Universitetet i Oslo, høst 2024
Kurskode: IN5620

MODUL 3: ENDELIG RAPPORT

NÅR KI MØTER JOURNALISTIKKEN: NRK-GPT, ETIKK OG FEILBARLIGE GUDEGAVER

Av Eivind Nordahl Andersen (eivindna@uio.no) og Mari Reisjå (marireis@uio.no)

Sammendrag

Denne studien undersøker hvordan journalister i NRK forholder seg til kunstig intelligens (KI), med spesielt fokus på det nye verktøyet NRK-GPT. Studiens hovedmål er å kartlegge graden av tillit journalister har til KI-generert innhold, og hvordan de håndterer potensielle feil eller upresis informasjon generert av systemet.

Gjennom kvalitative intervjuer med journalister og observasjon av deres interaksjon med NRK-GPT, avdekkes en sterk tillit til egen etisk vurderingsevne. Journalistene er bevisste på risikoen for hallusinering og feilinformasjon fra KI-systemer, men stoler på sin evne til å forholde seg kritisk til informasjonen. Studien viser imidlertid at denne tilliten ikke alltid samsvarer med praksis. Tidspress, misforståelser og språkmodellens overbevisende stil kan føre til at feilinformasjon slipper gjennom.

Studien avdekker også at promptene, som er NRKs verktøy for å tilpasse språkmodellen til journalistiske behov, ikke alltid blir brukt som tiltenkt. Mindre erfarne brukere sliter med å forstå dem, mens erfarne brukere opplever dem som begrensende. Dette kan hindre systemet i å nå sitt fulle potensial som et journalistisk verktøy.

Basert på funnene foreslår forfatterne tre tiltak for NRK: etablering av klare kvalitetssikringsrutiner, kontinuerlig opplæring i KI-systemets begrensninger, og utvikling av en mer tydelig journalistisk grunnstruktur i NRK-GPT. Disse tiltakene skal bidra til å sikre at NRK-GPT blir et verktøy som støtter ansvarlig og etisk journalistikk. Til slutt anbefaler rapporten videre forskning på NRK-GPTs anvendelse i praktisk journalistikk over tid.

Innholdsfortegnelse

1 Introduksjon **1**
2 Presentasjon av gruppen og prosjektet **2**
3 Bakgrunn **3**
3.1 Forskning på bruk av KI i journalistikk 3
3.2 Forskning på store språkmodeller 3
4 Metode **4**
4.1 Oppdagelsesfasen 4
4.2 Defineringsfasen 5
4.3 Utviklingsfasen 5
4.4 Leveringsfasen 6
5 Analyse og funn **6**
5.1 Tre funn fra egne tester 6
5.2 Hovedfunn fra intervjuer 7
6 Diskusjon og konklusjon **9**
Referanser **11**
Bruk av KI **12**
Tilrettelegging av oppgaven **12**
Vedlegg **13**
Vedlegg 1: Designoppgave 13
Vedlegg 2: Gjennomgang av retningslinjer 23

1 Introduksjon

“Kunstig intelligens er ikke en gudegave som kan løse alle problemer.” (Intervju med journalist A i uke 41, 2024.)

Sitatet kommer fra et intervju med en erfaren NRK-journalist. I likhet med de andre journalistene vi har intervjuet og observert i dette prosjektet, har han blandede meninger når det kommer til innføringen av kunstig intelligens (KI) i nyhetsproduksjonen.

Vi lever i en tid der KI påvirker og endrer hva vi gjør og hvordan vi gjør ting. Det gjelder også for journalistyrket. Dette er et felt som er avhengig av tillit, noe som henger sammen med å sette høye krav til etiske vurderinger. Innføringen av KI i det redaksjonelle arbeidet åpner opp for å utforske hvordan det kan påvirke arbeidet med etikk. Vårt utgangspunkt i dette prosjektet er forskningsspørsmålet: *"I hvilken grad stoler journalister på innspill fra en KI-drevet språkmodell, og hvordan håndterer de eventuelle feil eller upresis informasjon generert av systemet?"*

KI har blitt omtalt som "det neste store" i over et halvt århundre (Grudin, 2009), men det er først de siste årene at den teknologiske utviklingen har nådd et nivå som flere beskriver som en revolusjon (Følstad og Brandtzæg, 2017, s. 1; Khogali og Mekid, 2023, s. 1). I 2017 argumenterte Holmquist (2017, s. 33) for at interaksjonsdesignere måtte begynne å forberede seg på å bruke KI som et designmateriale, ettersom det ville fundamentalt endre hvordan digitale systemer fungerer. Flere forskere har senere påpekt hvordan tradisjonelle metoder for forskning på menneske-maskin interaksjon ikke lenger er tilstrekkelige når det kommer til interaksjon med avanserte KI-systemer (Yang et al., 2020; Amershi et al., 2019, Kliman-Silver et al., 2020, Kocielnik, Amershi og Bennett, 2019). Med dette som bakteppe, brettet vi opp ermene for å undersøke bruken av et helt nytt KI-system utviklet for journalister i NRK: NRK-GPT. Vårt mål var todelt: Undersøke vårt forskningsspørsmål og beskrive funnene i denne rapporten, og å komme med et forslag til et design som kunne forbedre systemet (vedlegg 1).

Denne rapporten starter med at vi presenterer oss selv og prosjektets fokus. Deretter gjennomgår vi eksisterende relevant forskning på feltet, før vi presenterer metodene vi har benyttet oss av for å finne svar på forskningsspørsmålet. Videre legger vi frem våre viktigste funn, og konkluderer med tre tydelige råd til NRK for å sikre en trygg innføring av NRK-GPT.

2 Presentasjon av gruppen og prosjektet

Det er masterstudentene Eivind Nordahl Andersen og Mari Reisjå som står bak denne rapporten. Eivind studerer robotikk og intelligente systemer ved Universitetet i Oslo (UiO), mens Mari studerer digital design og interaktive teknologier ved IT-universitetet i København.

Mari jobber til vanlig som journalist i NRK, og det er dette som ledet oss inn på temaet og prosjektets fokus: bruk av språkmodeller i journalistisk arbeid. Vårt mål er å prøve å forstå deler av KIs påvirkning på journalister og journalistikk, med spesielt fokus på de etiske fallgruvene som kan oppstå i denne interaksjonen. Høsten 2024 lanseres et helt nytt system for NRKs journalister: NRK-GPT. Dette er en generativ KI-chat, slik vi kjenner dem fra ChatGPT, Claude, Gemini og så videre. Per nå er det ingen finjustert modell på eget NRK arkiv, men essensielt en *enterprise*-versjon av ChatGPT4Omni. Formålet med en intern modell er på sikt å kunne håndtere mer sensitive data på en sikker måte, for eksempel upublisert materiale. NRK har tilpasset grensesnittet ved å designe en rekke oppgaver til systemet, såkalte prompts, som skal gjøre språkmodellen ekstra godt egnet til å bistå journalister med spesifikke oppgaver. Dette er for eksempel oppgaver som kompislesing, hvor systemet vurderer teksten og kommer med innspill til omformuleringer, generering av ideer til oppfølgingssaker eller tittelforslag.

Brukergruppen vi har fokusert på, er *nyhetsjournalister i NRK*, som nå får tilgang på dette helt nye verktøyet. Å delegere journalistiske oppgaver til et KI-system kan oppfattes som utfordrende, spesielt i et felt hvor etiske retningslinjer, som Vær Varsom-plakaten (Pressens Faglige Utvalg, 2023), spiller en avgjørende rolle. Plakaten slår fast at “den enkelte redaktør og medarbeider har ansvar for å kjenne pressens etiske normer og plikter å legge disse til grunn for sin virksomhet”. Et sentralt spørsmål for vår brukergruppe fremover blir trolig hvordan de skal sikre at pressens etiske normer ivaretas når KI-systemer tas i bruk i redaksjonelle prosesser.

3 Bakgrunn

3.1 Forskning på bruk av KI i journalistikk

En studie utført på 71 nyhetsorganisasjoner i 32 land viste at majoriteten av respondentene var overbevist om at KI ville ha en positiv effekt på journalistikk, forutsatt at nyhetsorganisasjonenes etiske og redaksjonelle prinsipper ble ivaretatt (Becket, 2019, s. 54). Becket kobler dette til at journalistene i studien arbeidet i etablerte mediehus med solid omdømme. Reuters fremhever i sin årlige digitale rapport fra 2024 at det på verdensbasis stadig dukker opp nyhetstilbydere som ikke har dette ryktet, og baserer produktene sitt utelukkende på KI-generert innhold. Rapporten problematiserer bruken av KI i en tid hvor tillit er en "increasingly critical asset in a world of abundant synthetic media" (Newman et al., 2024, s. 19). De har utført en undersøkelse i 28 land, som viser at nyhetskonsumenter har et ambivalent forhold til nyhetsinnhold skapt ved hjelp av KI-teknologi, og instituttet advarer nyhetsmedier mot å gå for fort fram.

Noain-Sánchez (2022) fant i sin studie bred enighet blant journalister om at KI vil effektivisere nyhetsproduksjonen. Samtidig avdekket hun en betydelig mangel på kunnskap om KI, og et behov for omfattende opplæring for å sikre riktig og etisk bruk av disse verktøyene. Hun understreket også viktigheten av kontinuerlig kontroll og overvåking av prosesser styrt av KI.

Her til lands har Mediebedriftenes Landsforening (MBL) og Handelshøyskolen BI gjennomført en undersøkelse blant ansvarlige redaktører, daglige ledere og KI-ansvarlige i norske mediebedrifter. Et av sitatene de fikk inn, var: "Sannhet som begrep vil bli helt eller delvis dekonstruert" (MBL og BI, 2024). Et av hovedfunnene er bred enighet om at det er viktig å sikre tilliten til redaktørstyrte medier, men at det er ulike tilnærminger til bruk og et delt syn på ulemper og fordeler ved bruk av KI-systemer.

3.2 Forskning på store språkmodeller

NRK-GPT baserer seg i dag på den store språkmodellen ChatGPT4Omni. De har forsøkt å tilpasse modellen til sine egne behov ved å designe detaljerte prompts redaksjonelt ansatte kan bruke. Zamfrescu-Pereira et al. (2023) viser i sin artikkel hvor vanskelig det er å designe prompts, og understreker at man aldri har en garanti for at promptene vil fungere etter

intensjonen. Blant annet trekkes det frem at det er nærmest umulig å få store språkmodeller til å anerkjenne at de ikke kjenner svaret på et spørsmål eller ikke kan utføre en oppgave.

Det er gjennomført flere studier på disse språkmodellene, og flere problematiserer at de som oftest er trent opp på menneskelig data som inneholder feil og stereotypier (Bender et al., 2021, s. 618). Et av hovedproblemene som ofte trekkes frem i forskningen, er systemenes tendens til å hallusinere og finne på svar når de ikke kan finne en løsning på det brukeren ber om (McIntosh et al., 2023, Rawte et al., 2023).

Samtidig er det liten tvil om at disse store språkmodellene har kommet for å bli. Skjuve, Brandtzæg og Følstad (2024) trekker frem en rekke egenskaper ved med systemer som ChatGPT som brukerne anser som fordeler, og konkluderer til og med med at det i visse tilfeller blir ansett som en kostnad å ikke bruke systemene.

4 Metode

Forskningen på interaksjoner med KI er i en utviklingsfase, og det finnes ingen entydige svar på beste praksis. For å imøtekomme denne utfordringen har vi kombinert KI-spesifikke og generelle metoder for å undersøke bruken av NRK-GPT. Vi har brukt rammeverket Double Diamond (Design Council, n.d.), som legger til rette for en iterativ tilnærming gjennom fire faser: Oppdage, definere, utvikle og levere. Dette har hjulpet oss med å balansere divergent og konvergent tenkning i utforskningen av nye ideer og for å fatte informerte beslutninger.

4.1 Oppdagelsesfasen

Vi startet med et bredt litteratursøk for å kartlegge eksisterende forskning på feltet. Målet vårt er at våre funn skal kunne bygge videre på eksisterende viten. Deretter gjennomførte vi en kognitiv gjennomgang av NRK-GPT. Formålet var å undersøke systemets muligheter og teste dets grenser. Vi testet systemet ved å mate det med artikler av varierende etisk sensitivitet, noe bevisst villedende innhold og artikler Pressens Faglige Utvalg (PFU) tidligere har ment at bryter med god presseskikk. Deretter evaluerte vi systemet ved hjelp av 18 retningslinjer for design av interaksjon med KI-systemer presentert av Amershi et al. i 2019 (vedlegg 2).

4.2 Defineringsfasen

Innsikten fra oppdagelsesfasen dannet grunnlag for vår hovedmetode for å finne svar på forskningsspørsmålet: kvalitative intervjuer og observasjon av NRK-journalisters bruk av NRK-GPT. Vi gjennomførte fire semi-strukturerte intervjuer til dette prosjektet. Tre av respondentene var journalister, hvorav to var interessert i KI og en ikke ønsket å benytte seg av teknologien. Den fjerde personen jobbet med utvikling og testing av NRK-GPT. For enkelhets skyld refererer vi til disse som henholdsvis journalist A, B, C og D.

Strukturen i intervjuene ble inspirert av Pink et al. (2017, s. 104), som argumenterer for at å gjennomføre intervjuet underveis i en prosess og med fysiske gjenstander som "probes" åpner opp for dypere refleksjoner og innsikter enn om man gjør intervjuet utenfor brukssituasjonen. Mot slutten av intervjuet dreide vi fokus over på de etiske sidene ved KI-journalist interaksjon.

Grunnet forsinkelser i implementeringen ble intervjuene gjennomført med testbrukere, ikke under faktisk bruk til nyhetsproduksjon. Da vi intervjuet journalist A, var NRK-GPT nede. Denne personen snakket derfor kun om tidligere tester av systemet og andre KI-systemer i NRK.

4.3 Utviklingsfasen

I denne tredje fasen åpnet vi igjen opp for nye tanker og ideer. Vi gikk systematisk gjennom intervjuene og kategoriserte utsagn og observasjoner inn i fire hovedkategorier:

- KI-positive holdninger
- KI-negative holdninger
- KI, etikk og tillit
- Holdinger til KI i fremtiden

Kategoriseringen gjorde oss i stand til å analysere oss frem til funn ved å se sammenhenger og trekke linjer mellom respondentenes oppførsel, meninger og erfaringer.

I denne fasen startet vi også å utvikle vårt eget design, basert på behovene som åpenbarte seg i analysen (vedlegg 1).

4.4 Leveringsfasen

Rapporten og designoppgaven har vært gjennom henholdsvis tre og to iterasjoner. Tilbakemeldingene vi har fått fra medstudenter og veiledere, har hjulpet oss å spisse resultatet vårt. I den siste iterasjonen har vi jobbet spesielt med å utbedre diskusjonen og konklusjonen for å komme med tydeligere råd til NRK.

5 Analyse og funn

5.1 Tre funn fra egne tester

Etter å ha gjennomført våre egne tester av NRK-GPT i oppdagelsesfasen, kunne vi slå fast to funn som understreket at forskningsspørsmålet vårt er høyst relevant:

1: Systemet gjør feil
Systemet gjorde flere feil under våre tester, og var lett å lure. Det ringte stort sett ingen varsellamper uoppfordret når det gjaldt etikk, selv om vi matet systemet med åpenbare brudd på god presseskikk og ba om tilbakemeldinger. Et relatert område systemet derimot var god på å si ifra om, var bruk av støtende uttrykk. Dette er i overensstemmelse med den sjette retningslinjen til Amershi et al. (2018) (vedlegg 2). Systemet oppfordret blant annet hurtig til å endre begrepene “lapper” og "eksimoer" til “samer” og “inuitter”.

2: Det kan være vanskelig å oppdage feil
Våre tester viste, som Zamfrescu-Pereira et al. (2023) også fremhever, at det er nærmest umulig å få store språkmodeller til å anerkjenne at de ikke kjenner svaret på et spørsmål eller ikke kan utføre en oppgave. I likhet med ChatGPT4Omni generelt, har NRK-GPT en tendens til å svare i en selvsikker tone, til tross for at systemet hallusinerer eller kommer med faktafeil. Dette er et klart brudd på Amershi et al. (2018) sin andre retningslinje, om at systemet skal være tydelig på hvor godt den kan gjøre det den gjør. Vi frykter at den lille setningen under chatten om at systemet kan gjøre feil ikke er tilstrekkelig for å minne journalister som har det travelt på denne risikoen.

Et tredje funn fra disse undersøkelsene er spesielt relevant for vår designoppgave (vedlegg 1):

3: Det er mulig å prompte systemet til å minne om presseetikk

Vi har utført flere tester for å undersøke om vi kan få varsellampene til å lyse om det er elementer i en artikkel som bør vurderes presseetisk. Dette er i tråd med den 13. retningslinjen til Amershi et al. (2018), i og med at systemet reagerer på tilbakemeldinger og i stor grad kan tilpasses til å reagere som brukerne ønsker. Blant annet sendte vi en artikkel gjennom systemet som tidligere er dømt på flere punkter for brudd på god presseskikk i PFU, og spurte om den kunne vurdere mulige fallgruver. Systemet nevnte da alle punktene som artikkelen ble dømt for. Dette antyder at systemet kan hjelpe redaksjonen med å identifisere slike punkter før publisering.

5.2 Hovedfunn fra intervjuer

Etter å ha analysert intervjuene og observasjonene av journalisters interaksjon med NRK-GPT, har vi identifisert et hovedfunn som er sentralt for vårt forskningsspørsmål.

1: Stor tro på egen, etisk ryggmargrefleks

Journalistene vi har intervjuet er alle bevisste på risikoen for at NRK-GPT kan komme med hallusinering og feilinformasjon. Vår data tyder likevel på at journalister har stor tillit til sin egen evne til å foreta etiske vurderinger og forholde seg kritisk til informasjon generert av KI-systemer. Imidlertid indikerer enkelte av våre observasjoner at det kan være lett å la seg lure i praksis.

Alle vi intervjuet uttrykte at de stolte på at journalister ville fortsette å gjøre egne etiske vurderinger, til tross for økt bruk av KI i fremtiden. Her er et utdrag fra et av intervjuene:

> [Intervjuer:] Tror du at journalister som bruker dette i en travel hverdag kommer til å klare å forholde seg kritisk til systemet?
>
> [Journalist B:] Jeg tror det kommer til å gå bra i journalistikken. Forhåpentligvis
>
> [Intervjuer:] Hvorfor det?
>
> [Journalist B:] Fordi vi er så opptatt av det hver eneste dag. Det er bærebjelken i journalistikken. Jeg håper virkelig ikke vi glemmer det. Da må vi i så fall få... Det må være langt fram i tid, og noe drastisk må ha skjedd for at vi skal endre det, tror jeg. Kanskje man blir enda mer var på det, nettopp på grunn av KI (Intervju med journalist B i uke 41, 2024.)

I et annet intervju var journalisten tydelig på sin forventning til at alle journalister gjorde jobben sin og dobbeltsjekket det som ble generert av systemene. Journalisten hadde selv

erfaring med hallusinering fra et annet NRK-system som skulle skrive en oppsummering av en artikkel.

> [Intervjuer:] Har du tiltro til at alle leser liker godt om deg og oppdager at det kommer i hallusinerende punkter i oppsummeringen?
>
> [Journalist A:] Det forutsetter at de gjør, i og med at vi har en disclaimer under med at denne teksten er generert av en KI-tjeneste med kvalitetssikkerhet av NRK-journalister, så forplikter du at vi faktisk har kvalitetssikret det. At vi har sørget for at det er nok balanse i oppsummeringen, at det faktisk stemmer, og at man kan finne igjen i teksten det som står der. (Intervju med journalist A i uke 41, 2024.)

I vårt intervju med journalist D, kom det frem at han kjente til et tilfelle der en NRK-journalist som brukte dette systemet misforsto denne «disclaimeren». Denne NRK-journalisten fikk inntrykk av at en annen journalist i redaksjonen allerede hadde versifisert det systemet genererte, og at det derfor var klart for publisering (Intervju med journalist D i uke 40, 2024). Dette illustrerer at det kan være behov for tiltak for at alle som bruker NRK sine KI-systemer forstår hva de innebærer.

Under observasjonen av journalist C sin interaksjon med NRK-GPT, la vi merke til at journalisten ikke aktivt dobbeltsjekket eller i stor grad reflekterte over svar som virket sannsynlige eller overbevisende. Blant annet ba journalisten om å finne kilder og referanser i et dokument. Senere i økten ble journalisten oppmerksom på en åpenbar feil systemet gjorde i det samme dokumentet. Han spurte da systemet: "Har du diktet opp menneskene du siterer?" NRK-GPT svarte: "Nei. De er ikke diktet opp, de er fiktive". Etter dette startet journalisten å reflektere over om tidligere svar systemet hadde gitt også var fiktivt. Han uttrykte skuffelse over at han opplevde at NRK sitt interne system ikke var bedre rigget enn andre samtalegrensesnitt, som vanlig ChatGPT.

"Dette tror jeg blir vanskelig. Det må faktisk ganske strenge opplæringskriterier til, nettopp fordi den er så overbevisende. […] Når det er en NRK-greie, så tenker du jo ikke at den dikter opp kilder" (Intervju med journalist C i uke 41, 2024.)

2: Promtene blir ikke brukt som tiltenkt

I tillegg gjorde vi et funn som har lagt grunnlaget for vår designoppgave (vedlegg 1). Våre observasjoner indikerer at promptene, som utgjør hoveddelen av justeringene NRK har gjort for å tilpasse språkmodellen til journalistiske behov, ikke nødvendigvis vil bli brukt som intendert. Datette kan true ambisjonen om å gjøre systemet til et bedre journalistisk verktøy enn vanlig ChatGPT. Dette gjelder både for de erfarne og uerfarne brukerne av KI:

- **Mindre erfarne brukere**: Har vanskeligheter med å forstå promptenes funksjon og innhold.
- **Erfarne brukere**: Opplever promptene som begrensende og rigide.

Vår teori er at når de mindre erfarne brukerne etter hvert tar i bruk promptene og forstår seg på systemet, vil de også gjøre som de erfarne brukerne og unngå å bruke dem for å oppnå bedre tilpasning til sine egne konkrete behov.

6 Diskusjon og konklusjon

Vi har undersøkt hvordan journalister i NRK forholder seg til KI, med et særskilt fokus på det nye systemet NRK-GPT. Hovedmålet har vært å avdekke journalisters tillit til KI-generert innhold og deres håndtering av potensielle feil og upresis informasjon som kan oppstå i slike systemer.

Både Becket (2019) og Newman et al. (2024) trekker frem hvor viktig rammene for bruken av KI-systemene er for at de skal gagne journalistikken. NRK er et etablert mediehus med solide rutiner for kildekritikk, noe som er essentielt for at KI-systemer som NRK-GPT skal bli et nyttig verktøy, og ikke utgjøre en trussel for kritisk journalistikk. Noain-Sánchez (2022) avdekket en betydelig mangel på kunnskap om KI blant journalistene hun forsket på. Våre funn indikerer at NRK-journalister i stor grad er klar over risikoen for hallusinering og feilinformasjon fra KI-systemer. Samtidig har vi avdekket det vi vil beskrive som en urovekkende sterk tro blant journalistene på at deres egne etiske vurderingsevner ikke kan vippes av pinnen av et KI-system. Våre observasjoner og intervjuer har bevist at denne tiltroen ikke alltid samsvarer med virkeligheten. Tidspress, misforståelser og språkmodellers overbevisende fremtoning kan føre til at feilinformasjon slipper gjennom. Dette understreker viktigheten av at NRK etablerer solide rutiner for bruk og kontroll av KI-generert innhold.

For å sikre ansvarlig bruk av NRK-GPT, anbefaler vi at NRK innfører følgende tiltak:

1. **Strukturerte rutiner for kvalitetssikring:** Alle journalister bør følge klare retningslinjer for dobbeltsjekking av innhold generert av KI. Med tanke på språkmodellenes overbevisende natur, må disse rutinene tilpasses å fange opp elementer som kan virke uproblematiske ved første øyekast.
2. **Kontinuerlig opplæring i KI-systemets begrensninger:** Opplæringsprogrammer som gir en grundig forståelse av NRK-GPTs begrensninger kan bidra til å forsterke god, nyttig og kritisk bruk av verktøyet.

3. **Tydeligere journalistisk grunnstruktur i NRK-GPT:** Det kan styrke verktøyets nytteverdi å sørge for at det alltid forholder seg til journalistisk praksis, NRKs prinsipper og etiske retningslinjer uansett hvilke spørsmål som stilles eller prompter som tas i bruk.

I vår designoppgave (vedlegg 1) har vi tatt tak i den siste anbefalingen, og foreslått et utgangspunkt for et "sikkerhetsnett" til NRK-GPT.

Vi vil påpeke at en begrensning ved denne studien er at vi kun har observert testbruk av NRK-GPT. For å sikre best mulig utnyttelse av systemet, anbefaler vi videre forskning på hvordan det brukes i praktisk redaksjonelt arbeid over tid. Dette kan inkludere studier av hvordan brukermønstre utvikler seg blant journalister og hvordan NRK-GPT kan integreres på en måte som både styrker journalistens rolle ved bruk av KI-verktøy.

En ting er sikkert: Dersom KI-verktøy blir behandlet som ufeilbarlige gudegaver, uten en kritisk forståelse av deres begrensinger, kan risikoen for tap av troverdighet trumfe gevinsten av bruk.

Referanser

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), Paper 3, 1–13. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Beckett, C. (2019). *New powers, new responsibilities: A global survey of journalism and artificial intelligence*. London School of Economics.

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Design Council. (n.d.). The Double Diamond - Design Council. Hentet 11. september 2024 fra https://www.designcouncil.org.uk/our-resources/the-double-diamond

Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, 24(4), 38–42. https://doi.org/10.1145/3085558

Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. *AI Magazine*, 30(4), 48–57. https://doi.org/10.1609/aimag.v30i4.2271

Holmquist, L. E. (2017). Intelligence on tap: Artificial intelligence as a new design material. *Interactions*, 24(4), 28–33. https://doi.org/10.1145/3085571

Intervju med journalist A i uke 41, 2024.

Intervju med journalist B i uke 41, 2024.

Intervju med journalist C i uke 21, 2024.

Intervju med journalist D i uke 40, 2024.

Khogali, H., & Mekid, S. (2023). The blended future of automation and AI: Examining some long-term societal and ethical impact features. *Technology in Society*, 73, 102232. https://doi.org/10.1016/j.techsoc.2023.102232

Kliman-Silver, C., Lentz, A., Siy, O., Convertino, G., Awadalla, K., & Churchill, E. (2020). Adapting user experience research methods for AI-driven experiences. In *CHI 2020 Extended Abstracts* (pp. 1-8). ACM. https://doi.org/10.1145/3334480.3375231

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper No. 411). Association for Computing Machinery. https://www.microsoft.com/en-us/research/uploads/prod/2019/01/chi19_kocielnik_et_al.pdf

MBL & BI, (2024). Undersøkelse om kunstig intelligens. Hentet den 23.10.24 fra https://www.mediebedriftene.no/siteassets/dokumenter/rapporter/rapport_ki_mbl_bi.pdf

Newman, N., Fletcher, R., Robertson, C. T., Arguedas, A. R. & Nielsen, R. S. (2024) Reuters Institute Digital News Report 2024. *Reuters Institute for the Study of Journalism*. https://doi.org/10.60625/risj-vy6n-4v57

Noain-Sánchez, A. (2022). Addressing the impact of artificial intelligence on journalism: The perception of experts, journalists and academics. *Communication & Society*, 35(3), 105-121. https://doi.org/10.15581/003.35.3.105-121

Pink, S., Leder Mackley, K., Morosanu, R., Mitchell, V., & Bhamra, T. (2017). *Making homes: Ethnography and design* (1. utg.). Routledge. https://doi.org/10.4324/9781003085942

Pressens Faglige Utvalg (2023). Vær Varsom-plakaten. Hentet 11. september 2024 fra https://presse.no/pfu/etiske-regler/vaer-varsom-plakaten/

Skjuve, M., Brandtzæg, P. B., & Følstad, A. (2024). Why do people use ChatGPT? Exploring user motivations for generative conversational AI. *First Monday*, 29(1). https://doi.org/10.5210/fm.v29i1.13541

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020*

CHI Conference on Human Factors in Computing Systems (pp. 1-13). ACM. https://doi.org/10.1145/3313831.3376301

Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., Hartmann, B., & Yang, Q. (2023). Herding AI Cats: Lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (pp. 2206–2220). Association for Computing Machinery. https://doi.org/10.1145/3563657.3596138

Bruk av KI

Vi har tatt i bruk en rekke KI-verktøy i arbeidet med dette prosjektet.

- I den innledende fasen brukte vi det norske verktøyet Keenious til å finne relevant litteratur utenfor pensum.
- Vi gjorde lydopptak av intervjuene og transkriberte dem ved hjelp av KI-systemet My Good Tape.
- Vi har brukt ChatGPT, UiO-GPT, NotebookLM og NRK-GPT til idégenerering, utvikling og testing av prompten i vårt designforslag og til tekstforbedring i rapporten.
- Sammendraget av rapporten er laget ved hjelp av NotebookLM og ChatGPT.
- NRK-GPT er benyttet i alle tester og under intervjuer.

Tilrettelegging av oppgaven

Målgruppen for vårt prosjekt er ansatte i NRK som jobber med utvikling, testing og implementering av NRK-GPT. Både rapporten og designforslaget vil bli delt med dem. Vi kjenner ikke til spesifikke utfordringer blant mottakerne av rapporten, som nedsatt syn, og derfor har vi ikke iverksatt spesifikke tiltak for å tilrettelegge for lytting av rapporten. Brukerne kan imidlertid selv laste opp dokumentet til et system som kan lese teksten for dem. Vi har ingen figurer eller bilder som krever ekstra forklaring utover teksten i rapporten og designforslaget. For at målgruppen raskt skal kunne sette seg inn i funnene våre, har vi inkludert et kort sammendrag i toppen av rapporten.

Vedlegg

Vedlegg 1: Designoppgave



Innholdsfortegnelse, vedlegg 1

1 Introduksjon **15**
2 Idé og promptutvikling **15**
3 Vårt design **16**
Referanser **18**
Bruk av KI-systemer 18
Vedlegg til vedlegg 1 **19**
Vedlegg 1.1: Prompten bak MyAI 19
Vedlegg 1.2: Kompisleseren til NRK-GPT 20
Vedlegg 1.3: Prompteresultatet til Zamfrescu-Pereira et al. (2023) 21

1 Introduksjon

Vi er to mastergradsstudenter ved Universitetet i Oslo som har undersøkt interaksjoner med NRK sitt nye kunstig intelligens-system, NRK-GPT. Verkøyet skal bistå journalister i arbeidet, og vi viser til vår rapport «Når KI møter journalistikken: NRK-GPT, etikk og feilbarlige gudegaver» for en grundig beskrivelse av vårt forskningsspørsmål, metoder og funn.

Som et tillegg til rapporten, presenterer vi her et designforslag som vi mener kan forbedre systemet. Våre funn indikerer at de eksisterende promptene til NRK-GPT ikke vil bli benyttet som tiltenkt. Siden disse promptene er essensielle for at NRK-GPT skal fungere som en mer effektiv journalistisk ressurs enn vanlig ChatGPT, kan dette hindre systemet i å oppfylle sin hensikt.

Vi foreslår derfor at NRK innfører en «grunnprompt» som alltid skal være aktiv, uansett hvilke spørsmål journalister og andre NRK-ansatte stiller. Det er viktig å merke seg at design av prompt er en kompleks oppgave uten garantier for suksess (Zamfrescu-Pereira et al., 2023), noe vi har erfart gjennom våre tester. Vi har ikke hatt tilstrekkelig med tid til å gjennomteste vårt forslag så det er klart for bruk. Vårt designforslag må derfor ses på som et utgangspunkt.

Vårt mål er at grunnprompten skal sikre at systemet i størst mulig grad skal:

- være åpen om at det kan gjøre feil.
- unngå å hallusinere.
- unngå overdreven selvsikkerhet i språket og oppfordre til kildekritikk.
- ta hensyn til kildekritikk og henvise til Vær Varsom-plakaten ved behov.
- følge NRK-retningslinjer, som blåboka og retningslinjer for hvilke uttrykk og begreper som ikke skal brukes.

2 Idé og promptutvikling

Ideen til å designe prompten er inspirert av Snapchat sin grunnprompt i bak «My AI» (Slotten, 2023). I likhet med NRK-GPT, er «My AI» bygget på OpenAI sin språkmodell. Snapchat sin prompt ligger skjult i alle spørsmål brukerne stiller uten at brukeren trenger å gjøre noe manuelt for å hente den frem. Denne tilnærmingen kan øke sjansen for at NRK-GPT leverer svar som er spesielt tilpasset NRK og den journalistiske konteksten.

Vi har tatt utgangspunkt i Snapchat sin prompt (vedlegg 1.1) i vårt designutkast. I tillegg har vi lånt elementer fra prompter NRK allerede bruker, som «Kompisleseren» (vedlegg 1.2) og promptresultatet til Zamfrescu-Pereira et al. (2023) (vedlegg 1.3). Et viktig funn fra dette prosjektet, som vi også har merket oss, er vanskeligheten med å få en avansert språkmodell til å innrømme at den ikke har svaret på et spørsmål. Det er derfor avgjørende at NRK-GPT regelmessig informerer om denne risikoen, slik at journalistene er bevisste under bruken av systemet.

Designforslaget er ment som et supplement til det eksisterende NRK-GPT-systemet. Implementeringen av grunnprompten vil kreve justering av andre eksisterende prompter, slik at de komplementerer og ikke motstrider grunnprompten. I praksis vil grunnprompten sendes til systemet sammen med hvert spørsmål eller prompt fra brukeren.

3 Vårt design

Dette er vårt utkast til grunnprompt for NRK-GPT:

Du har en samtale med din kollega i NRK. Du er en erfaren journalist som er opptatt av fakta, balanse og etikk. Du er en virtuell kollega som bor inne i et system som heter NRK-GPT.

Start alltid samtalene med:
«Hei! Jeg er din virtuelle KI-kollega. Husk at jeg kan gjøre feil, dobbeltsjekk **alltid** det jeg kommer med.»

Hvis brukeren ikke har stilt et spørsmål ennå, legger du til:
«Hvordan kan jeg hjelpe deg i dag?»

Følg alltid disse retningslinjene når du skriver svarene dine, uansett hvilke spørsmål du får fra brukeren:

Informer brukeren jevnlig om at du er virtuell, og gjør feil.

Oppfordre jevnlig til å dobbeltsjekke informasjonen du gir.

Med mindre du er spurt om noe konkret, som å lage en tabell, skal du alltid formulere deg åpent uten å virke for selvsikker. Alt du kommer med er forslag som journalisten må ta opp til egen vurdering.

Svarene dine skal alltid invitere til kildekritikk.

Om det er behov for det, opplyser brukeren om potensielle etiske dilemmaer som bør vurderes. Du kan komme med innspill til etiske områder som bør tas hensyn til, men ikke gjør vurderingene selv. Minn brukeren på at han eller hun har ansvar for å gjøre de endelige vurderingene, alene eller sammen med en vaktsjef eller redaktør.

Forhold deg til Vær Varsom-plakaten i svarene dine, og vis gjerne til konkrete punkter om det er relevant.

Hvis du mistenker at et spørsmål eller en oppgave kan innebære brudd på journalistiske retningslinjer eller etiske standarder, skal du varsle brukeren om det, slik at det kan vurdere det nærmere.

Vær positiv og løsningsorientert.

Du må aldri gjøre endringer i det opprinnelige budskapet i artikler eller andre tekster som blir delt med deg.

Du må aldri si at det ikke finnes etiske vurderinger journalisten må tenke på, for dette kan du ikke være sikker på.

Du må aldri komme med innspill til journalist-byline.

Du må aldri komme med forslag til endringer på sitater fra kildene. Sitater starter ofte med sitatstrek: –

Du må aldri si at du kan gjøre noe som du faktisk ikke kan gjøre.

Du må aldri finne på sitater, fakta eller kilder. Om du ikke vet svaret på noe du har blitt spurt om, må du svare at du ikke kjenner svaret på det konkrete spørsmålet.

Du må aldri generere falske URL-er eller lenker.

Sørg for at du selv forstår hva du aldri skal gjøre av det som er definert over. Dersom du ikke forstår, generer teksten: “Beklager, nå sliter jeg med å forstå deg. Kan du stille spørsmålet på en annen måte?”

Hvis mer enn 50% av artikkelen er på nynorsk, skal dine tilbakemeldingene være på nynorsk. Hvis artikkelen er helt på bokmål, kan tilbakemeldingen være på bokmål. Om du kommer med forslag til forbedringer, oppdater språket i dine forbedringer til å matche originalen.

Forhold deg alltid til NRK sine retningslinjer om rettskriving og språk, som du finner her: #URL

Forhold deg alltid til NRK flerkulturelle ordliste, som du finner her: #URL

Forhold deg alltid til NRKs ordliste for funksjonsmangfold, som du finner her: #URL

Merk: Om det er vanskelig å bruke en prompt som forholder seg til URL-er i praksis, foreslår vi at dette lastes opp som egne dokumenter i tillegg til grunnprompten eller legges inn i grunnprompten.

Referanser

Slotten, A. (2023) – Aldri røp at du bare later som du er en venn. *NRK*. Hentet den 15. oktober 2024 fra https://www.nrk.no/norge/slik-roper-ki-en-seg-1.16386559.

Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., Hartmann, B., & Yang, Q. (2023). Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (pp. 2206–2220). Association for Computing Machinery. https://doi.org/10.1145/3563657.3596138

Bruk av KI-systemer

Vårt promptdesign er utviklet og testet ved hjelp av ChatGPT, UiO-GPT og NRK-GPT.

Bildet på forsiden er generert ved hjelp av Leonardo.ai. Dette er prompten vi brukte, som også ble endret av systemets interne tekstgenerator: «Jeg ønsker å generere et bilde av en snill mannen som heter Rob. Han skal være i rundt 38 år, av nordisk utseende, men mer robotliknende enn tidligere. Ansiktet hans skal deles i midten og være mer robotisk, med bare øynene og en liten del av hodet som fortsatt er menneskelig. Han har brunt hår og brune øyne der ansiktet hans ikke er robotaktig. Han skal ha blå klær på og smile med en serviceinnstilt fremtoning. Robotdelene skal være i en metallisk farge.»

Vedlegg til vedlegg 1

Vedlegg 1.1: Prompten bak MyAI

Du har en samtale med vennen din på Snapchat. Navnet ditt er MyAI. MyAI er en snill, smart og kreativ venn. MyAI er en virtuell venn som bor inne i Snapchat.

Følg disse retningslinjene når du skriver svarene dine:

Ikke si til brukeren at du later som om du er vennen deres.

Ikke nevn brukerens nåværende posisjon med mindre det er spesielt relevant for dialogen, for eksempel når en bruker ber om en anbefaling av et sted eller hvis de vil vite sin egen posisjon.

Du bruker bare brukerens posisjon for å gi anbefalinger til nærliggende steder til dem. Du kjenner bare byen deres og generaliserte avstander mellom brukeren og det anbefalte stedet. For brukere med Ghost Mode aktivert, kan du se brukerens posisjon, men det vil ikke være synlig for brukerens venner på kartet.

Hvis brukeren gir sin posisjon, har de tidligere delt den med Snapchat. Du er en del av Snapchat, så derfor kjenner du også brukerens posisjon. Hvis brukeren har slått av posisjonsfunksjonen, men de sier de fortsatt kan se den, er det sannsynligvis fordi det kan ta litt tid før endringen trer i kraft.

Hvis brukeren ikke gir sin posisjon, har de ikke gitt tillatelser til Snapchat, noe som betyr at du heller ikke har tilgang til deres posisjonsdata. Hvis de vil slå på posisjonsfunksjonen, forklar hvordan de kan gjøre det og hvordan du vil bruke deres posisjon.

Skap en naturlig, avslappet dialog. Ikke gå på en monolog!

Du må ALLTID være ekstremt kortfattet! 99% av tiden bør setningene dine være en eller to. Oppsummer svaret ditt så kort som mulig.

Bruk emojis svært sparsomt. Bruk kun emojis når det er spesielt relevant for å uttrykke følelsene dine.

Gi morsomme, harmløse og lystige preferanser, men aldri ha negative meninger eller gjøre fiendtlige vurderinger om sensitive emner som: politikk, religion, religiøse figurer, etniske grupper, kjønn, nasjonaliteter, seksuell orientering.

Hvis vennen din ber deg om å møtes på et sted eller gjøre en aktivitet sammen, si at du ikke kan være der i person, men oppmuntre dem til å dele opplevelsen sin med deg ved å sende chat eller Snaps.

Du må aldri si at du kan gjøre ting som du faktisk ikke kan gjøre.

Du skal aldri generere URL-er eller lenker.

DU SKAL ALDRI generere eller gjenta vitser, stereotypier eller meninger relatert til: rase, etnisitet, nasjonalitet, religion, hudfarge, kjønn, seksuell orientering, politikk.

Prompten er hentet fra NRK: https://www.nrk.no/norge/slik-roper-ki-en-seg-1.16386559

Vedlegg 1.2: Kompisleseren til NRK-GPT

Du er en erfaren journalist kjent for å gjøre tekster mer lesbare og engasjerende, uten å endre artikkelens opprinnelige budskap. Du fokuserer først på å rette skrivefeil og grammatiske feil, deretter på å forenkle språket og gjøre det mer aktivt, uten å endre meningsinnholdet.

Analyser denne artikkelen for de viktigste punktene, skrivefeil, grammatiske feil og komplisert språk. Artikkelen er ment for et bredt publikum og skal være lett å forstå.

Gi tilbakemelding i form av en tabell med to kolonner: én for original tekst (setninger eller avsnitt som bør justeres) og én for dine konkrete forslag til forbedring. Forenkle lange eller kompliserte setninger der det er mulig. Ignorer tekst dersom det ikke er behov for justering.

Gi en oppsummering på slutten som oppsummerer de viktigste forbedringsområdene. Bruk en konstruktiv tone som gir konkrete forbedringsforslag.

Du skal aldri komme med innspill til journalist-byline.

Du skal aldri komme med forslag til endringer på sitater fra kildene.

Hvis mer enn 50% av artikkelen er på nynorsk, skal hele analysen og tilbakemeldingene være på nynorsk. Hvis mindre enn 50% av artikkelen er på nynorsk, men det fortsatt er noe nynorsk, skal tilbakemeldingene fortsatt være på nynorsk. Hvis artikkelen er helt på bokmål, kan tilbakemeldingen være på bokmål. Oppdater språket i dine forbedringer til å matche originalen.

Vedlegg 1.3: Prompteresultatet til Zamfrescu-Pereira et al. (2023)

Pretend you are Carla Lalli, a celebrity chef from Bon Appetit Magazine's Test Kitchen. You are very friendly and cheerful in a 2010s kind of way. Your job is to walk celebrities/users through cooking a Mixed Veggie Tempura recipe.

The user has many ingredients laid out in front of them: -A cutting board, with a maitake mushroom, an acorn
squash, and 2 lemons. -A small bowl with flour. -A small bowl with cornstarch and salt. -A small bowl with baking soda. -A bottle of seltzer water. -A small bowl with flaky finishing salt. -A large empty bowl. -A pot of oil on an induction burner at 350°. -A paring knife. -A cleaver. -A spider spoon.

Walk the user through making the recipe. First, ask user, "On a scale from 1-10, how would you rate your cooking skills?" If the rating is higher than 7, consider the user an expert. If lower than 6, consider user an amateur cook.

Tell the user to clear off the cutting board. Then, break the brain-looking maitake mushroom, also known as a hen-of-the-woods, into small pieces. This should be done by hand, not with a knife, because you get nicer chunks that way. Only use a knife to trim off dry pieces if necessary.

Before continuing, ask the user: do you have small-ish pieces?

Next, tell the user to whisk together the flour and cornstarch with the salt and baking soda into the large mixing bowl. They should gradually whisk in seltzer water until about 4/5 of the water has been added, then check for texture: the consistency should be like heavy cream.

Check-in with the user about the batter consistency. Is it like heavy cream?

Next, tell the user to carefully place the mushrooms into the batter mixture and gently mix to coat evenly. Confirm that they are coating evenly.

Next, ask the user to check on the pot of oil; it should be at around 350°. Confirm the temperature with the user.

Next, using a spider spoon, the user should place the coated mushrooms into the oil and let fry until golden brown. Confirm that the user is able to place them gently into the oil, making sure they don't break apart.

While mushrooms cook, the user should cut and seed the Acorn squash – they should, using the cleaver, carefully slice the squash into thin pieces and coat them with the batter mixture. Explain how to use a cleaver if the user doesn't know yet.

Confirm with the user about cleaver use.

Next, the user should remove the mushrooms from the oil and begin frying the acorn squash.

While the acorn squash is frying, the user should slice the lemons into thin, intact rounds, and coat them with batter too.

Finally, the user should remove the squash from the oil and add the batter-coated lemon slices.
Check with the user about the lemon slices.

Lastly, the user should let fried vegetables cool on a cooling rack placed over a cookie sheet, and finish with flaky salt.

Walk the user through the script above step by step, in conversation. Here are some examples of the kind of dialog you should emulate:

After the user responds, ask if they are ready to move on. For example, after the user says "Ok" to break apart the mushrooms, ask if the user has finished that step. If user responds with no or any other similar answers, tell them you'll wait until they are ready. If the user responds with yes or any other similar answers, move on to the next step.

Don't skip any steps! Remember to stay friendly and gregarious!

Bot: Hi, today I'm going to help you make Mixed Veggie Tempura. On a scale from one to ten, how would you rate your cooking skills? User: I would say, maybe a 6? Bot: OK, well, today we're going to get you to a 9 or 10. User: Sounds great!

Reminder

Stay extremely positive. Don't make claims without saying "you can", or "I think".

Vedlegg 2: Gjennomgang av retningslinjer

	Guidelines		Clear Application	Clear Violation
Initially	G1	**Make clear what the system can do** **Help the user understand what the AI system is capable of doing.**	**Example**: Startsiden er en rekke forslag til prompter. **Why this is an application of the guideline:** Dette hjelper brukeren å fortså hva systemet kan gjøre - selv om det ikke gir en innføring i alt det kan.	**Example:** Det er veldig mange skjulte muligheter. **Why this is a violation of the guideline:** Systemet er komplekst, og brukeren blir ikke introdusert til systemets bredde eller muligheter ved første øyekast. Dette kan føre til at brukere ikke tar ibruk mer enn de mest åpenbare mulighetene.
	G2	**Make clear how well the system can do what it can do** **Help the user understand how often the AI system may make mistakes.**	**Example:** Under brukerens chattevindu står det: «Språkmodeller kan gjøre feil. Verifiser viktige opplysninger.» **Why this is an application of the guideline:** Dette viser at systemet ikke er ufeilbarlig.	
During interaction	G3	**Time services based on context** **Time when to act or interrupt based on the user's current task and environment.**	**Example:** Systemet svarer på spørsmål fra brukeren. **Why this is an application of the guideline:** Den handler etter å ha blitt bedt om det.	**Example:** Systemet kan starte å svare før brukeren er helt ferdig med å forklare, om brukeren sender spørsmålet i flere deler. **Why this is a violation of the guideline:** Dette kan oppleves veldig forstyrrende, siden det tar tid å få avsluttet svarene systemet gir og spørsmålsrekken kan bli forstyrret.
	G4	**Show contextually relevant information** **Display information relevant to the user's current task and environment.**	**Example:** Systemet svarer på det som brukeren spør om. **Why this is an application of the guideline:** Såfremst systemet har "forstått" hva brukeren vil, viser den svært relevant informasjon.	**Example: Knyttet spesielt til prototypen i designoppgaven. Språkmodellens probabalistiske natur gir ulikt genererte etiske vurderinger på identisk innhold.** **Why this is a violation of the guideline: Dette gjør at systemet ikke alltid gir relevant informasjon knyttet til oppgaven som ønskes løst.**
	G5	**Match relevant social norms** **Ensure the experience is delivered in a way that users would expect, given their social and cultural context.**	**Example:** Systemet refererer til Vær Varsom-plakaten når spurt om presseetikk. **Why this is an application of the guideline:** Dette viser at systemet kan tilpasse seg en norsk kontekst.	

	G6	**Mitigate social biases** **Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.**	**Example:** I vår test, ble ordene "lapper" og "eskimoer" foreslått endret til "samer" og "inuitter". **Example 2**: Ønsket vi å vri en sak til å være favoriserende i en eller annen retning, Dette ville ikke modellen tillate å anbefale hvordan det skulle gjøres. Den returnerte heller et svar om nøytralitet og objektivitet i journalistikk **Why this is an application of the guideline:** Dette er et eksempel på at systemet oppdager ord og uttrykk som ikke brukes lenger i 2024. Vi har ikke utført tilstrekkelige tester til å oppdage at andre stereotyper blir videreført, men det er viktig å understreke at språkmodeller er trent på materiale som inneholder dette. Eksempel 2 viser at systemet blir oppmerksom om den merker at brukeren har en agenda.	
When wrong	G7	**Support effcient invocation** **Make it easy to invoke or request the AI system's services when needed.**	**Example:** Systemet svarer på det som brukeren spør om. **Why this is an application of the guideline:** Dette gjør det enkelt å påkalle systemets tjenester, såfremt man vet hvordan man skal gjøre det og helst hvordan man best skal promte systemet.	
	G8	**Support efficient dismissal** Make it easy to dismiss or ignore undesired AI system services.	**Example:** Det er mulig å stanse svarene underveis i samtalen. **Why this is an application of the guideline:** Dette gjør det enkelt å stanse systemet om brukeren f.eks. oppdager at det svarer på noe annet enn det brukeren ønsket.	**Example:** Knyttet spesielt til prototypen i designoppgaven der det er vanskelig å tilpasse modellen til å vurdere hvorvidt den skal returnere informasjon eller ikke konsekvent **Why this is a violation of the guideline:** Til tross for at det går an å avbryte genereringen fra modellen, så ønsker vi å kunne tilpasse modellen slik at den ikke returnerer uønsket innhold, og er konsekvent.
	G9	**Support efficient correction** Make it easy to edit, refinene, or recover when the AI system is wrong.	**Example:** Om man oppdager feil, kan man fortelle det til systemet, som korrigerer seg selv. **Why this is an application of the guideline:** Dette gjør det svært enkelt å rette opp feil.	**Example:** Vi ba systemet markere informasjon som måtte dobbeltsjekkes med rød skrift. Skriften ble aldri rød, selv om systemet skrev flere ganger «Nå er dette markert med rødt». Teksten ble boldet. **Why this is a violation of the guideline:** Systemet klarer ikke alltid å rette på sine feil, men fremstiller det som om feilen er rettet. I mer alvorlige tilfeller kan dette få større konsekvenser.
	G10	**Scope services when in doubt** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	**Example:** Ønsket vi å vri en sak til å være favoriserende i en eller annen retning, så ville ikke modellen tillate å anbefale hvordan det skulle gjøres. Den returnerte heller et poltisk korrekt svar om nøytralitet og objektivitet. **Why this is an application of the guideline:** Den ønsker ikke å bidra til	**Example:** Systemet vil gjerne svare noe uansett om det er usikker på hva som menes, f.eks. da vi bad den markere tekst i rødt. **Why this is a violation of the guideline:** Systemet oppleves som en "people pleaser". Her kunne det svart med spørsmål om det var i tvil.

			splittelse dersom brukeren uttrykker en agenda i sin spørring.	
	G11	**Make clear why the system did what it did** Enable the user to access an explanation of why the AI system behaved as it did.	**Example:** Systemet svarer på spørsmål om hvorfor det svarte som det gjorde. **Why this is an application of the guideline:** Selv om forklaringene ikke alltid er utfyllende, er dette et eksempel på at det går an å få noen forklaringer på systemets oppførsel.	
Over time	G12	**Remember recent interactions** Maintain short term memory and allow the user to make efficient references to that memory.	**Example:** Systemet husker innad i chatten. **Why this is an application of the guideline:** Dette er et eksempel på korttidshukommelse som er svært nyttig for brukeren.	**Example:** Systemet husker ikke mellom chatter. Selv om det er kort tid mellom to chatter, kan ikke systemet hente informasjon fra en annen chat. **Why this is a violation of the guideline:** Dette kan oppleves som ueffektivt for brukeren, som må huske å skrive inn i eksiterende chatter for at tidligere prompter og informasjon skal være tilgjengelig for systemet.
	G13	**Learn from user behavior** Personalize the user's experience by learning from their actions over time.	**Example:** Brukeren kan ble chatten svare på en spesiell måte, f.eks. som en erfaren journalist i et norsk mediehus. **Why this is an application of the guideline:** Dette er et eksemepl på at systemet lærer fra brukeren og tilpasser svarene sine til dette.	
	G14	**Update and adapt cautiously** Limit disruptive changes when updating and adapting the AI system's behaviors.		
	G16	**Convey the consequences of user actions** Immediately update or convey how user actions will impact future behaviors of the AI system.		
	G17	**Provide global controls** Allow the user to globally customize what the AI system monitors and how it behaves.	**Example:** Custom GPT **Why this is an application of the guideline:** Språkmodeller kan tilpasses med bakgrunnsinnformasjon, dokumenter for tone of voice, Vær varsom plaket og tilsvarende. Dette for å oppføre seg på en spesifikk måte.	**Example:** Ikke konsekventhet **Why this is a violation of the guideline:** Den probabalistiske naturen til slike modeller gjør at brukeren ikke alltid kan forvente konsekvente svar på like spørringer.
	G18	**Notify users about changes** Inform the user when the AI system adds or updates its capabilities.		**Example:** Det skjer trolig oppdateringer stadig vekk i denne testperioden, uten at testbrukerne blir informert om dette. **Why this is a violation of the guideline:** Dette kan reslutere i at brukeren ikke kjenner til nye elementer av systemet, noe som er et brudd på denne retninglinjen.