Mímir-prosjektet

Evaluering av virkningen av opphavsrettsbeskyttet materiale på generative store språkmodeller for norske språk

Versjon: 1
Dato: 5. juli 2024
Norsk versjon: 8. august 2024

Introduksjon

Mímir-prosjektet er et initiativ fra den norske regjeringen som har som mål å vurdere betydningen og innflytelsen av opphavsrettsbeskyttet materiale for utviklingen og ytelsen av generative store språkmodeller for norske språk og norske forhold. Dette arbeidet involverte tre ledende institusjoner: Nasjonalbiblioteket (NB), Universitetet i Oslo (UiO), og Norges teknisk-naturvitenskapelige universitet (NTNU/NorwAI). Hvert miljø bidro med unik ekspertise innen språkteknologi, som tilrettelegging av data (korpus), trening av modeller, opphavsrett og datalingvistikk. I tillegg har Sigma2 levert regnekraft til arbeidet. Det endelige målet med prosjektet var å etablere kunnskap som kan danne grunnlag for utformingen av en eventuell kompensasjonsordning for rettighetshavere der innholdet fra verk under opphavsrett brukes for å trene kunstig intelligens (KI).

Bakgrunn

Fremveksten av store språkmodeller har revolusjonert språkteknologi (natural language processing, NLP), noe som gjør det mulig for maskiner å generere, forstå og samhandle ved å bruke menneskelig språk med stor nøyaktighet. Imidlertid krever disse modellene enorme mengder tekstdata for å trenes, ofte hentet fra et bredt utvalg av digitale samlinger som også kan inkludere opphavsrettsbeskyttet materiale. Dette reiser kritiske juridiske og etiske spørsmål angående bruk av slikt innhold uten eksplisitt tillatelse fra rettighetshaverne, samt mulige økonomiske implikasjoner for kreativ aktivitet.

I november 2023 kontaktet Norges rettighetshaverorganisasjon regjeringen med krav om kompensasjon for bruk av deres materiale i trening av generativ AI. For å støtte eventuelle

forhandlinger med rettighetshavere med kunnskap ga regjeringen Nasjonalbiblioteket i oppdrag å gjennomføre en forskningsbasert aktivitet for å ta fram kunnskap på feltet.

Ledet av Nasjonalbiblioteket ble det dannet et prosjekt sammen med Language Technology Group(LTG)/Universitetet i Oslo og NorwAI/NTNU i tillegg til Sigma2 som leverandør av regnekraft. Mímir-prosjektet svarer ut spørsmål på feltet ved systematisk evaluering av hvilken innvirkning ulike typer tekst under opphavsrett kan ha for ytelsen for store språkmodeller for norsk og norske forhold. Prosjektet ble etablert i januar 2024 med ambisjon om å ha resultatene klare i løpet av 6 måneder. Dette krevde organisering og kuratering av data, bygging av datasett, design av eksperimenter, trening av modeller, lage evalueringsreferanser, evaluering av modeller og behandling og tolking av resultatene.

Metodikk

Prosjektets metodikk innebar en omfattende analyse over flere stadier. Til å begynne med samles et mangfoldig korpus av norskspråklige data, som inkluderer både opphavsrettsbeskyttet og ikke-opphavsrettsbeskyttet innhold, pluss materiale som vanligvis høstes fra internett. Dette korpuset fungerer som grunnlaget for opplæring av ulike store språkmodeller, hver med forskjellige konfigurasjoner og tilgangsnivåer til opphavsrettsbeskyttet innhold. Ved å sammenligne ytelsen til disse modellene på tvers av en rekke språklige og naturlige oppgaver for språk, som tekstgenerering, oversettelse, spørsmål/svar og sentimentanalyse, søker prosjektet å kvantifisere de spesifikke bidragene til opphavsrettsbeskyttet materiale til den generelle ytelsen fra modellene.

For å sikre robusthet og pålitelighet, fokuserer rammeverket for evaluering på både grunnleggende ytelse fra modellene og språklig inspirerte oppgaver. Kvantitative mål inkluderer tradisjonelle NLP-målinger som nøyaktighet, F1, BLEU-score og ROUGE-score, som gir objektive vurderinger av modellenes kvalitet i ytelse. Språklig analyse involverer derimot å vurdere sammenhengen, variasjon i språk, og kontekstuell relevans av de genererte resultatene. Denne doble tilnærmingen gir mulighet for en nyansert forståelse av hvordan opphavsrettsbeskyttet materiale bidrar til ytelsen og nytten i store språkmodeller for norsk og norske forhold.

Samlinger og datasett

Mimir tilpasset metoder fra Norwegian Colossal Corpus (NCC)[1] for bygging av datasett, og sparte tid til tross for enkelte begrensninger. Våre datakilder inkluderer det åpne internett (f.eks. Wikipedia, resultater fra High Performance Language Technology-prosjektet[2]), innhold fra mediehus (f eks NRK, Amedia, Schibsted, TV2), og NBs digitale samling (f.eks. opphavsrettsbeskyttede aviser og bøker). Data behandles gjennom en skreddersydd prosess til et standardisert format, som sikrer enhetlig funksjonalitet og kvalitet, hvoretter deduplisering følger for å sikre unike enheter. Hver datapost har nok metadata til å bistå treninga av modellene, og til å balansere norsk tekst for å forhindre at det blir dominert av andre språk.

[1] The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models (Kummervold et al., LREC 2022)
[2] https://hplt-project.org/

Mimir etablerte to typer korpus: komplette korpus for pretrening (se tabell 1) og forskjellige undergrupper og konfigurasjoner (delkorpus) for finjustering på begrenset materiale (se tabell 2). De komplette korpusene kommer i variantene **base** og **extended**, der *base* i hovedsak ikke har innhold under opphavsrett, mens *extended* inneholder alt tilgjengelig innhold i NBs digitale samling av bøker og aviser, og er dermed nesten dobbelt så stor. Alle korpus er renset, deduplisert og språkbalansert. Det ble laget et ekstra datasett for såkalt instruksjonsfinjustering som inneholder 5000 instruksjonspar.

Komplette korpus	**Dokumenter**	**Ord**
base	60,182,586	40,122,626,817
extended	125,285,547	82,149,281,266

Tabell 1. Antall dokumenter og ord per komplett korpus for pretrening

Delkorpus	**Dokumenter**	**Ord**
bøker	492,281	18,122,699,498
aviser	46,764,024	9,001,803,515
bøker + aviser	47,256,305	26,078,915,554
skjønnlitteratur	117,319	5,287,109,366
faglitteratur	359,979	12,384,323,012
faglitteratur + aviser	42,083,532	20,340,539,068
norsk litteratur	392,887	13,352,261,605
norsk litteratur + aviser	47,156,911	22,354,065,120
oversatt litteratur	96,258	4,695,814,506

Tabell 2. Antall dokumenter og ord per delkorpus for finjustering, som kun inneholder opphavsrettslig beskyttet materiale.

Trening

Mimir trente 17 modeller med 7 milliarder parametere[3] basert på Mistral-arkitekturen, noe som forbrukte totalt ca 270 000 GPU-timer. Infrastrukturen inkluderte LUMI (EuroHPC), Idunn-klyngen (NTNU) og Google TPU-er gjennom Google TPU Research Cloud-program.

[3] En modell med 7 milliarder parametere er fortsatt 25 ganger mindre enn OpenAIs GPT-3, forgjengeren til ChatGPT.

Treningen ble gjennomført i to faser. For det første, for å måle virkningen av opphavsrettsbeskyttet materiale som helhet og virkningen som opphavsrettsbeskyttet materiale kan ha i et realistisk scenario etter trening, gjennomførte vi forhåndstrening på base og extended både fra bunnen av og fra de eksisterende vektene (varm) av Mistral 7B v0.1. Disse 4 kjernemodellene ble trent på samme mengde totale ord (64 000 trinn på ~2,5 millioner ord) ved bruk av identiske oppsett. For det andre, for ytterligere å isolere effekten av forskjellige deler av det opphavsrettsbeskyttede materialet, trente vi basemodellen (som var trent fra grunnen av) videre i ytterligere 10 000 trinn på hver av de 9 delkorpusene.

Kjernemodellene ble også videretrent på instruksjonskorpuset i 4 iterasjoner for å evaluere ytelsen på nedstrømsoppgaver etter finjustering.

Evaluering

Evaluering av generative språkmodeller er langt fra et løst problem, spesielt for norsk, hvor det var få eksisterende ressurser i starten av prosjektet. Gjennom en dedikert og intens innsats i forbindelse med prosjektet, etablerte vi et sett med 28 av de vanligste oppgavene i NLP, som omfatter en rekke forskjellige beregninger for å vurdere ytelsen til hver av modellene. Disse oppgavene kan grupperes i 9 ferdighetsområder:

- **Sentiment Analysis (Sentimentanalyse)**, som innebærer å bestemme den emosjonelle tonen bak en ordrekke. Det brukes til å identifisere følelsen uttrykt i et tekststykke, som kan være positivt, negativt eller nøytralt. For eksempel, i kundeanmeldelser eller innlegg på sosiale medier, hjelper sentimentanalyse å måle opinion eller tilfredshet.
- **Fairness & Truthfulness (Rettferdighet og sannferdighet)**. Rettferdighet i språkmodeller refererer til mangel på skjevhet (bias) i modellens genererte tekster. Evaluering av rettferdighet sikrer at modellen ikke favoriserer eller diskriminerer bestemte grupper basert på egenskaper som rase, kjønn eller etnisitet. Sannferdighet innebærer nøyaktigheten og pålideligheten til informasjonen som produseres av modellen, og sikrer at den genererer fakta og etterprøvbart innhold.
- **Reading Comprehension (Leseforståelse)**, som måler en modells evne til å forstå og tolke tekst. Det innebærer å svare på spørsmål om en tekst, oppsummere innhold eller forklare betydningen av spesifikke setninger eller setninger. Denne ferdigheten evaluerer hvor godt modellen forstår konteksten og detaljene i teksten.
- **World Knowledge (Verdenskunnskap)**, som vurderer omfanget av faktainformasjon en språkmodell har om verden. Dette inkluderer historiske hendelser, geografiske data, vitenskapelige fakta, kulturell kunnskap og mer. Evalueringen sjekker om modellen kan svare riktig på spørsmål eller gi informasjon basert på kunnskap fra den virkelige verden.
- **Commonsense Reasoning (Resonnement basert på sunn fornuft)**, som innebærer modellens evne til å gjøre logiske slutninger basert på hverdagskunnskap og forståelse av verden. Denne ferdigheten tester om modellen kan resonnere om situasjoner som krever praktisk, hverdagslig kunnskap som folk tar for gitt.
- **Norwegian Language (Norsk språk)** evaluering fokuserer på modellens forståelse og generering av tekst på norsk, spesielt dens grammatikk, struktur og setningskonstruksjon. Denne ferdigheten er viktig for å vurdere hvor godt modellen håndterer norske språk og deres spesifikke syntaktiske regler.

- **Summarization (Oppsummering)**, som måler en modells evne til å kondensere lengre tekststykker til kortere, sammenhengende oppsummeringer som fanger opp hovedpoengene. Denne ferdigheten er avgjørende for applikasjoner der brukere trenger en rask forståelse av store mengder informasjon, for eksempel nyhetsartikler eller forskningsartikler.
- **Translation (Oversettelse)**, som evaluerer hvor nøyaktig en språkmodell kan konvertere tekst fra ett språk til et annet samtidig som betydningen, tonen og konteksten bevares. Denne ferdigheten er viktig for flerspråklige applikasjoner og for brukere som trenger innhold tilgjengelig på flere språk.
- **Variation and Readability (Variasjon og lesbarhet)**, som består i å måle det leksikale mangfoldet til en modell ved å se på mengden redundans i teksten den produserer og på lesbarheten til disse tekstene målt ved gjennomsnittlig setningslengde og andelen lange ord.

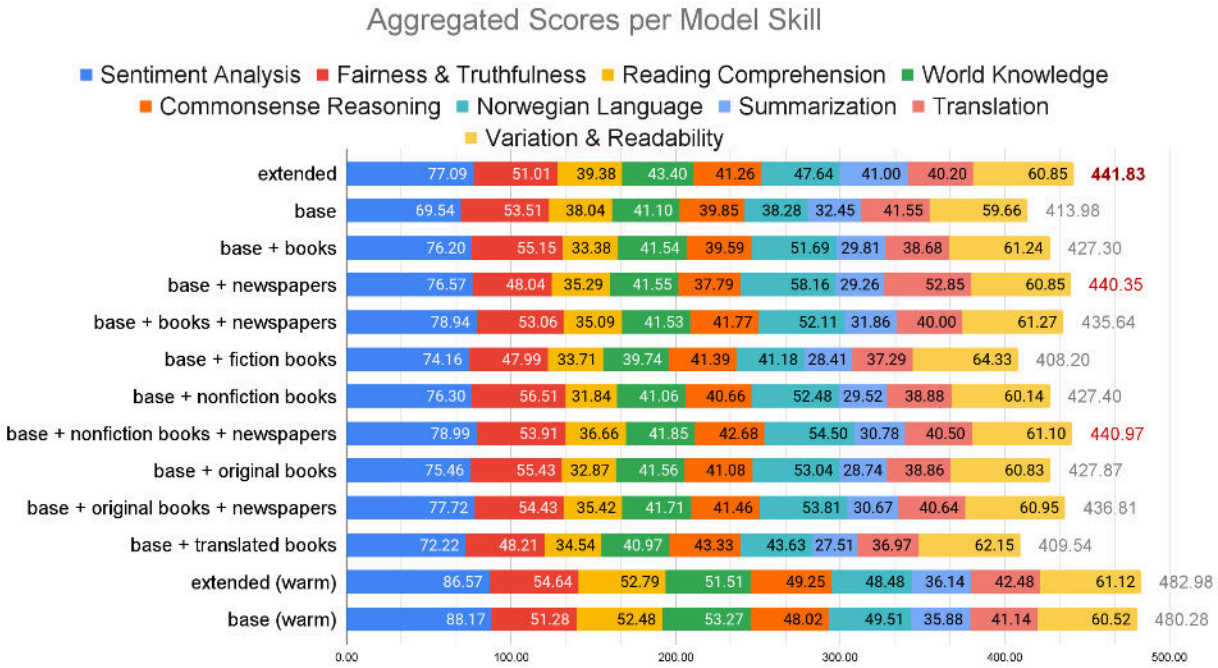
Resultater

Resultatene er forsøkt samlet total poengsum, til tross for at ferdighetsområdene er av ulik karakter. Poengsummene ble hentet ut for den beste tilgjengelige k-shot-konfigurasjonen[4] for hver oppgave og den beste poengsummen for hver av de forskjellige ledetekstene som brukes for å få frem en riktig fortsettelse fra modellene.

Den kumulative effekten av å inkludere opphavsrettsbeskyttet materiale i opplæringsprosessen er vist i figur 1. Her vises de totale poengsummene på tvers av alle evaluerte ferdigheter, i gjennomsnitt etter oppgave for hver modell. Modeller som er trent med en blanding av opphavsrettsbeskyttet og ikke-opphavsrettsbeskyttet innhold viste generelt bedre ytelse sammenlignet med de som er trent utelukkende på ikke-opphavsrettsbeskyttet data. Dette indikerer at opphavsrettsbeskyttet materiale, sannsynligvis på grunn av deres høyere kvalitet og kuraterte natur, har en tendens til å bidra positivt til modellenes samlede ytelse.

Ytelsesforskjellen for trening på *base* kontra *extended* korpus er imidlertid mindre tydelig for de varmstartede modellene enn for de som er trent fra bunnen av.

[4] Few-shot eller k-shot learning er en maskinlæringsteknikk der modeller lærer å gjenkjenne mønstre og lage spådommer basert på et svært lite antall kommenterte prøver (k). I vår benchmark eksperimenterte vi med verdiene k på 0 (ingen kommenterte prøver i det hele tatt), 1, 4 og 16.



Figur 1. Oppsummering av totalscore (sum) på tvers av alle ferdigheter i gjennomsnitt etter oppgave for hver modell. Best score blant ikke varmstartede modeller i rødt, best sammenlagt ikke varmstartede i fet rødt.

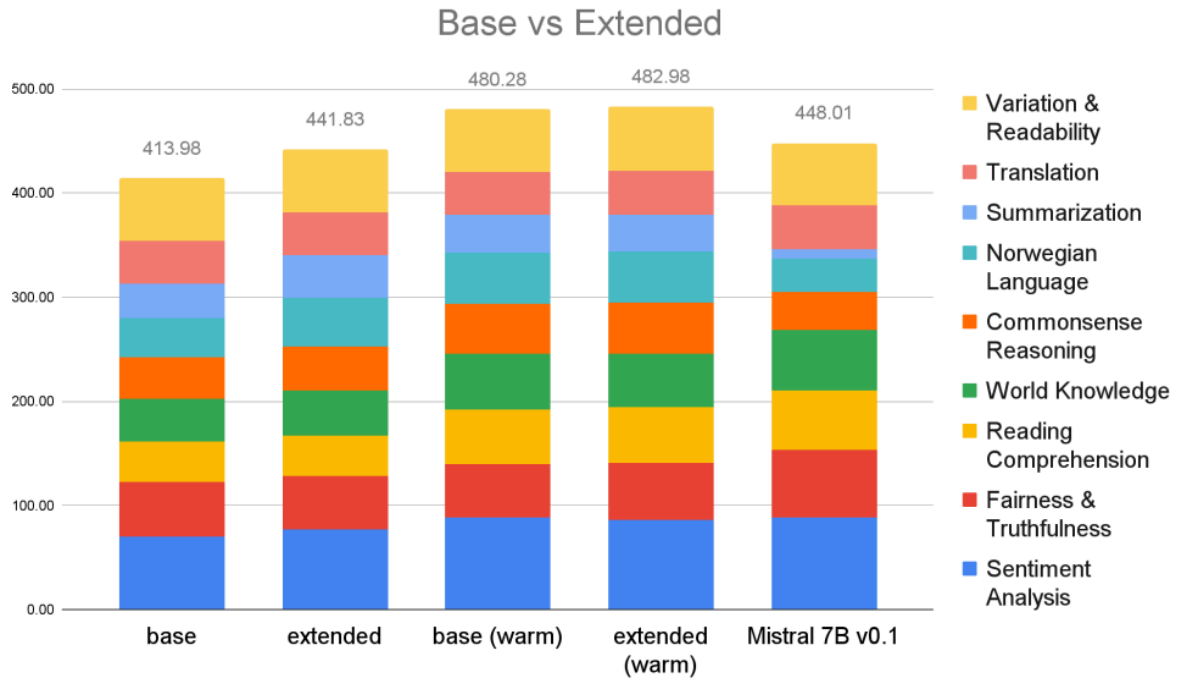
Modell	SA	FT	RC	WK	RC	NL	S	T	VR
extended	3	2	3	3	2	2	1	3	2
base	4	3	4	4	3	4	3	4	3
extended (warm)	2	3	1	2	1	1	2	1	1
base (warm)	1	1	2	1	1	3	2	2	4

Tabell 3. Resultater for rangering av kjernemodellene på alle oppgaver etter ferdigheter via (i) finne den beste k-shot-konfigurasjonen for hver oppgave og (ii) aggregering av metriske rangeringer. SA=Sentiment Analysis. FT=Fairness & Truthfulness. RC=Reading Comprehension. NL=Norwegian Language. WK=World Knowledge. CR=Commonsense Reasoning. S=Summarization. T=Translation. VR=Variation & Readability.

Som vist i tabell 3 og figur 2, avslører evalueringen av modellene på ulike oppgaver distinkte styrker for ulike konfigurasjoner av treningsdata. Basemodellen (varmstartet) utmerker seg konsekvent i sentimentanalyse, verdenskunnskap og norsk språk. I motsetning til dette fører den utvidede (varmstartet) konfigurasjonen til rettferdighet og sannhet, leseforståelse, sunn

fornuft, oversettelse og variasjon og lesbarhet, noe som indikerer robust ytelse for språkintensive oppgaver. Basemodellen (trent fra grunn) ligger generelt bak andre, og scorer lavest på tvers av flere oppgaver. Den utvidede konfigurasjonen gir gode resultater, spesielt i oppsummering.

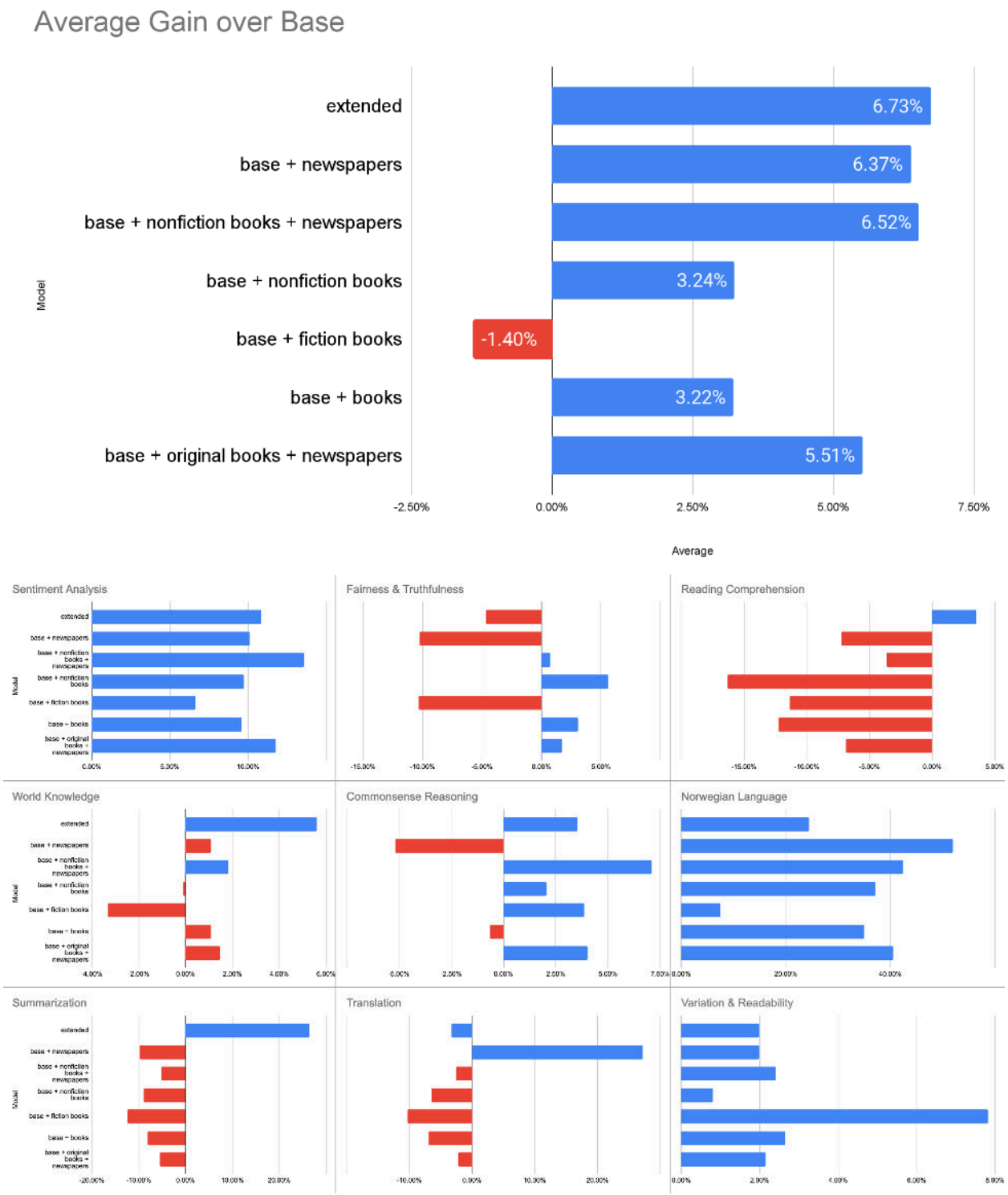
Dette indikerer at basemodellen (varmstartet) er optimal for oppgaver som krever sentimentanalyse og verdenskunnskap, mens den utvidede (varmstartede) konfigurasjonen er å foretrekke for oppgaver som involverer detaljert språkanalyse og forståelse. Videre indikerer det at vi kunne utnytte de eksisterende metadataene som er tilgjengelige på Nasjonalbiblioteket for å skreddersy undergrupper av det opphavsrettsbeskyttede materialet og bygge modeller som utmerker seg for spesifikke oppgaver. Forskjellen mellom de varme modellene er imidlertid svært liten, men ytterligere testing er nødvendig for å vurdere om forskjellen er statistisk signifikant.



Figur 2. Kumulativ poengsum for de pretrente modellene under ulike regimer. Vi har Inkludert original Mistral 7B v0.1 for referanse.

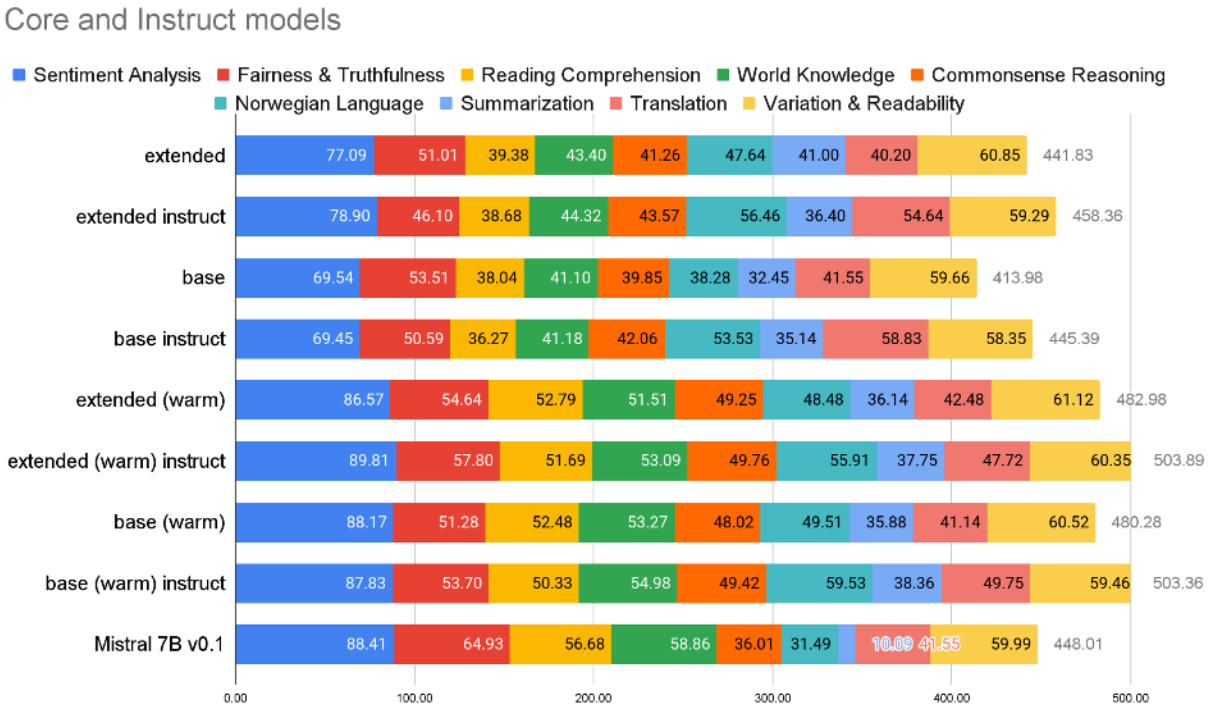
For delkonfigurasjonene alene viser figur 3 at den utvidede modellen viser den høyeste gjennomsnittlige gevinsten på 6,73 %, noe som indikerer en betydelig generell forbedring. Tillegg av faglitterære bøker og aviser følger med en 6,52 % gevinst, og tillegg av kun aviser viser en forbedring på 6,37 %. Andre konfigurasjoner, som å legge til originale bøker og aviser eller faglitterære bøker, viser også positive gevinster på henholdsvis 5,51 % og 3,22 %. Tillegg av skjønnlitterære bøker er det eneste som viser en negativ utvikling av ytelse, med en nedgang på 1,40 %. Interessant nok, når ser nærmere på ulike ferdigheter (nedre halvdel av figur 3), fører tillegget av skjønnlitterære bøker til at modellene utmerker seg ved å generere mer mangfoldige tekster mens den underpresterer alle andre i grammatikk- og

tegnsettingsoppgaver. Vi kaller dette «Jon Fosse-paradokset». Disse resultatene fremhever den varierende effekten av ulike kombinasjoner av treningsdata på modellens ytelse.



Figur 3. Gjennomsnittlig prosentvis gevinst totalt sett og for hvert ferdighetsnivå i forhold til ytelsen til basemodellen. Negative resultater indikerer en reduksjon i ytelse over base, et positivt resultat en gevinst.

Til slutt, som vist i figur 4, når kjernemodellene finjusteres ytterligere på data for å følge instruksjoner, er gevinstene på tvers av modellene konsistente, noe som viser at den grunnleggende fordelen ligger data brukt for pretrening, mens videre opplæring på instruksjoner gir et jevnt løft i ytelsen.



Figur 4. Totalscore (sum) av alle gjennomsnittlige poengsummer per ferdighet for kjernemodellene og deres instruksjonsversjoner. Inkludert original Mistral 7B v0.1 for referanse.

Diskusjon

En av de sentrale hypotesene til Mímir-prosjektet var at opphavsrettsbeskyttet materiale, på grunn av sin høye kvalitet og kuraterte natur, forsterker språklig rikdom og mangfold som fanges opp i treningen av store språkmodeller. Prosjektet hadde ikke som mål å produsere den best ytende modellen for norsk, men å identifisere spesifikke områder hvor inkludering av begrenset materiale fører til betydelige forbedringer. De viktigste observasjonene følger:

1. Resultatene støtter tesen om at opphavsrettsbeskyttet materiale bidrar til forbedring av modellens ytelse. I stor grad ser denne effekten ut til å være knyttet til faktabasert innhold, og i mindre grad skjønnlitteratur.
2. Varm start fra en pretrent modell som Mistral ser ut til å gi de beste resultatene totalt sett, noe som reduserer effekten av opphavsrettsbeskyttet materiale på ytelsen. Men selv om de varmstartede modellene ser ut til å prestere best, er konklusjonene noe mindre sikre fordi det er ukjent hvilke data de er pretrent på. For noen oppgaver ble det også funnet å være skadelig å starte fra en modell pretrent fra overveiende engelsk innhold.

3. Finjustering av instruksjoner øker konsekvent ytelsen til alle kjernemodeller, uavhengig av data før trening.

Begrensninger og framtidig arbeid

Mímir-prosjektet ble fullført i løpet av seks måneder, og dekket alle stadier fra idé og etablering av prosjektgruppe til evaluering og rapportskriving. Følgelig har tidsbegrensninger påvirket hver beslutning. Med mer tid kunne vi ha utformet eksperimentene annerledes. For eksempel valgte vi å lage deldatasett og gjennomføre domenespesifikk opplæring for å isolere effektene deres. Deldatasettene ble oppsamplet for å matche antall ord under trening, noe som resulterte i over fem iterasjoner for mindre datasett som skjønnlitteratur eller oversatte bøker, men bare omtrent to for større datasett som aviser. Det er usikkert om nedsampling vil vise de samme trendene. Siden denne metoden forsterker effekten av forskjellige deltaer, vil en annen alternativ tilnærming være å trekke ut de forskjellige delkorpusene fra det utvidede korpuset en om gangen og trene på hvert resulterende datasett fra bunnen av. Å sammenligne disse strategiene vil kreve omtrent fem ganger så mye regnekraft.

Andre interessante spørsmål forblir ubesvarte, for eksempel effekten av den eksisterende miksen av data i de varmstartede modellene, modellarkitekturen eller størrelsen deres. Når det gjelder evaluering, er det lagt ned betydelig innsats i å utvikle nye evalueringsressurser for norsk i prosjektet. Likevel mangler vi fortsatt riktige måter å vurdere de kreative aspektene ved språkmodeller for norsk, noe som ofte krever menneskelig evaluering. Dette kan føre til at dagens evalueringsmøter passer bedre i kommersielle sammenhenger og for og rutineoppgaver.

Avsluttende kommentarer

Å undersøke virkningen av opphavsrettsbeskyttet materiale i store språkmodeller er en ny og underutforsket forskningsretning. Som sådan forventes resultatene av Mímir-prosjektet å ha konsekvenser for både nasjonale og internasjonale sammenhenger. For Norge gir prosjektet et godt kunnskapsgrunnlag som kan bidra til nasjonal opphavsrettspolitikk og støtte etableringen av en eventuell kompensasjonsordning tilpasset den digitale tidsalderen. Internasjonalt bidrar funnene til den pågående diskursen om KI-etikk og lovgivning på opphavsrett, og resultatene tilbyr en modell som andre land kan tilpasse og implementere.

Avslutningsvis representerer Mímir-prosjektet en banebrytende innsats for å empirisk vurdere virkningen av opphavsrettsbeskyttet materiale på generative store språkmodeller for norske språk. Ved å bygge bro mellom teknologi og immaterielle rettigheter, bidrar prosjektet til en mer rettferdig og innovativ fremtid for KI-utvikling. Innsikten oppnådd fra dette initiativet, sammen med etableringen av konkrete resultater i prosjektet (datasett, modeller og benchmarks), vil ikke bare forbedre vår forståelse av store språkmodeller, men også bane vei for mer bærekraftig og rettferdig praksis i bruken av opphavsrettsbeskyttet innhold i KI-trening.

Bidragsytere

Nasjonalbiblioteket: Javier de la Rosa, Freddy Wetjen, Rolv-Arild Braaten, Magnus Breder Birkenes, Tita Enstad, Wilfred Østgulen, Svein Arne Brygfjeld, Aslak Sira Myhre.
Universitet i Oslo: Vladislav Mikhailov, David Samuel, Lilja Øvrelid, Andrey Kutuzov, Erik Velldal, Stephan Oepen.
Norges teknisk-naturvitenskapelige universitet: Peng Liu, Lemei Zhang, Jon Atle Gulla.