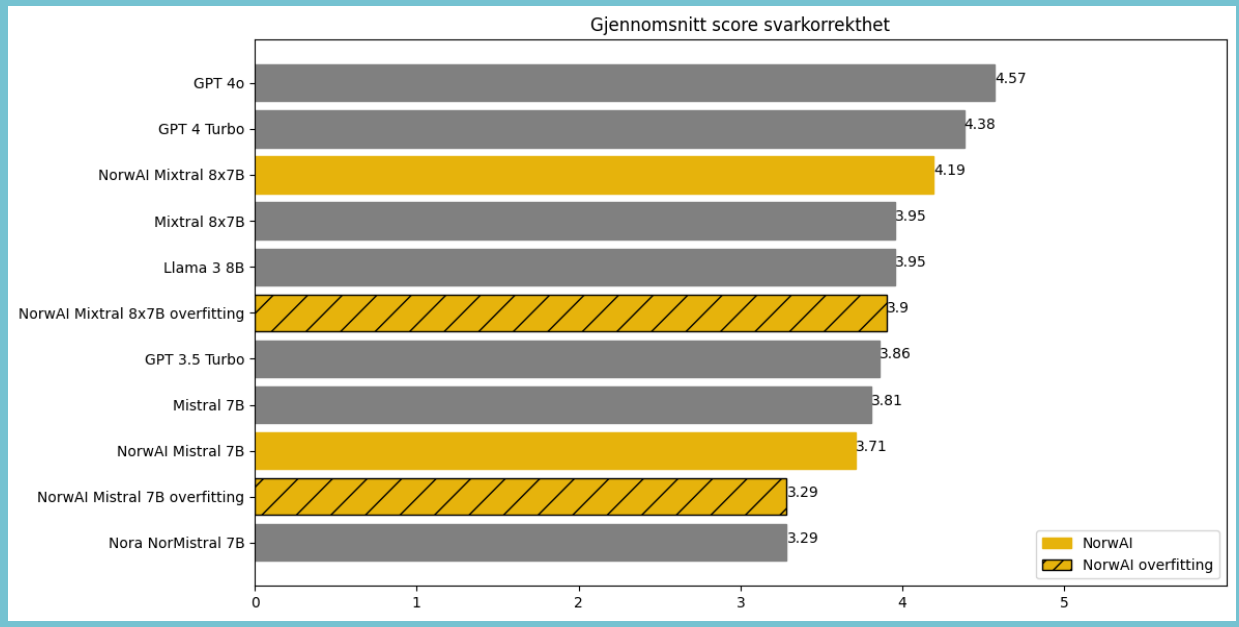
Overraskande klare resultat

- Vidare pretrening med norske data viktig
- Vidare norsk pretrening med fintuning betre enn største internasjonale modellane på nyheitsdomenet
- Store modellar betre enn små

- *Fintuning verkar alt på 1000-2000 eksempel. Titlar i VG fintuna med 10.000 eksempeltitlar.*

Retrieval-Augmented Generation i Sparebank1 SMN





Truskap: Kor mykje av det genererte svaret er basert på interne data? Mål på hallusinasjonar.

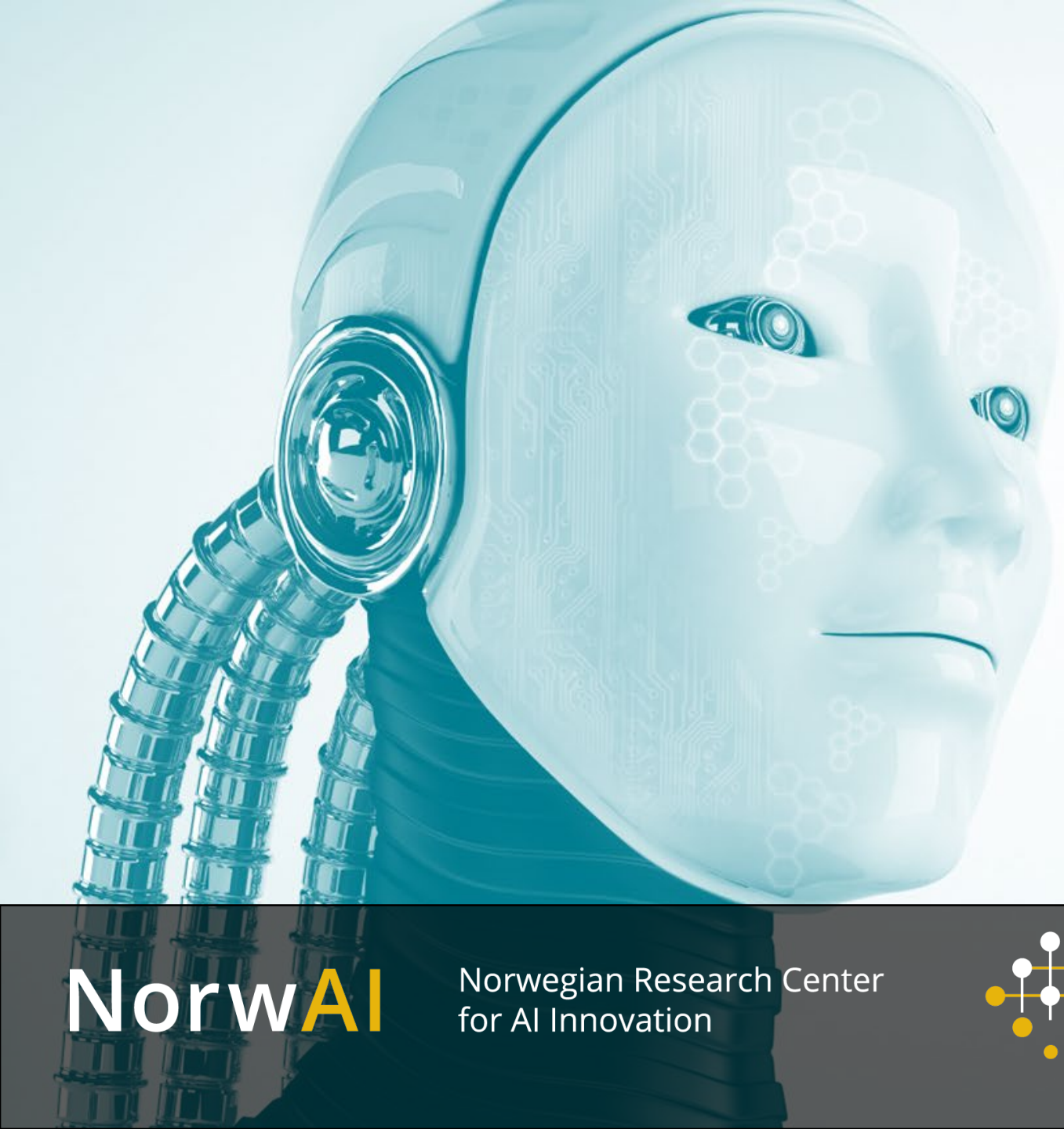
Nøyaktigheit: Kor korrekt er svaret? Mål på faktuell likskap mellom generert respons og korrekt svar.

NorwAI

Fintuning er viktig

- Ingen fintuning i eksperimenta til SMN
- Utan fintuning er internasjonale store modellar ofte litt betre
- Relativt små norske modellar er overraskande konkurransedyktige når trena på gode norske data

- Vidare pretrening på norsk + fintuning = SANT



Takk

Jon Atle Gulla
jag@ntnu.no

NorwAI Norwegian Research Center for AI Innovation

NTNU

sfi Centre for Research-based Innovation

The Research Council of Norway