Fra kvalitetspåpekning til revisjon: KI-støttet gjennomgang av flervalgsoppgaver

Omid Mirmotahari[1], Yngvar Berg[1], Inger Skrede[2], Meh Sameen Nawaz[3], and Lex Nederbragt[2]

[1] Institutt for informatikk, Universitetet i Oslo
[2] Institutt for biovitenskap, Universitetet i Oslo
[3] Institutt for helse og trening, Høyskolen i Kristiania

Sammendrag Vi undersøker hvordan et KI-støttet system leste gjennom flervalgsoppgaver skrevet av faglærere og læringsassistenter, og hva faglærerne gjorde med påpekningene. Systemet kombinerer 22 regelkontroller med en språkmodell som leser oppgavene mot en fast rubrikk. For 554 oppgaver ga det 1889 påpekninger, hvorav 1120 gjaldt faglige og didaktiske forhold. I tolv oppgavesett sammenlignet vi oppgavene før gjennomgangen med faglærerens senere reviderte versjon. Felt med påpekninger om faglige feil ble endret i 15 av 17 tilfeller, felt med formatpåpekninger i 3 av 29. Materialet er lite, og vi har verken sammenligning med vanlig revisjon eller med ren språkmodellbruk.

Nøkkelord kvalitetssikring · flervalgsoppgaver · store språkmodeller · informatikkutdanning · høyere utdanning

1 Innledning

Språkmodeller lager syntaktisk feilfrie flervalgsoppgaver på sekunder, og studenter lager allerede egne øvingsoppgaver på denne måten. Kvaliteten varierer. Sindre mfl. [9] konkluderer med at KI kan levere kandidatoppgaver til eksamen i norsk sammenheng, men at ekspertkuratering er nødvendig. Denne artikkelen handler om den kurateringen: hva skjer når faglæreren bruker de samme modellene til å kvalitetssikre oppgavene, og hva gjør faglæreren med det modellen peker på?

Flervalgsformatet har kjente svakheter. Binær retting belønner gjetting, og konstruksjonsfeil som at fasit er det lengste alternativet gjør at studenter kan gjette seg til riktig svar uten å kunne faget [3]. Slike feil rammer særlig de sterkeste studentene [10]. Formatet vi studerer er et diagnostisk svar på dette. Hvert svaralternativ klassifiseres av oppgaveforfatteren som CORRECT (fasit), PLAUSIBLE (galt, men bygget på et rimelig, delvis riktig resonnement), MISCONCEPTION (galt som følge av en kjent feiltanke) eller INCORRECT (galt uten bestemt feiltanke, fyller ut alternativrommet), og hvert alternativ har egen tilbakemelding. Formatet brukes i konfidensbasert vurdering (CBA), der studenten også angir hvor sikker hun er [2,4]. CBA er bare bakgrunn her. Vi har ingen studentdata og evaluerer ikke CBA som vurderingsmetode, bare kvalitetssikringen av oppgavene formatet bygger på.

Klassifiseringen er et skjønn, og taksonomien har svakheter vi må ta på alvor. Grensen mellom PLAUSIBLE, MISCONCEPTION og INCORRECT er uskarp: en kjent misoppfatning kan bygge på et delvis riktig resonnement, og «uten bestemt feiltanke» kan bety at forfatteren ikke kjenner den. Fra konseptinventar-tradisjonen er det dessuten kjent at feil tankegang noen ganger gir riktig svar, og at et galt svar ikke uten videre kan knyttes til én navngitt misoppfatning. Vi behandler derfor etiketten som forfatterens vurdering og systemets uenighet om etiketter som forslag, ikke som fasit.

Forskningen på KI og oppgavekvalitet har i hovedsak målt hvor godt modeller oppdager feil mot ekspertvurderinger. Moore mfl. [5] fant at en regelbasert metode ga høyere samsvar med menneskelige vurderinger enn GPT-4 (91 mot 79 prosent), mens GPT-4 markerte flere mulige feil per oppgave (4,2 mot 1,6). Oppfølgeren SAQUET kombinerer regelkontroller og språkmodell [6], og Ramesh mfl. [7] laget et annotert evalueringssett for innføringsprogrammering. Når lærere får KI-generert tilbakemelding til studenter, beholdt de om lag 80 prosent uendret [1]. Vi har ikke funnet tilsvarende undersøkelser av hva lærere gjør med KI-kritikk av sitt eget oppgavemateriale.

Vi avgrenser artikkelen til to spørsmål. Systemet og formatet er bakgrunn og ikke gjenstand for evaluering.

RQ1: Hvilke kvalitetspåpekninger gir systemet i oppgaver skrevet av faglærere og læringsassistenter?

RQ2: Hvor ofte endres påpekte felt i faglærernes reviderte oppgavesett, og varierer det mellom typer påpekninger?

2 Systemet

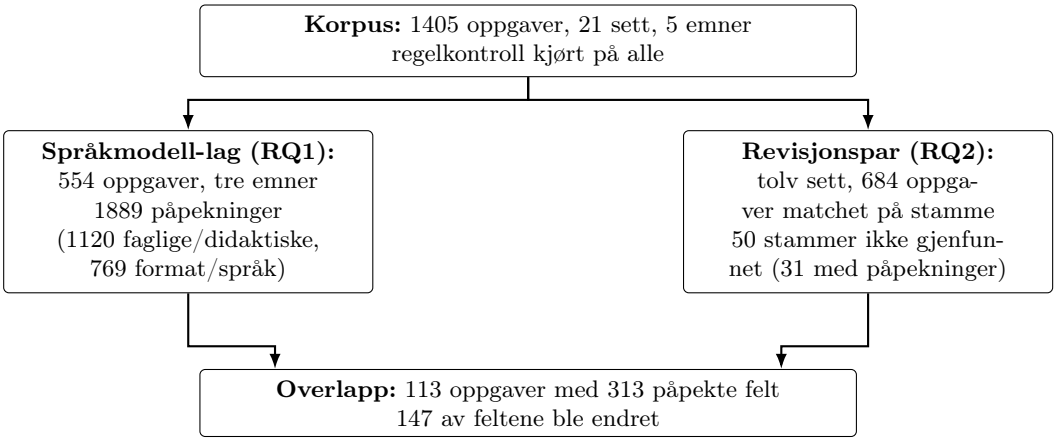
Systemet brukes som en førstelesing etter at faglæreren har levert et ferdig oppgavesett og før settet publiseres til studentene. Det har tre deler. Først 22 regelbaserte kontroller for det som lar seg regelfeste: strukturfeil, tomme felt, tegngrenser, klassiske ledetråder som at fasit er det lengste alternativet, og etiketter utenfor formatets vokabular. Deretter en språkmodell som leser hver oppgave mot en rubrikk med sju akser: faglig korrekthet, distraktorkvalitet, klassifisering av alternativene, tilbakemeldingskvalitet, språk, nivåplassering og egnethet for gruppebesvarelse. Til sist samles alt i ett dokument til faglæreren, med hver påpekning knyttet til ett felt i én oppgave og en begrunnelse i klartekst.

Hver påpekning får en alvorlighetsgrad. **RØD** betyr at forholdet bør håndteres før oppgaven publiseres, **GUL** at det er en anbefaling faglæreren kan velge å følge, og **BLÅ** at systemet allerede har rettet en skrivefeil synlig i arket. Fargen settes av språkmodellen etter rubrikkens definisjoner. Skillet er definert ut fra hvor viktig det er å handle, og vi har ikke operasjonalisert det mer presist. Fargene er ikke validert mot en uavhengig vurdering (se kapittel 5). Systemet endrer aldri faglig innhold selv. Faglæreren bestemmer. Arkitekturen lar seg kjøre med ulike språkmodeller, og vi omtaler systemet som beskrivelse av hva som ble kjørt, ikke som en sammenligning av designvalg.

3 Data og metode

Korpuset består av 1405 unike flervalgsoppgaver i 21 oppgavesett, skrevet av faglærere og læringsassistenter i fem innføringsemner ved to norske institusjoner: fire emner i biovitenskap og ett i programmering. Vi oppgir ikke institusjon eller emnekode. 1177 oppgaver har full firedelt klassifisering, og 228 foreligger bare som binært riktig/galt. Hver oppgave har også et nivå (1–3). Det er forfatterens merking av oppgavens kognitive krav på en tredelt skala, og det er én av rubrikkens sju akser om merkingen stemmer med oppgaven. Korpuset er satt sammen etter et manifest med én autoritativ fil per oppgavesett, slik settet forelå da det gikk til kvalitetssikring. Maskingenererte sett er ekskludert og duplikater fjernet.

Vi bruker fire måleenheter. Et *felt* er én celle i oppgavesettets regneark, for eksempel stamme, alternativtekst, etikett eller tilbakemelding. En *påpekning* er én



Figur 1. Inkluderings- og analyseløp. Påpekningene (RQ1) og endringene (RQ2) bygger på to utvalg som overlapper i 113 oppgaver. De 50 stammene som ikke ble gjenfunnet, er ikke med i endringstallene.

kommentar knyttet til ett felt, og et felt kan ha flere. En *endring* er at et felts innhold er ulikt mellom to versjoner. Kategorier som *Faglig feil* er analytiske koder vi har satt i etterkant ut fra påpekningsteksten, med regelbasert klassifisering og manuelle stikkprøver.

Regelkontrollene er kjørt på hele korpuset. Språkmodell-laget ble kjørt settvis i august 2026 med Anthropic-modellen Claude Opus 5, én gang per sett, med hele settet og hele rubrikken i konteksten. Det omfatter bare de 554 oppgavene som gikk gjennom kvalitetssikringen i strukturert rapportform i studieperioden, fordelt på tre av fem emner. Utvalget fulgte kvalitetssikringsløpet og er ikke trukket tilfeldig. Figur 1 viser hvordan analyseutvalgene henger sammen.

For RQ2 sammenligner vi tre versjoner av hvert sett: forfatterens innleverte versjon (T0), versjonen faglæreren fikk tilbake med språkrettelser utført og øvrige påpekninger merket (T1), og faglærerens reviderte versjon (T2). Endringssporingen sammenligner T0 med T2, og BLÅ språkrettelser rapporteres separat fordi de var utført i T1. Oppgaver matches på uendret spørsmålsstamme. Det gir 684 matchede oppgaver og 11 756 feltsammenligninger, og endringsraten beregnes for påpekte felt per kategori og alvorlighetsgrad. Feltene er gruppert i oppgaver, sett og et lite antall forfattere og er ikke uavhengige observasjoner. Vi rapporterer derfor tellere og andeler og trekker ikke statistiske slutninger.

Validiteten av påpekningene ble kontrollert manuelt av førsteforfatter alene, uten blinding og uten andrebedømmer. Vi vurderte 93 påpekninger trukket med faste kvoter per kategori for om de var berettigede, alle 147 endrede påpekte felt for om endringen adresserer påpekningen, og de 13 oppgavene som ikke fikk noen påpekning for oversette problemer. Studien bruker ingen studentdata. Materialet er kollegers og læringsassistenters oppgaver og revisjoner, samlet inn som del av ordinært kvalitetssikringsarbeid, og emneansvarlige er med i forfattergruppen og har sett funnene for eget materiale.

4 Resultater

4.1 Påpekningene (RQ1)

Av de 1889 påpekningene gjaldt 769 (41 prosent) format eller språk: for få alternativer, overskredne tegngrenser og språkrettelser levert ferdig utført. Mange av disse

Tabell 1. Faglige og didaktiske påpekninger fra språkmodell-laget etter kategori, 554 oppgaver. RØD + GUL summerer ikke alltid til antallet, fordi 70 BLÅ språkrettelser og to ugraderte markeringer hører til kategoriene.

Kategori	Antall	RØD	GUL	Typisk innhold
Klassifisering	322	21	300	gal eller manglende etikett på alternativ
Tilbakemelding	309	87	170	bekreftende, tom, eller knyttet til feil alternativ
Nivå	133	16	117	nivåmerking i utakt med oppgavens krav
Distraktorkvalitet	114	32	78	ikke-plausible eller identiske distraktorer
Faglig feil	84	35	44	gal tilbakemelding, begrepsfeil, dobbel fasit
Ledetråd	56	4	43	fasit kan identifiseres uten fagkunnskap
Gruppeegnethet	34	2	32	ren gjenkjenning
Forklaring/annet	68	13	54	

kan kontrolleres automatisk, men vi har ikke målt tidsbruken. Resten av analysen gjelder de 1120 faglige og didaktiske påpekningene (tabell 1).

Påpekninger om faglige feil forekom i alle tre gjennomgåtte emner: 84 påpekninger på 77 oppgaver gjelder faglig gale tilbakemeldinger, upresis begrepsbruk eller flere riktige alternativer. Dette er en feiltype som kan villede studenten, siden en faglig gal tilbakemelding forklarer feil, og den kan ikke fanges av regler. Eksempelet er valgt som tydelig tilfelle, ikke som tilfeldig utvalg:

> **Faglig feil (RØD).** "*'It never copies non-primitive objects: it only creates a new name pointing at the same object.' Dette er faglig galt og lærer bort en misforståelse: Python har ikke primitive typer, og tilordning kopierer aldri noe som helst, heller ikke tall og strenger.*"

Klassifisering er den største kategorien, og her er systemets påpekninger minst entydige. 322 påpekninger gjelder etiketter som ikke stemmer med alternativets karakter, og 300 av dem er GUL. Det typiske mønsteret er at skillet mellom INCORRECT og MISCONCEPTION settes feil vei, for eksempel:

> **Klassifisering (GUL).** "*'The boundary values like exactly 70 fall through because* `>=` *excludes them' er merket PLAUSIBLE, men dette er en spesifikk, gjenkjennelig misforståelse. Forslag: endre til MISCONCEPTION.*"

Her kan en rimelig faglærer mene at begge etiketter er forsvarlige, og vi har ingen uavhengig fasit for slike tilfeller. Tilbakemelding er nest største kategori (309), og den gjelder tilbakemeldinger som bekrefter i stedet for å forklare [8], gjentar alternativteksten, står tomme eller hører til et annet alternativ. Kontrollaget viser et samsvarende mønster over alle 1405 oppgavene: fasit er det klart lengste alternativet i 464 av dem, og absolutte ord som «alltid» og «aldri» står bare i de gale alternativene i 163.

Alle de 93 kontrollerte påpekningene ble vurdert som berettigede, og ingen av de 13 upåpekte oppgavene viste kvalitetsproblemer. Kontrollen er gjort av én bedømmer som er førsteforfatter og systemansvarlig. 93 av 93 dokumenterer ikke presisjonen i alle 1889 påpekninger, 13 oppgaver gir ikke et estimat på hva systemet overser, og kontrollen sier ikke om RØD og GUL er riktig fordelt.

Tabell 2. Påpekte felt som ble endret i faglærerens reviderte versjon, etter kategori. Alle 313 påpekte felt i revisjonsparene er med.

Kategori	Endret	Andel	Kategori	Endret	Andel
Faglig feil	15/17	88 %	Ledetråd	10/17	59 %
Tilbakemelding	29/33	88 %	Distraktorkvalitet	8/32	25 %
Klassifisering	43/54	80 %	Språk (rest)	17/96	18 %
Nivå	18/27	67 %	Format	3/29	10 %
Gruppeegnethet/annet	4/8	50 %			

4.2 Endringer i reviderte oppgavesett (RQ2)

I settene med revisjonspar hadde 313 felt påpekninger fra språkmodell-laget, og 147 av dem ble endret. Fordelt på alvorlighetsgrad ble 101 av 153 GUL-felt endret (66 prosent) og 33 av 62 RØD-felt (53 prosent). At RØD ligger under GUL henger sammen med at formatkrav er den største gruppen blant RØD-feltene (28 av 62), og formatpåpekninger ble sjelden fulgt av endring. BLÅ språkrettelser var utført i T1, og de 13 av 98 som likevel er endret mellom T0 og T2, kan ikke leses som lærerrespons.

Etter kategori er mønsteret tydelig selv om tellerne er små (tabell 2). Faglige feil og mangelfulle tilbakemeldinger ble som regel fulgt av endring (15 av 17 og 29 av 33), mens distraktorforslag (8 av 32) og formatkrav (3 av 29) sjelden ble det. Regnet per felt ble 139 av 571 felt flagget av regelkontrollene endret, mot 914 av 11 185 uflaggede. Forskjellen er størst for forklarings- og tilbakemeldingsfelt og snur for etikettfelt og alternativtekster, som ble endret sjeldnere når de var flagget. Alle 147 endrede, påpekte felt ble lest side om side med påpekningen. Alle ble vurdert å adressere påpekningen, ingen å være endret av annen grunn eller reversert. Dette viser samsvar mellom endring og råd, men ikke at de reviderte oppgavene ble faglig bedre.

Matchingen på spørsmålsstamme har en kjent svakhet. 50 stammer ble ikke gjenfunnet i revidert versjon, og 31 av dem hadde påpekninger. De kan være slettet eller omskrevet og er utelatt fra endringstallene. Matchingen utelukker dermed trolig noen av de største revisjonene. Retningen på skjevheten kjenner vi ikke, men tallene kan undervurdere hvor mye som ble gjort med påpekningene.

5 Diskusjon og begrensninger

Oppfølgingen varierte mellom kategoriene: faglige feil og mangler ved tilbakemeldinger ble oftere fulgt av endring enn formatkrav og distraktorforslag. Vi har ingen data om hvorfor. Forklaringene spenner fra faglig uenighet til tidsbruk og hvor mye omskriving en endring ville kreve. Vi kan heller ikke si om lærerne fulgte gode råd og avviste dårlige, siden rådenes kvalitet ikke er uavhengig fastslått, og dermed heller ikke noe om hvor godt kalibrert tilliten til systemet var. Borchers mfl. [1] fant at lærere beholdt det meste av KI-generert tilbakemelding, men det gjaldt KI-generert innhold og ikke kritikk av lærerens eget arbeid.

Fargekodene er en reell svakhet. Fargen settes av språkmodellen, og vi har ikke undersøkt om den er riktig eller konsistent. Siden RØD og GUL ikke er validert, er forskjellen på 53 og 66 prosent oppfølging vanskelig å tolke. Det samme gjelder klassifiseringspåpekningene. Taksonomiens uskarpe grenser gjør at «riktig» etikett mellom PLAUSIBLE, MISCONCEPTION og INCORRECT i mange tilfeller ikke er veldefinert, og at systemets uenighet er et forslag til faglæreren, ikke en feil i

oppgaven. Det er rimelig å lese at 80 prosent av klassifiseringsfeltene ble endret som et uttrykk for at faglærerne fant forslagene brukbare, ikke som at etikettene var gale. En enklere taksonomi, eller en der misoppfatninger knyttes til dokumentert litteratur, kan være nødvendig.

Studien har to mangler som begrenser hva den kan si om systemet. Vi har ingen sammenligning med en vanlig revisjon. Vi vet ikke hvor mye faglærerne ville ha endret uten systemet, hvor lang tid en vanlig gjennomgang tar, eller om endringene er av samme art. Vi har heller ingen sammenligning med ren språkmodellbruk, der råmaterialet leveres til en nyere modell sammen med en passende ledetekst. Rubrikk, regelkontroller og feltnivåleveranse er derfor ikke isolert som bidrag, og det er mulig at en enklere oppskrift gir omtrent det samme. Tidsbesparelsen er ikke målt.

Flere begrensninger kommer i tillegg. Valideringen er gjort av én bedømmer uten blinding, og de feilfrie resultatene (93/93, 147/147, 0/13) må leses i lys av det. Språkmodell-laget dekker tre av fem emner uten tilfeldig utvalg. Kjøringene brukte én modellkonfigurasjon i én periode, og eksakt modellversjon og dekodingsparametre ble ikke registrert. Rubrikk og kildekode er beskrevet, men ikke offentlig tilgjengelige. Endringsanalysens tellere er små, og kategoriene er satt i etterkant. Til sist har plattformen ikke logget studentbesvarelser, så ingen funn kan kobles til studentatferd eller læring. Korpuset ble også brukt til en eksplorativ sammenligning av oppgaver skrevet for formatet og oppgaver konvertert fra binært format. Den er ikke tatt med her, fordi produksjonsvei, emne og klassifiseringsprosedyre ikke lar seg skille, og den behandles separat.

6 Konklusjon

Vi har koblet rubrikkstyrte kvalitetspåpekninger til faglæreres faktiske revisjoner av flervalgsoppgaver. Systemet påpekte forhold et rent regelbasert system ikke ser, som faglige feil, tilbakemeldinger som bekrefter i stedet for å forklare, og etiketter som kan diskuteres. Felt med påpekte faglige feil ble endret i 15 av 17 tilfeller, mot 3 av 29 for formatkrav, og hvorfor noen påpekninger følges og andre ikke, er ikke undersøkt. Neste steg er en sammenligning med vanlig revisjon og med en enkel ledetekst til en språkmodell, en uavhengig dobbeltbedømming av påpekninger og farger, og en tydeligere taksonomi for distraktorer.

Referanser

1. Borchers, C. et al. (2026). Understanding teacher revisions of large language model-generated feedback. In *Artificial Intelligence in Education. AIED 2026*, pp. 60–74. Springer.
2. Gardner-Medwin, A. R. (2006). Confidence-based marking: Towards deeper learning and better exams. In Bryan, C. and Clegg, K. (eds.), *Innovative Assessment in Higher Education*, pp. 141–149. Routledge.
3. Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333.
4. Mirmotahari, O., Perminova, I., and Berg, Y. (2025). Fra gjetting til refleksjon på eksamen med flervalgsoppgaver. *Norsk IKT-konferanse for forskning og utdanning*, 37(4).
5. Moore, S., Nguyen, H. A., Chen, T., and Stamper, J. (2023). Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In *EC-TEL 2023*, pp. 229–245. Springer.
6. Moore, S. et al. (2024). An automatic question usability evaluation toolkit. In *Artificial Intelligence in Education. AIED 2024*, LNCS 14830, pp. 31–46. Springer.

7. Ramesh, A. et al. (2024). A benchmark for testing the capabilities of LLMs in assessing the quality of multiple-choice questions in introductory programming education. In *SIGCSE Virtual 2024*, pp. 193–199. ACM.
8. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.
9. Sindre, G., Kolås, L., Munkvold, R. I., Nowostawski, M., Ringen, H. E. S., and Vassbakk, J. P. (2025). Generative AI tools to propose summative assessment tasks. *Norsk IKT-konferanse for forskning og utdanning*, 37(4).
10. Tarrant, M. and Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2), 198–206.