

Tilbakemelding frå ChatGPT i Moodle

Jonas Julius Harang¹, Hans Georg Schaathun¹ og Kenny Solevåg-Hoti²

¹ Institutt for IKT og realfag

² Universitetsbiblioteket

Noregs Teknisk-Naturvitskaplege Universitet, Larsgårdsvegen 2, Postboks 1517, 6025 Ålesund

Samandrag Kunstig intelligens tok verda med storm i desember 2022.

Høgare utdanning har eit stort behov for kostnadseffektiv vurdering, både summativ og formativ, og der er stor etterspurnad etter KI som kan retta og gje tilbakemelding på oppgåver som elles krev fagleg skjønn. Praktiske løysingar lét derimot venta på seg.

Her skal me presentera ei enkel løysing for å gje umiddelbar tilbakemelding på oppgåver i optikk, vha. Moodle og ChatGPT. Me vurderer potensialet basert på faktiske studentsvar og tilbakemeldingar, samt dei etiske innvendingane og behov for vidare utvikling.

1 Introduksjon

Med den sokalte Kvalitetsreformen i 2003/2004 skulle studentane i høgare opputdanning få meir, grundigare og hyppigare tilbakemelding og formativ vurdering. Universitets- og høgskulesektoren er derimot ikkje finansiert for å gje individuell oppfylgjing i særleg omgang, og i alle fall ikkje på dei fyrste årstrinna.

Somme fag har brukt automatisk vurdering i mange år, men tradisjonelle løysingar krev strenge vurderingsreglar, som er mogleg med fleirvalsoppgåver, og på algebraiske og numeriske oppgåver, men er vanskeleg på fritekstsvar. Store språkmodellar har vist seg effektive til å løysa mange generelle problem i naturleg språk, og håpet er at dei også kan brukast til formativ og summativ vurdering på meir eller mindre vilkårlege oppgåver i vilkårlege fag.

Ei særleg utfordring er *umiddelbar* tilbakemelding. Me veit at arbeidsminnet har ei levetid på 20–30 sekund (Baddeley & Hitch, 1974), og dvs. at om tilbakemeldinga tek lengre tid, vil studenten ha gløymt mykje av korleis dei tenkte. Kunstig intelligens gjev ei von om slik umiddelbar tilbakemelding som er uoppnåeleg for ein menneskeleg førelesar.

Der er mange utfordringar før kunstig intelligens faktisk kan brukast i undervising. Der er ingen KI-tenester som er ferdig kvalitetssikra og klart for bruk i undervisning. I denne artikkelen skal me presentera ei praktisk løysing som er utprøvd i liten skala, med alle sine lytar. Dette er eit døme frå *praxis* (SoTL) og ikkje i utgangspunktet organisert som forskning. Hensikta er å dela eit døme på kva som er mogleg å gjera og reflektera over kva som er verd å gjera igjen og kva som krev vidare forskning og utvikling. Koden som vart brukt er publisert oppe på github (Harang & Schaathun, 2025).

Etter ein liten litteraturgjennomgang om vurdering (2) og kunstig intelligens i vurdering (3), vil me presentera det didaktiske forsøket (4) og den tekniske løysinga (5). Deretter går me gjennom og vurderer dei tilbakemeldingane som språkmodellen gjev (6) og drøftar potensialet for vidare utvikling og bruk (7).

2 Definisjonar: vurdering av ulike slag

Vurdering er eit kontroversielt konsept. Det er vanleg å skilja mellom formativ og summativ vurdering, men formativ vurdering er òg tvitydig, og vert gjerne forveksla med kontinuerleg tilbakemelding sjølv om denne er summativ. Hansen og Ringdal (2018) har vist at kontinuerleg summativ tilbakemelding kan ha ein *negativ* formativ effekt, som driv studentane over i ein *performance-avoidance approach* der dei søkjer å unngå å visa kva dei ikkje kan og dermed lærer mindre.

Summativ vurdering er oppsummerande og deskriptiv, og fortel kva studenten kan på det aktuelle tidspunktet. Hensikta er hovudsakeleg eit vitnemål på kva studenten duger til. Formativ vurdering skal derimot *forma* studenten. Hensikta er å endra åtferd, slik at studenten utviklar seg og i framtida vil duga til meir.

Universiteta har tradisjonelt fokusert på summativ sluttvurering, eller det som gjerne vert kalt *high-stake exams*, dvs. store vurderingsøvingar der studenten vert vurdert mot store pensumeningar, gjerne eitt eller to studiesemester, sjølv om 7½-15 studiepoeng er vanlegast i dag. Andre land har i større grad basert karaktersetjinga på fleire og mindre vurderingsøvingar spreidd utover semesteret, og dei siste tjue åra ser ein aukande bruk av slik kontinuerleg vurdering her i landet. Der er i alle fall tre grunnar for denne utviklinga. For det fyrste vert studentane tvungne til å arbeida jamnt. For det andre får dei meir tilbakemelding på eigne prestasjonar. Til sist vert karakterane mindre tilfeldige og mindre avhengig av dagsformen, ved at dei er basert på fleire målepunkt.

Me skal likevel merka oss at dersom den kontinuerlege vurderinga fyrst og framst fortel studentane kor mykje dei kan og ikkje kan, so er ho stadig summativ. Dette er ikkje berre eit spørsmål om karakter eller ikkje. Ein friteksttilbakemelding kan vera like oppsummerande som ein karakter. Formativ tilbakemelding må designast målretta for å gje studenten innsikt ikkje berre i kva dei treng arbeida meir med, men korleis dei lyt arbeida og ikkje minst kva dei har å byggja på (jf. Hansen & Ringdal, 2018). Formativ tilbakemelding må oppmuntra studenten til å halda fram med å forsøkja, og gjerne halda fram med å gjera feil (jf. Boaler, 2015, kap. 1).

For å oppsummera, lét me formativ vurdering tyda tilbakemelding som er transformativ og egna til å endra åtferd for å fremja læring. Summativ tilbakemelding er deskriptiv, og fortel berre kor dugleg studenten er i forhold til læringsmåla og/eller resten av klassa.

Me kan òg skilja mellom umiddelbar og forseinka tilbakemelding. Manuell tilbakemelding er nesten alltid forseinka, og kjem dagar eller veker etter at studenten leverte sitt svar. Det er ikkje interessant for analysen om det tek ein dag eller ein månad. Det vesentlege skiljet for oss er om studenten får tilbakemel-

ding utan mentalt å ha forlate oppgåvekonteksta. Har studenten stadig oppgåva og løysinga si i arbeidsminnet, eller må han grava i langtidsminnet for å forstå tilbakemeldinga? Me definerer difor umiddelbar tilbakemelding som vurdering som kjem medan studenten stadig hugsar løysinga si i arbeidsminnet. Som nemnd gjev det oss ei tidsfrist på om lag eit halvt minutt.

3 Litteratur: vurdering ved kunstig intelligens

Ikkje overraskande er der tett med studiar av vurdering vha. KI i litteraturen. Det er naturleg å skilja mellom store språkmodellar (LLM) slik som me kjenner dei i dag, og meir tradisjonelle formar for KI, slik me kjente dei før ChatGPT vart lansert som gratisteneste i 2022. Store språkmodellar er ofte generelle nok til at dei kan gje nyttig tilbakemelding på vilkårleg tekst i vilkårleg fagdisiplin. Meir tradisjonelle formar for KI kan ofte vera meir presis og etterprøvbar, men krev at modellane vert trent eller programmert spesifikt det aktuelle problemet. Me treng ikkje her å skilja mellom maskinlæring og regelbasert KI, men skiljet mellom generell generativ KI og problemspesifikke modellar er viktig. Dei generelle språkmodellane støttar umiddelbart eit langt større repertoar av oppgåvetypar og tema.

3.1 KI før ChatGPT

Både Cope mfl. (2020) og Bernius mfl. (2022) slår fast at maskin-generert og maskin-assistert tilbakemeldingar gjer det mogleg å vurdera studentarbeid i eit omfang som langt forbigår alle manuelle metodar. Potensialet for umiddelbar tilbakemelding kan verka endelaust. Samtidig påpeikar dei begge at kvaliteten på tilbakemeldingane framleis krev menneskelig innsats. Liknande funn kjem også fram i den systematiske litteraturgjennomgangen til Gao mfl. (2024) som tek føre seg mange studiar frå tida før ChatGPT. Dei fortel at studiane ofte rettar merksemda si mot tekniske mål, snarare enn å sjå på korleis tilbakemeldingane faktisk styrkjer læring. Modellane ser heller ikkje ut til å kunne støtte djupare læring i særleg grad og gjev lite ny innsikt i studentsvara. Der er stadig behov for menneske som pedagogar og kan henda rekk det med automatisk gjenkjenning av akseptable svar (Gao mfl., 2024). Vurderingsansvarlige kan ta seg av resten.

3.2 Store språkmodellar

Sidan store språkmodellar berre har vore allment tilgjengelege i knappe tre år, er studiane so langt relativt enkle og utforskande. Det er for tidleg å teoretisera og generalisera, men der er eit rikt omfang av fallstudiar og utforskande studiar av ei lang rekkje ulike problemstillingar.

Rudolph mfl. (2023) omtaler artikkelen sin som ei av dei første forskingsartiklane til å tematisera ChatGPT i høgare utdanning. Forfattarane går gjennom litteraturen og testar modellen på nokre få ulike oppgåver. Dei rår høgare utdanningsinstitusjonar til å nytta verktøyet, og å forska vidare på verknadene desse

store språkmodellane har på læring og undervising. Chang og Ginter (2024) set ut for å finna kva disiplinær og typar spørsmål som passer best for automatisk retting med ChatGPT. Dei finn ikkje nokon spesielle mønster (og det er i seg sjølv eit funn) anna enn at språkmodellen gjer det betre når spørsmåla er snevrare. Forfattarane ser ikkje at modellen er god nok (på den tida) til direkte å kunna brukast på det dei kallar *Automatic Short Answer Grading*. Dei understrekar likevel at resultatet kunne ha vore eit anna om ein gav modellen betre døme på svar og vurderingskriterium eller om ein gjorde om på ordlyden i spørsmåla. Eit aspekt av dette var undersøkt av Klyshbekova og Abbott (2024) då dei ba ChatGPT både gjera oppgåva, laga vurderingskriteria og deretter evaluera sitt eige arbeide. Like fullt var resultatet heller lite tilfredsstillande. Korkje kriteria eller vurderinga heldt den akademiske standarden som forfattarene la til grunn. Ei tilleggsutfordring var at modellen gav ulike poengsummer på den same oppgåva. Variansen var ikkje stor, men blei brukt som eit argument for manglande pålitelighet i vurderingane. Dette er også Jukiewicz (2024) innom i sin studie, og dei argumenterer for å gjere minst sju vurderingar for deretter å bruke type-talet som endeleg karakter. Dei meiner då at modellens hallusinasjonar skal ha liten effekt på karakteren som gis.

Kasneci mfl. (2023) kommenterer på fleire moglegheiter og utfordringar ved bruk av store språkmodellar i utdanning. Dei oppmodar til ei semi-automatisering av vurderinga, der lærarane bruker modellane til å dra fram moglege styrker og lytar i studentarbeidet, for so å gjera vurderinga sjølv. Dei ytrar likevel uro for at lærarane kjem til å verta avhengig av og prisgjewe modellane. Dei er redd modellane ikkje kan erstatta dei menneskelege instruksjonane som utviklar studentane si kreativitet, kritiske tenking og problemløysingsevner. Dette er viktige element i formativ vurdering. Ogso dei umiddelbare tilbakemeldingane kan ein vere uroa over, då dei kan forsterka studentane si latskap, og hindra dei i å gjennomføra eigne undersøkingar som kunne vore med på å auka læringsutbytte deira. Dei tek til orde for meir opplæring og retningslinjer for bruken av store språkmodellar.

Chiang og Lee (2023) presenterer nokre retningslinjer for bruken av store språkmodellar til automatisk vurdering. Frå studien deira kjem det fram at å be språkmodellen berre gje karakter eller poengverdi ikkje er optimalt. Det er betre å be modellane rasjonalisera vurderinga si for å betre samsvaret med menneskelege vurderingar. Eit døme på korleis ein kan gjere dette finn me i arbeidslivet, der bedrifta hjå Pinto mfl. (2023) brukar Chat-GPT til å vurdere opne spørsmål i teknisk opplæring av utviklarar. Dei brukar fylgjande *prompt*

«What grade would you give to the student's response, considering the expert's answer, on a scale from 0 to 10? Return the response in a JSON format, with a variable 'grade', containing your grade, and another variable 'explanation' with the explanation for this grade. Your explanation must have at least 20 words. In the explanation, identify knowledge gaps and explain how to minimize these gaps using real-world examples.»

Latif og Zhai (2024) kjem også med nokre anbefalingar som gjerne er relevante, sjølv om dei hadde utgangspunkt i ungdoms- og vidaregåande skule.

Studentane brukar ikkje nødvendigvis det same språket som modellane er trent på. Det kan vera avgjerande for nytteverdien å finjustera og ettertrena modellane på meir kontekstuelle data. Treningsdata må gjerne vere ei mengd tidlegare spørsmål, studentsvar og lærarvurderingar med varierende kompleksitet, frå same kurs.

Ein studie av den utforskande sorten er Koutchme mfl. (2024) som brukar ein stor språkmodell til å evaluere tilbakemeldingane gitt av andre språkmodellar. Det interessante er at GPT-4 evaluerer tilbakemeldingane frå dei små opne Zephyr-7B modellane som like gode som dei frå GPT-3.5 og endå betre enn dei frå dei store CodeLlama modellane. Dette er nyttig fordi desse mindre og opne modellane vil vera enklare å finjustera med kontekstuell data. Dei kan òg vera meir kostnadseffektiv og kan køyrast på eit lukka system som sikrar personvern. Til sist nemnar forfattarane (som så mange andre) at når modellane raskt genererer tilbakemeldingar førespurnad, vert dei òg lett tilgjengelege støtteverktøy for studentane i læringsprosessen. Språkmodellane gjer ein tilfredsstillande jobb i mange fall, men dei strever når fagstoffet vert komplekst.

4 Forsøket

Utgangspunktet for studien er eit quiz-opplegg, utvikla av ein av forfattarane, til eit kursopplegg i optikk for biologiske fag. Dette er ikkje eit eige emne, men eit opplegg som inngår i eit større emne elles undervist av andre. Opplegget omfattar førelesingar og øvingar over fire veker. Målet for opplegget er å gje studentane ei djupare forståing for mikroskop og andre optiske instrument som dei bruker i biologiske fag.

Sjølv om den aktuelle quizen er for eit spesifikt fag, skal me leggja vekt på den generelle løysinga, som skal vera overførbar til andre fag. Quizen bruker Moodle, som er ein populær LMS (Learning Management System) med open kjeldekode. Moodle står for *Modular Object-Oriented Dynamic Learning*, og modulariteten er ein stor styrke. M.a. har Moodle eit rikt monn av ulike spørsmålstypar, med ulike formar for automatisk retting. Moodle vert mykje brukt i matematikk- og fysikkmiljø pga. spørsmålstypen STACK, der matematiske svar vert validert algebraisk. Der er også spørsmålstypar for fleirvals svar, numerisk validerte svar (*Calculated*) og programmering (*CodeRunner*). Quizen i det faktiske kursopplegget omfatta ulike spørsmålstypar, og hensikta med forsøket var å utvida repertoiret med nokre fritekstsvaer som òg vert automatisk retta.

Sjølv om *CodeRunner* er utvikla for programmeringsoppgåver, er spørsmålstypen generell nok til å køyra vilkårleg kode på vilkårlege svar frå studentane. Det vanlege er sjølvsagt å køyra ein kompilator eller tolk der studentsvaret vert handtert som programkode, men me kan like gjerne ta studentsvaret i naturleg språk og køyra det gjennom ein språkmodell. I quizen vår bruker me nettopp *CodeRunner* med valideringskode som brukar ChatGPT gjennom OpenAPI.

Øvingsopplegget omfatta fem spørsmål med tilbakemelding frå KI, blant totalt 21 spørsmål fordelt på to *quizzar*. Eitt spørsmål vart brukt på den fyrste

quizen, men som fyrste forsøk var det relativt rudimentært. Fire spørsmål vart brukt på den andre *quizen* og det er dei som vert drøfta her.

Språkmodellen vart fora med éi lang *prompt*, som omfattar ei kontekst med utdrag frå pensumlitteraturen, oppgåveteksta og evt. resultat frå tidlegare forsøk av same student på same oppgåve, samt instruksjonar om korleis svaret skulle formatterast. På éi oppgåve inkluderte me eit løysingsforslag, og på to oppgåver ei liste over punkt som burde vera med. På den fjerde oppgåvene stod språkmodellen heilt fritt i å velja vurderingskriterium. Lite tid vart investert i *prompt engineering*, og her ligg eit uutnytta potensiale.

5 Den tekniske løysinga

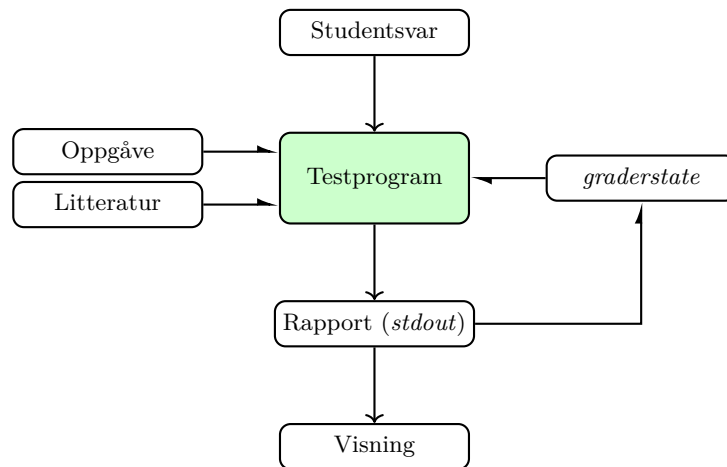


Fig. 1. Interaksjon med testprogrammet.

I den enklaste bruken av *CodeRunner* svarer studentane med programkode som vert køyrd gjennom ein tolk, og vurderinga er basert på utskrifta frå programmet. I denne samanhengen skal me derimot bruka *CodeRunner* med *custom template*, der me kan skriva inn vårt eige testprogram som behandlar student-svaret som data.

I tillegg til oppgåveteksta og svaret frå studenten, opererer Moodle med ein *graderstate* – ein vurderingstilstand – som kan innehalda tidlegare svar frå og tilbakemeldingar til studenten på det aktuelle spørsmålet. I formativ vurdering kan det t.d. vera nyttig å tilpassa tilbakemeldinga ikkje berre til det siste svaret, men òg tidlegare svar, for å fokusera på det som er nytt og unngå å gjenta vurderingar som er gjevne før.

Figur 1 viser interaksjonen rundt testprogrammet. På venstre side er data som høyrer til den spesielle oppgåva, sjølve oppgåveteksta og relevante utdrag

frå pensumlitteraturen. Til høgre er vurderingstilstanden, som testprogrammet kan lesa. På toppen er sjølve svaret som studenten leverer. Resultatet frå testprogrammet inneheld både tilbakemelding til studenten og ein ny vurderingstilstand. Av sikkerheitsomsyn køyrer *CodeRunner* testprogrammet i ei sandkasse, og utdata og tilstandsendingar vert difor skrivne til *stdout*. Moodle tolkar dette for å oppdatere vurderingstilstanden og visa tilbakemelding på skjermen.

Testprogrammet kommuniserer med ein stor språkmodell over HTTP. I utprøvinga brukte me ChatGPT v4 frå OpenAI, men det er òg mogleg å bruka t.d. Ollama. Me kjem tilbake til val av språkmodell i avsnitt 7. Meldinga til språkmodellen ser slik ut (i JSON):

```
{ "model": "gpt-4o",
  "messages": [ { "role": "system",
                  "content": chatrolle
                },
                { "role": "user",
                  "content": studentsvar
                }
  ]
}
```

Meldinga definerer to roller, ei for studenten og ei for språkmodellen. Variabelen **studentsvar** er svaret slik som studenten har gjeve det. Systemrolla, **chatrolle**, forklarar korleis KI skal vurdera svaret, og omfattar oppgåveteksta, litteraturoversikta (i JSON), evt. forrige svar frå studenten (frå *graderstate*), samt sjølve bestillinga som spesifiserer eit JSON-formatet for vurderinga.

Vurderinga er ei liste med testar, der kvar test har namn, forklaring, karakter (greidd/ikkje greidd) og tilbakemelding i fritekst. Det er opp til språkmodellen å avgjera kva og kor mange testar som trengst. Testprogrammet er i prinsippet generisk, og berre oppgåveteksta og litteraturutdraget treng tilpassing til den aktuelle oppgåva og emnet.

Testprogrammet er delt opp i tre delar. Hovuddelen må kopierast inn i ein tekstboks når ein opprettar spørsmålet i *CodeRunner*, men denne har me freista å halda kort. Sjølve testprogrammet vert lasta inn i python som ein teiknstreng, for å køyra i ei sandkasse. Det inneheld kodar som vert substituert med variablar frå Moodle før køyring. Den siste delen er eit bibliotek med klasse- og funksjonsdefinisjonar. Ved å bruka eit separat bibliotek kan me halda hovuddelen kort og unngå duplisering av kode.

Løysinga er publisert på github (Harang & Schaathun, 2025) med brukarvegleiing og eit oppgåvedøme. Ho omfattar python-modulen *ChatRunner* som kommuniserer med språkmodellen og formatterer resultatet, saman med ein testkonfigurasjon for *docker compose*, med Moodle og databasetenar, samt *jobe* som *CodeRunner* bruker til å køyra koden i ei sandkasse.

Kvart spørsmål krev to ulike filer som vedlegg, ei for oppgåveteksta og ei for utdraget av relevant pensumlitteratur. Elles er det meste av koden og *prompt*-teksta lagt inn i *ChatRunner*-modulen som vert installert på *jobe*-tenaren vha. pip. Berre ei kort standardkode må limast inn i *CodeRunner*-spørsmålet for å lesa studentsvaret og andre data frå *Moodle* og kalla funksjonane frå *ChatRunner*.

6 Vurdering av tilbakemeldingane

Etter øvinga spurte me studentane om lov til å bruka deira svar til evaluering av forsøket. Om lag halvparten sa ja, noko som gav 94 oppgåvesvar fordelt på fire spørsmål og på 24 studentar. Sidan studentane kunne svara fleire gongar (sjå figur 2) hadde me totalt rundt 250 forsøk. Førelsesaren gjekk gjennom desse svara og vurderingane og samanlikna dei til forventingane han ville ha hatt til ein læringsassistent. Me skal ta dei dårlege nyhenda fyrst og dei gode etterpå.

6.1 Mangelfulle tilbakemeldingar

Me identifiserte 15 tilbakemeldingar som ikkje var heilt tilfredsstillande. Blant dei var der eitt grovt tilfelle av feilinformasjon. Dei øvrige 14 innvendingane er innenfor det som ein normalt vil måtte akseptera frå læringsassistentar, sjølv om dei krev rettmessig og konstruktiv vegleiing. Innvendingane er oppsummert i tabell 6.1, og me ser at dei spennar over mange ulike feil. Dei mest uskuldige er språkblanding, der modellen svarer til dels på engelsk og til dels på norsk. Andre gongar vert studentsvaret mistyda eller informasjon vert oversedd.

Det grove tilfellet oppstår i tilbakemeldingane til student 6 i spørsmålet om spektrometriske kilder, som vist i tabell 2. Det fyrste forsøket er ikkje ille, men etter det andre forsøket blander ChatGPT saman ei deuteriumlampe og ei hydrogenlampe. Fyrstnemnde er ei vanlig kontinuerlig kjelde i UV-vis spektroskopi, medan sistnemnde er ei linjekjelde som ikkje er særleg mykje brukt i dei spektrografiske metodane i kurset.

I tredje forsøk tek studenten opp vranglæra frå ChatGPT og blandar òg saman dei to lampetypane. ChatGPT er i alle fall konsekvent når han godkjenner svaret i tråd med tidlegare tilbakemelding, men vranglæra er då ogso fullendt.

6.2 Gode tilbakemeldingar

Studentane leverte i stor grad ganske tilfredsstillande svar på fyrste forsøk. Figur 2 viser at 44% av spørsmåla vart besvarte med eitt forsøk, 65% med to forsøk eller færre, og 78% etter høgst tre forsøk. Dei bruker altså ikkje tilbakemeldings-systemet som ein *chatbot*, som skal gje omfattande bidrag til teksta. Stort sett legg dei altso ned mykje sjølvstendig arbeid før dei sender inn det fyrste svaret. Figur 3 viser infoboksen som studentane ser under øvinga. Mellom anna bed me dei om å

«Utvis(e) litt moderasjon i bruk da det ikke er satt inn ubegrenset midler på chatGPT kontoen, men ikke vær redd for å svare mange ganger »

Dette, i tillegg til litt treig responstid i språkmodellen, kan ha vore med på å redusera talet på forsøk per spørsmål.

Språkmodellen gjev gode tilbakemeldingar på studentsvar som er nær fullendte. Det dreier seg då gjerne om mindre manglar som studenten kan retta ved å leggja til ei setning eller to. Til dømes viser tabell 3 eit studentsvar som fekk

Stud.	Spørsmål	Forsøk	Beskriving
1	Kilder	2	Påstår at studenten forveksler to kilder, noe den ikke gjør
1	Kilder	6	Påstår igjen studenten feilkarakteriserer laserkilde, noe den ikke gjør
3	Kilder	Alle	Sliter med å gi riktig tilbakemelding på en upresis påstand om bølgelengder til kilde
3	Transducere	1	Plukker ikke opp alvorlig feilutsagn om foton i fotocelle
5	Mikroskopet	Alle	Ber om detaljer som allerede er tilstede (uklart språk fra student)
5	Aperturblender	1	Påstår difraksjon ikke er nevnt - Skulle heller bedt om utdypning
6	Kilder	1	Mener laser ikke er en linjekilde (egentlig riktig) - For streng på dette punktet
6	<i>Kilder</i>	2→	<i>Blander deuterium/hydrogenlampe og roter med kontinuerlig å linjekilde - villeder student</i>
8	Transducere	1	Gir full uttelling, men 1/3 av oppgaven er besvart
8	Transducere	3	Engelsk ord i overskrift
10	Mikroskopet	1	Engelske overskrifter
11	Mikroskopet	6	For høyt detaljnivå - utenfor oppgaven
14	Transducere	1	Ber om sammenlikning med PMT på unaturlig måte
17	Mikroskopet	1/2	Studenten svarer ikke på delspørsmål, men får full uttelling
22	Apperturblender	2→	Påstår at numerisk appertur påvirker diffraksjon - Forståelig men alvorlig feiltolkning

Tabell 1. Mer eller mindre mangelfulle tilbakemeldinger

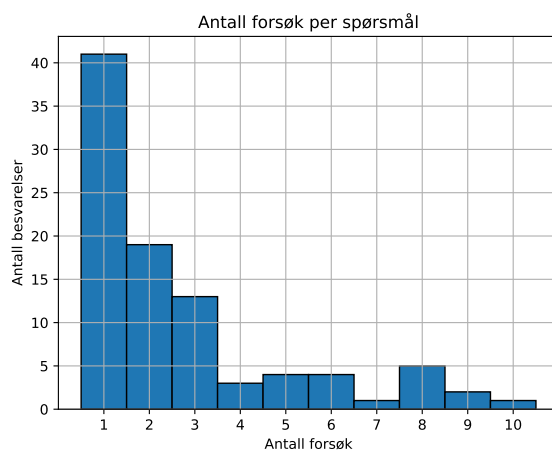


Fig. 2. Antall forsøk per spørsmål

Student: c. En hydrogenlampe er en vanlig strålingskilde i spektroskopi npr man skal studere amotære emisjonsspektre. Ved eksitering av hydrogenatomer med en elektrisk utladning hopper elektronene opp til et høyere energinivå, og emitterer så lys når de faller ned igjen. En hydrogenlampe er en linjekilde.

ChatGPT: Eksempel på en strålingskilde

Du har valgt en hydrogenlampe, som er korrekt identifisert som en linjekilde. Beskrivelsen av hvordan den fungerer er riktig vedrørende elektronovergangene. Utforsk gjerne hvordan denne prosessen kan påvirke praktiske anvendelser og resultater i spektroskopi for en mer fullstendig beskrivelse.

Student: c. En hydrogenlampe er en vanlig strålingskilde i spektroskopi npr man skal studere amotære emisjonsspektre. Ved eksitering av hydrogenatomer med en elektrisk utladning hopper elektronene opp til et høyere energinivå, og emitterer så lys når de faller ned igjen. Dette resulterer i de karakteristiske linjene i hydrogenspekteret. En hydrogenlampe er en linjekilde.

ChatGPT: Eksempel på linjekilde

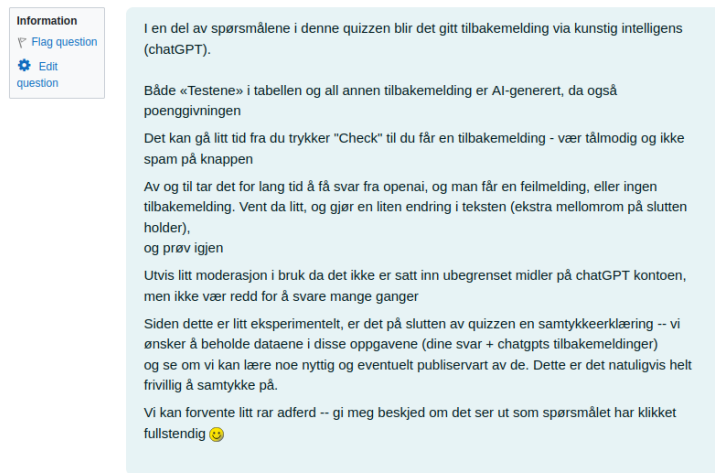
Beskrivelsen din av hydrogenlampen er nesten korrekt, men husk å klargjøre at hydrogenlamper vanligvis brukes som kontinuerlige kilder i UV-området, og gi mer detaljer om hvordan utladningen spesifikt fører til utsendelse av de spesifikke bølglengdene. En detaljert forklaring av emisjonsprosessen kan styrke ditt svar ytterligere. Husk å nevne f.eks. Balmer-serien for å knytte det mer til spektroskopi.

Student: c. En hydrogenlampe er en vanlig strålingskilde i spektroskopi npr man skal studere amotære emisjonsspektre. Hydrogenlampen består av et glassrør fylt med hydrogen under lavt trykk. Deretter påføres en høyspenningskilde til elektrodene i hver ende av glassrøret. Når elektronene i glassrøret påføres spenning eksiteres de, altså hopper elektronene opp til et høyere energinivå, og så emitterer de lys når de faller ned igjen. Dette resulterer i de karakteristiske linjene i hydrogenspekteret. Disse deles inn i Lyman-serien som ligger i UV-området, Balmer-serien som er synlig lys, og Paschen-serien som er infrarødt lys. Ved spektroskopi er det lyset fra Balmer-serien som er mest brukt.

ChatGPT: Eksempel på strålingskilde

God beskrivelse! Du har forklart hvordan en hydrogenlampe fungerer og har inkludert detaljer om seriene i hydrogenspekteret, som viser forståelse for emnet. Veldig bra! Husk å rette små skrivefeil som 'npr' til 'når'.

Tabell 2. Døme på vranglære frå ChatGPT.



Information
 Flag question
 Edit question

I en del av spørsmålene i denne quizen blir det gitt tilbakemelding via kunstig intelligens (chatGPT).

Både «Testene» i tabellen og all annen tilbakemelding er AI-generert, da også poenggivningene

Det kan gå litt tid fra du trykker "Check" til du får en tilbakemelding - vær tålmodig og ikke spam på knappen

Av og til tar det for lang tid å få svar fra openai, og man får en feilmelding, eller ingen tilbakemelding. Vent da litt, og gjør en liten endring i teksten (ekstra mellomrom på slutten holder), og prøv igjen

Utvis litt moderasjon i bruk da det ikke er satt inn ubegrenset midler på chatGPT kontoen, men ikke vær redd for å svare mange ganger

Siden dette er litt eksperimentelt, er det på slutten av quizen en samtykkeerklæring -- vi ønsker å beholde dataene i disse oppgavene (dine svar + chatgpts tilbakemeldinger) og se om vi kan lære noe nyttig og eventuelt publisert av de. Dette er det naturligvis helt frivillig å samtykke på.

Vi kan forvente litt rar adferd -- gi meg beskjed om det ser ut som spørsmålet har klikket fullstendig 😊

Fig. 3. Studentene ble vist følgende informasjon i om KI under øvingen

full poenguttelling etter to forsøk. Sveret er ganske godt etter første forsøk, men studenten nemner ikkje at transducerar gjerne er mynta på ein spesifikk del av EM-spekteret. Dette er eit viktig poeng, og KI fangar det opp i tilbakemeldinga og poengvurderinga (75% poenguttelling). Studenten endrar deretter første avsnitt og sender inn eit nytt svar som får full uttelling.

Nokon av resultatane med mange innsendingar kjem frå studentar som tilsynelatande synest dei har svart tilfredsstillande på oppgåva, og som leverer inn på nytt i håp om full uttelling. I mikroskopspørsmålet bed me studentane leggja ved ein lekk til skisse av optikken, noko som ikkje alltid er lett for KI å tolka riktig. Det fører til at somme studentar må prøva å feila ein del for å velja kva lekk dei kan leggja ved.

6.3 Evaluering

ChatGPT ser ut til å gje riktige og nyttige tilbakemeldingar i minst samme grad som me forventar frå læringsassistentane våre. Andre nyttige observasjonar dreier seg om kor mykje instruksjon og kontekst me bør gje til KI-modellen. I spørsmålet om mikroskopet fekk KI ingen instruksar rundt forventa omfang i studentsvaret, og heller ikkje om innhald eller detaljnivå. KI måtte byggja på spørsmålsteksta og pensumlitteraturen aleine. I dei andre spørsmåla la me ved kva punkt eit godt svar burde innehalda, men utan instruksjonar om omfang.

Eit interessant fenomen er at KI tilpassar vurderingskriteria til svaret frå studenten. Når svaret har mykje detaljar, stort omfang, eller eit visst (akademisk) språk, vil KI gje kriterium som krev eit omfangsrikt svar, gjerne med innleiing, hovuddel og avslutting. Studentar som leverer knappare svar på det same spørsmålet, kan få full uttelling. Då oppgjev KI heilt andre kriterium. Dette treng

Student: Oppgåva til ein transducer er å omforme lysenergi, til elektrisk energi. Transduceren registrerar lyset med tanke på til dømes bølgelengd og intensitet. den omgjer dette til elektriske signal som igjen kan bli digitalt framstilt.

Egenskapar som er viktige er at den kan registrera lys godt, og oppfatta lågintensitetslys. Den må ha høg oppløysing og kunna skilja mellom små forskjellar. Den bør helst og gje lite elektrisk støy.

Eit døme på ein transducer er ein CCD-brikke. Denne kan brukast i mikroskop. Lyset treff overflata til brikken, og frigjer elektron i frå materialet. Dei frigjordte elektrona samlast i små lommer. Det er ladninga til desse lommene som vart avlest, og omgjordt til eit elektrisk signal.

ChatGPT: Beskrivelse av transducerens oppgave

God beskrivelse av transducerens grunnleggende funksjon: å omforme lysenergi til elektrisk energi, som så kan digitalt fremstilles.

Registrering av lys med tanke på bølgelengde og intensitet

Det er fint at du nevner registrering av lys, men det er viktig å også nevne at transducere ofte behandler en spesifikk del av spekteret (f.eks. UV, synlig, IR) heller enn å registrere hele spekteret samtidig.

Egenskaper for en transducer

Du nevner viktige egenskaper som høy følsomhet og evne til å skille mellom små forskjeller, samt lav støy. Dette er viktige egenskaper for en transducer.

Eksempel på en transducer

CCD-brikken er et godt eksempel på en transducer. Din forklaring på hvordan den fungerer, med frigjøring og samling av elektroner, gir en god forståelse av prinsippet bak dens virkemåte.

Student: Oppgåva til ein transducer er å omforme lysenergi, til elektrisk energi. Transduceren registrerar lyset med tanke på til dømes bølgelengd og intensitet. den omgjer dette til elektriske signal som igjen kan bli digitalt framstilt. Transducera behandlar ofte ein spesifikk del av spekteret. Til dømes UV eller IR (...)

Tabell 3. Student 21, spørsmål om transducere

ikkje vera eit stort problem i formativ vurdering, der hensikta er å leida studenten vidare med utgangspunkt i der dei faktisk er. Gjerne er det ein god idé å forma kriterium i samråd med studenten. Som summativ vurdering er det derimot openbert urettvist, og det kan det òg vera i formativ vurdering dersom ein ikkje er tydeleg på at kriteria ikkje kan overførast til summativ sluttvurdering.

Me har ikkje grunnlag for å vurdera kor godt egna språkmodellen er til å gje formativ tilbakemelding. Det som me har demonstrert er *umiddelbar* tilbakemelding på reine kunnskapsspørsmål. Det må bli eit ope spørsmål både kva som er formativ tilbakemelding på slike oppgåver, og i kva grad og med kva slags *prompt* KI kan bidra formativt.

7 Manglar og behov for vidare utvikling

Me har vist at det er mogleg, og relativt enkelt, å implementera eit øvingsopplegg som gjev umiddelbar tilbakemelding frå KI. ChatGPT gjev i all hovudsak tilbakemeldingar som er minst like gode som det læringsassistentar kan gje, og KI klarer det på eit minutt eller mindre, der læringsassistentane treng dagar eller veker. Der er likevel fleire utfordringar som krev vidare utvikling.

Dei fleste manglane i tilbakemeldingane kan ein truleg utbetra ved å investera meir tid i *prompt engineering* og utnytta meir av funksjonaliteten i API-et. Me kan derimot ikkje rekna med å eliminera hallusinasjonar, sjølv om me kan redusera førekomsten. Datasettet viser eitt grovt døme, der KI kjem med vranglære som fører studenten på villspor. Det er ein alvorleg risiko, og dersom hallusinasjonane ikkje kan eliminerast, må ein setja studentane i stand til å handtera feilinformasjon, noko som heller ikkje er enkelt.

Eit større problem er personvern. Me må rekna med at studentane ser på sine svar som fortrulege, og me kan ikkje forsvare å gå vidare med denne løysinga der svara vert sende til ChatGPT utan ein databehandlaravtale med OpenAI som eig KI-systemet. Det har neppe hatt negative, reelle konsekvensar i dette eksperimentet, med få og upersonlege svar, men vidare utvikling må søkja løysingar som skalerer og som kan generaliserast.

Der er ingen enkle løysingar. Sjølv om universiteta tek til å kjøpa KI-tenester som tilsette kan bruka på interne data, so manglar dei ofte API-tilgang. Til dømes satsar NTNU på MicroSoft CoPilot, men utan eit API som me kunne ha brukt saman med *ChatRunner*.

Det er mogleg å køyra språkmodellar på eigen maskinvare, men då er det tale om reduserte modellar som ikkje kan handtera like lang kontekst som ChatGPT eller MicroSoft CoPilot. Dei kraftigaste opne modellane krev ogso kostbar maskinvare; dei siste llama-modellane frå Meta krev t.d. fire eller åtte GPU-ar. Me har gjort innleiande eksperiment ved å setja opp Ollama¹ på ein arbeidsstasjon som tilfeldigvis hadde eit skjermkort i 100.000-kroner-klassen (Nvidia RTX A6000) frå eit tidlegare prosjekt. Det er relativt enkelt å setja opp, men slett ikkje enkelt å få modellane til å gje svar tilsvarende det som me har sett frå

¹ <https://ollama.com/blog/ollama-is-now-available-as-an-official-docker-image>

ChatGPT. Det største problemet vårt er at dei opne modellane sjelden gjev oss det føreskrivne JSON-formatet, med fleire vurderingskriterium og tilhøyrande vurdering. Somme modellar vekslar mellom å gje brukbare og syntaktisk meiningslause svar. Me reknar med at hovudårsaka er at *prompten* er for lang for dei små språkmodellane. Det er mogleg at betre *prompt*-design eller ettertrening av modellen vil gje betre resultat. Der er ingen tvil om at opne modellar krev meir innsats enn dei kommersielle modellane, før me får tilfredsstillande resultat.

Det er for tidleg for nokon fullstendig konklusjon. Når me presenterer dette rudimentære arbeidet er det fordi me meiner at god utvikling av ny *praxis* krev at praktikarane sjølv tek ansvar for utviklinga og testar ut nye idéar i breidast mogleg omfang. Me håper at forsøket vårt kan inspirera fleire, og at den tekniske løysinga som me har presentert vert eit verkty for fleire.

Referansar

- Baddeley, A. D., & Hitch, G. (1974). Working Memory. I G. H. Bower (Red.). Academic Press. [https://doi.org/https://doi.org/10.1016/S0079-7421\(08\)60452-1](https://doi.org/https://doi.org/10.1016/S0079-7421(08)60452-1)
- Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, 3, 100081.
- Boaler, J. (2015). *Mathematical mindsets: Unleashing students' potential through creative math, inspiring messages, and innovative teaching*. Jossey-Bass.
- Chang, L.-H., & Ginter, F. (2024). Automatic short answer grading for Finnish with ChatGPT. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23173–23181.
- Chiang, C.-H., & Lee, H.-y. (2023). A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.
- Cope, B., Kalantzis, M., & Sears-Smith, D. (2020). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational philosophy and theory*, 53(12), 1229–1245.
- Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6, 100206.
- Hansen, G., & Ringdal, R. (2018). Formative assessment as a future step in maintaining the mastery-approach and performance-avoidance goal stability. *Studies in Educational Evaluation*, 56, 59–70.
- Harang, J. J., & Schaathun, H. G. (2025). moodle-chatrunner [Open kjeldekode]. <https://github.com/iirevu/moodle-chatrunner>
- Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, 52, 101522.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., mfl. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274.
- Klyshbekova, M., & Abbott, P. (2024). Chatgpt and assessment in higher education: a magic wand or a disruptor? *Electronic Journal of e-Learning*, 22(2).
- Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J., & Denny, P. (2024). Open Source Language Models Can Provide Feedback: Evaluating LLMs' Ability to Help Students Using GPT-4-As-A-Judge. I *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (s. 52–58).
- Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100210.
- Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using chatgpt. *Proceedings of the XXXVII brazilian symposium on software engineering*, 293–302.
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching*, 6(1), 342–363.