

Kva seier generativ KI om kvalitet i IT-utdanning?

Guttorm Sindre² [0000-0001-5739-8265], Line Kolås¹ [0000-0003-4908-3241],
Robin Isfold Munkvold¹ [0000-0002-2524-279X]

¹Fakultet for samfunnsvitskap, Nord Universitet, Levanger, Noreg
line.kolas@nord.no, robin.munkvold@nord.no

²Institutt for datateknologi og informatikk, NTNU, Trondheim, Noreg
guttorm.sindre@ntnu.no

Samandrag. Dei siste åra har generativ KI basert på store språkmodellar fått stor popularitet til ulike formål, både innan akademien og elles i samfunnet. Éin mogleg bruk kan vere å få kjapt innblikk i eit nytt tema gjennom eit konsist samandrag. I denne artikkelen prøver vi å finne ut kva generativ KI seier om kvalitet i høgare IT-utdanning. Forskingsspørsmålet er ikkje primært å finne ut kva utdanningskvalitet *er* (som fagfolk allereie burde ha eigne idear om), men vel så mykje å finne ut noko om desse systema og – indirekte – kva dei har lært om utdanningskvalitet frå treningsmaterialet sitt. To ulike prompt blei brukte, der det første spør om kvalitet i høgare IT-utdanning på ein open måte, medan det andre spesifikt vinklar spørsmålet mot samstemd undervisning (*constructive alignment*). Tre ulike KI-system vart brukte første måndag i kvar måned gjennom eit halvt år, for også å kunne sjå eventuelle skilnader mellom systema og utvikling over tid. Dei primære funna er at det var ganske stor skilnad på svar mellom dei to ulike prompta, men mindre skilnad i svar på same prompt mellom ulike KI-system, og ingen klar tendens til utvikling i svara over den tidsperioden studien føregjekk.

Nøkkelord: IT-utdanning, kvalitet, samstemd undervisning, KI, språkmodellar

1 Introduksjon

Dei siste åra har generativ KI basert på store språkmodellar fått stor popularitet til ulike formål, både innan akademien og elles i samfunnet. Ein mogleg bruk kan vere å kjapt setje seg inn i eit nytt tema ved å få eit konsist samandrag. Tidlegare ville ein typisk ha brukt Wikipedia eller søk på internett til dette, men no vil mange i staden gi eit prompt til KI-system som t.d. ChatGPT. Kvaliteten på det som kjem ut, kan variere, men det kan likevel vere interessant å sjå, fordi dette seier noko om kor gode (eller dårlege) systema er, og kva informasjon som finst i treningsmaterialet dei har vore eksponert for.

Eitt omgrep som det kan vere interessant å sjå KI'ane sine påstandar om, kan vere *utdanningskvalitet*. Mange norske universitet og høgskular har nettsider om dette, men felles for dei fleste er at dei vegrar seg for å definere kva utdanningskvalitet er. Til

dømes seier NTNU si nettside¹ at definisjonen av kvalitet kan variere mellom fagfelt og må framkome gjennom lokale dialogar, og Nord si nettside² seier meir om kvalitets-systemet enn om kva god kvalitet er i seg sjølv. Dette er for så vidt også i tråd med internasjonal faglitteratur. Til dømes inneheld antologien til Green (1994) mange ulike perspektiv på utdanningskvalitet, og litteraturstudien av Welzant et al. (2015) konkluderer også med at det er eit uklårt omgrep. Som faglærarar innan IT-fag og forskarar innan IT fagdidaktikk bør vi rimelegvis allereie ha våre egne idear om kva som er god kvalitet i IT-utdanning. Det kan likevel vere interessant å sjå kva generativ KI seier om dette, som kan gi ein slags syntese av kva som er skrive om utdanningskvalitet relevant for høgare IT-utdanning i den litteraturen som desse systema har blitt trena opp på. Ved å prøve ut dei same prompta på fleire system og over tid kan vi dessutan sjå om det er skilnader mellom systema, og om svara viser noka utvikling over tid. Forskingsspørsmåla våre er:

FS1: Kva perspektiv har KI-systema på kvalitet i høgare IT-utdanning, versus indikatorar i Studiebarometeret?

FS2: I kva grad har svara innhald som er spesifikt for IT-utdanningar, versus innhald som kunne gjelde kvaliteten i all slags høgare utdanning?

FS3: I kva grad fører to ulike prompt – eitt meir ope, eitt meir spesifisert – til skilnader i svara som blir gitt?

FS4: Kva skilnader er det mellom svara frå ulike KI-system?

FS5: I kva grad viser svara utvikling over tid?

Det finst mykje anna enn Studiebarometeret som kunne ha vore brukt som rammeverk for ein analyse av KI-tekstar om kvalitet i utdanning. Desse indikatorane - Undervisning, Tilbakemeldingar, Vurderingsformer, Læringsmiljø, Organisering, Tilknytning til yrkeslivet, Studenten sitt eige engasjement – gir likevel ein peikepinn på kva norske styresmakter via NOKUT meiner er viktige aspekt ved utdanningskvalitet, og gir dermed ein mogeleg måte å strukturere analysen på. Resten av artikkelen er strukturert som følgjer: Seksjon 2 forklarar forskingsmetoden, og seksjon 3 presenterer resultat. Seksjon 4 diskuterer funna med tanke på kva slags konklusjonar som kan trekjast og i kva grad funna minner om eller skil seg frå funn i relatert arbeid. Til slutt kjem seksjon 5 med ein konklusjon, inkludert veikskapar med studien og korleis dette eventuelt kan følgjast opp i vidare forskning.

¹ <https://i.ntnu.no/wiki/-/wiki/Norsk/Hva+er+utdanningskvalitet>

² <https://www.nord.no/om/kvalitetssystem/utdanningskvalitet>

2 Relatert forskning

2.1 Kvalitet i høgare utdanning

Det er fleire som har undersøkt definisjonen av utdanningskvalitet. Slette og Johansen (2025) analyserte dokument frå det norske initiativet Senter for framifrå utdanning (SFU) for å identifisere korleis kvalitet og eksellens vert forstått. Dei finn fire grupper av kvalitetsforståingar: Fagspesifikke (knytt til disiplinenes normer); Utdanningscentrerte (fokus på undervisning og læring); Sektorretta (nasjonale mål og politikk) og Sektorovergripande (samfunnsperspektiv). Artikkelen problematiserer at omgrepa ofte er vage og kontekstavhengige, og etterlyser meir presise definisjonar og metodar for å måle kvalitet. Dei peikar også på at studentperspektivet har fått aukande innverknad på kva som vert rekna som kvalitet. Biggs (1996) hadde fokus på kvalitet innan læring, med sitt omgrep «constructive alignment» (samstemt undervisning), der det skal vere ein samanheng mellom læringsutbytte, undervisningsmetodar og vurderingsmetoder.

Det er publisert fleire artiklar om forskning relatert til Studiebarometeret. Karlsen et al. (2017) analyserer spenningsforholdet mellom vitskapelege og politiske omsyn i utforminga og bruken av Studiebarometeret. Wiers-Jensen og Hovdhaugen (2019) undersøkte kva barometeret fortel om studieinnsats på bachelornivå, og diskuterer metodiske utfordringar med studentar sin eigenrapporterte studieinnsats. Munthe (2020) analyserte korleis studentengasjement er forstått i spørjeundersøkinga og korleis data om dette kan nyttast i kvalitetsarbeid. Sindre (2021) analyserte om vurderingsformer og obligatoriske øvingar hadde effekt på kor nøgde studentar var med studieprogramma dei gjekk på. Hansen og Sindre (2023) såg på resultatata for indikatoren «Tilknytning til yrkeslivet» for ein del IT-studieprogram ved NTNU, og i kva grad dette samsvarer med programma sin arbeidslivsrelevans.

Den noverande artikkelen har lite til sams med dei som er nemnde i førre avsnitt, sidan vår artikkel her ikkje analyserer data frå Studiebarometeret på nokon måte – vi berre bruker indikatorer frå Studiebarometeret i deler av analysen av dei KI-genererte tekstane.

2.2 Analyse av tekstar laga av generativ KI

Bandi og Zeng (2024) foreslår eit rammeverk for å evaluere effektiviteten av KI-prompt og svara som blir generert. Rammeverket har fleire ulike dimensjonar, mellom anna kor rask responsen er, og kva kvalitet svara har, til dømes målt ved kor korrekt informasjon som blir gitt, og kor nøgde brukarane er med svara.

Vår artikkel skil seg frå dette ved at vi ikkje prøver å evaluere kvaliteten til svara frå KI-systema. Vi har ingen spesifikk fasit for kva som burde ha vore sagt om kvalitet i høgare IT-utdanning, og har heller ikkje prøvd å måle kvalitet gjennom kor nøgde mot-takarane er med svara. Dette kunne ha vore ein alternativ framgongsmåte – til dømes samla saman ei gruppe personar med erfaring frå kvalitetsarbeid i utdanning og bede dei gje poeng til svara (til dømes utan å vite kva system dei var laga av) – for å sjå om det vart nokon skilnad i score mellom systema, eller over tid. Det vi ønskte å finne ut,

var meir kva tema svara var innom, og i kva grad det vart skilnad mellom dei to prompta, og mellom ulike system, i kva tematikk som dukka opp.

3 Forskingsmetode

Studien tek utgangspunkt i eit interpretivistisk perspektiv, der tekstar vert forstått som meiningsberande uttrykk snarare enn objektive data. Tolkinga av tekstane er såleis ikkje nøytral, men situert i den konteksten dei er produserte i, og krev at forskaren aktivt reflekterer over kva slags forståingar og perspektiv som kjem til syne. Eit slikt tolkingsteoretisk utgangspunkt eignar seg særleg godt når målet er å undersøkje korleis KI-system formidlar ulike førestillingar om kvalitet i høgare utdanning.

Som metodisk ramme har vi valt innhaldsanalyse (Krippendorff, 2018; Cohen et al, 2018). Innhaldsanalyse gjev høve til å systematisere og samanlikne tekstmateriale på ein strukturert måte, samstundes som det opnar for tolking av meiningsinnhald. Metoden eignar seg òg for kombinasjonar av kvantitative og kvalitative tilnærmingar: den kvantitative delen gjer det mogleg å kartlegge frekvensar og semantiske mønster, medan den kvalitative delen gjer det mogleg å identifisere tema og perspektiv som ikkje kan reduserast til talfesting.

I det følgjande presenterer vi først korleis datamaterialet vart samla inn, før vi skildrar dei konkrete analysestega som vart gjennomførte.

3.1 Datainnsamling.

Tre ulike KI-system vart valde ut til undersøkinga: ChatGPT (v4.0), Google Gemini og MS Bing Copilot, slik desse var i perioden juli 2024 – mars 2025. Vi kunne ha valt andre eller fleire system (t.d., Claude, DeepSeek), men for å gjere studien overkomeleg med tanke på tidsbruk men samstundes ha ei viss breidde, verka 3 som eit passande tal. Dei tre nemnde går også for å vere dei mest brukte på verdsbasis³.

To liknande prompt vart utvikla for undersøkinga. Jamvel om denne artikkelen er på norsk, valde vi å ha prompta på engelsk, sidan mykje meir av treningsmaterialet til KI-systema vil ha vore på engelsk enn på norsk, og utdanningskvalitet er eit tema av internasjonal interesse. Dei to prompta var som følgjer:

P1: *Write a 100-word article about how to define quality in higher computing education?*

P2: *Write a 100-word article about how to define quality in higher computing education with regards to constructive alignment?*

Som ein kan sjå, er prompta like bortsett frå at P2 gir ei føring om at svaret skal fokusere på samstemd undervisning (“constructive alignment”) medan P1 er meir ope. Føremålet med å ha to ulike prompt var å sjå kva skilnader dette førte til i svaret. Med

³ <https://firstpagesage.com/reports/top-generative-ai-chatbots/>

berre 100 ord unngår ein å bli overlessa med tekst, men dette har også ei potensiell ulempe ved å vere nokså knapt for å seie noko substansielt om kvalitet i utdanning.

Prompta P1 og P2 vart utført første måndag i kvar måned frå juli 2024 til januar 2025 av to ulike forskarar uavhengig av kvarandre. Dette skulle gjere det mogeleg å sjå på eventuell utvikling i svara over tid, pluss også eventuelle skilnader mellom svar gitt til ulike personar på same dag. Mottatte svar vart lagra i eit Teams filarkiv for påfølgjande analyse.

3.2 Dataanalyse.

Deler av analysen vart gjort automatisk, mellom anna måling av semantisk distanse mellom tekstane ved ein lettvekts *sentence-BERT* metode (Reimers & Gurevych, 2019). I nokon grad brukte vi også automatisk ordteljing for å sjå frekvens av ulike ord i tekstmaterialet. Statistiske t-testar vart gjort mellom talseriar for semantisk likskap for svar på P1 vs. P2, og for samanlikningar av svar på same prompt vs. ulike prompt, for å undersøkje om skilnader var signifikante. I tillegg vart tekstane analysert manuelt for å sjå kva perspektiv dei har på kvalitet, dvs. i kva grad kunne ulike setningar i desse tekstane relaterast til dei ni samansette indikatorane i Studiebarometeret (Undervisning, Tilbakemeldingar, Vurderingsformer, Læringsutbyte, Sosialt Læringsmiljø, Fysisk Læringsmiljø, Organisering, Tilknytning til yrkeslivet, Studentengasjement). I tillegg såg vi i analysen at tekstane hadde ein del andre tema som ikkje passa inn i Studiebarometeret sine indikatorer. Dei mest markante av desse var Faglærarkompetanse, Student-suksess (at studentane gjennomfører studiane med gode resultat, får jobb etterpå, etc., typisk kalla «student outcomes» på engelsk), Kvalitetssikring (at universitetet har gode prosedyrer for å evaluere og oppdatere studietilbodet), og Forsking – at studentar vart involverte i forskingsaktivitetar i løpet av studiane. I tillegg registrerte vi om teksten nemnde spesifikke kompetansar som studentane skulle ha, anten det galdt Generiske evner (t.d. problemløysing, kritisk tenking) eller IT-spesifikk fagkompetanse (t.d., programmering, kunstig intelligens).

4 Resultat

Som innleiing på presentasjonen av resultat gir vi nokre konkrete døme på svar frå KI-systema. Det følgjande er eit svar frå ChatGPT på P1, juli 2024.

Sitat 1:

Defining quality in higher computing education involves several key factors. Firstly, a robust curriculum that evolves with technological advancements is essential. This ensures students gain up-to-date knowledge and skills. Secondly, experienced faculty with industry and research expertise enhance learning through practical insights and

innovative teaching methods. Thirdly, access to cutting-edge resources, such as advanced laboratories and software tools, enables hands-on experience. Additionally, strong industry partnerships facilitate internships and job placements, bridging the gap between academia and real-world applications. Finally, fostering a collaborative and inclusive learning environment promotes critical thinking, problem-solving, and creativity, preparing students for successful careers in the dynamic field of computing.

I manuell analyse vart denne teksten klassifisert til å vere innom tema Læringsutbytte med passasjen «robust curriculum ... up-to-date knowledge and skills», Faglærarkompetanse («experienced faculty ... expertise»), Undervisning («enhance learning ... innovative teaching methods»), Fysisk læringsmiljø («access to cutting-edge resources... hands-on experience»), Tilknytning til yrkeslivet («strong industry partnerships ... real-world applications»), Sosialt læringsmiljø («collaborative and inclusive learning environment») og til slutt nokre spesifikke evner studentane skal ha: «critical thinking, problem-solving, and creativity»). Det vart ikkje registrert nokon IT-spesifikke tema. Rett nok står ordet «computing» heilt i starten og igjen mot slutten, men dette ordet var gitt i promptet. Likeeins kunne «evolves with technological advances» ha vore uttrykt for mange andre fagfelt, ikkje berre IT, så det blir sagt lite spesifikt om kva studentane bør kunne.

Det neste er eit svar frå Bing Copilot på P2, frå august 2024:

Sitat 2:

Quality in higher computing education, with regards to constructive alignment, hinges on the coherence between learning objectives, teaching methods, and assessment tasks. Firstly, clear learning outcomes should be defined, outlining the skills and knowledge students are expected to acquire. Secondly, teaching activities must be designed to actively engage students in achieving these outcomes, using methods like problem-based learning and collaborative projects. Lastly, assessment tasks should directly measure the attainment of the specified outcomes, ensuring that students are evaluated on their ability to apply what they've learned. This alignment fosters a more effective and meaningful educational experience.

Setningane som startar “Firstly, ... Secondly, ... Lastly...” passar fint inn i Studiebarometeret sine indikatorer for læringsutbytte, undervisning og vurderingsformer. På same måte som det første dømet, er det derimot ikkje noko IT-spesifikt i teksten. Bortsett frå det eine ordet «computing» i første setning, kunne dette ha vore eit svar om kvalitet i høgare utdanning generelt.

Tabell 1: Lengde på svara frå KI-systema

System	P1, snitt	Min, maks	P2, snitt	Min, maks
ChatGPT	98	94-103	99	96-102
Bing Copilot	94	84-100	100	89-161
Gemini	106	75-146	106	79-149

7 Kva seier generativ KI om kvalitet i IT-utdanning?

Tabell 1 viser snitt og spenn i responslengde. Dei fleste svara ligg rundt den etter-spurde lengda på 100 ord, men det er litt skilnad. ChatGPT ser ut til å ha prøvd å halde seg nær 100 ord både i svara på P1 og P2 – aldri langt under, aldri langt over. Bing hadde éin tekst som var altfor lang (eit svar på P2 i juli 2024 som var 161 ord), men bortsett frå dette eine høvet, har Bing levert svar \leq 100 ord. Gemini har generelt større variasjon både med vesentleg kortare og lengre svar.

Tabell 2 viser resultat av automatisk analyse med parvise samanlikningar av semantisk likskap mellom ulike KI-svar. Ein verdi på 0.0 ville tyde at det ikkje er nokon semantisk overlapp mellom to tekstar i det heile tatt (t.d., handlar om to heilt ulike tema), medan 1.0 ville tyde at tekstane uttrykkjer heilt same meining (om ikkje naudsynt med identiske tekstar). Snittverdiar på 0.8 er dermed ganske høgt og indikerer at dei ulike responsane på same spørsmål i stor grad uttrykkjer dei same ideane. Som ein kan sjå, er gjennomsnittleg likskap noko større for P2 enn for P1, og denne skilnaden er statistisk signifikant ($p < 0.0001$), dvs. inkludering av «constructive alignment» i promptet gjorde svara likare.

Tabell 2: Gjennomsnittleg semantisk likskap i parvise samanlikningar

Prompt	P1, snitt	Min, maks	P2, snitt	Min, maks
Alle parvise samanlikningar	0.84	0.60-1.00	0.89	0.45-1.00
ChatGPT vs ChatGPT	0.85	0.73-0.92	0.91	0.85-0.95
Bing Copilot vs Bing Copilot	0.89	0.77-1.00	0.93	0.82-1.00
Gemini vs Gemini	0.84	0.63-0.98	0.84	0.44-0.96
ChatGPT vs Bing Copilot	0.85	0.73-0.93	0.91	0.80-0.96
ChatGPT vs Gemini	0.83	0.60-0.91	0.88	0.63-0.97
Bing Copilot vs Gemini	0.84	0.63-1.00	0.88	0.63-0.96

Dei tre radene som er markerte med lyst grått i Tabell 1, viser snitt av parvise samanlikningar mellom svar frå same system, medan dei nedste radene viser snitt mellom svar frå ulike system. Bing Copilot skil seg ut ved å ha dei mest internt like svara både for P1 og P2, med andre ord noko mindre variasjon i kva aspekt svara inkluderer.

Tabell 3 viser resultat av samanlikning mellom svar på ulike prompt, dvs. P1 vs. P2. Som ein kan sjå, ligg snittet her monaleg lågare for likskapen mellom svara, og dette er sterkt statistisk signifikant ($p < 0.000001$), dvs. ulike prompt (P1 vs P2) førte til større ulikskap mellom svara enn bruk av ulike verktøy på same prompt.

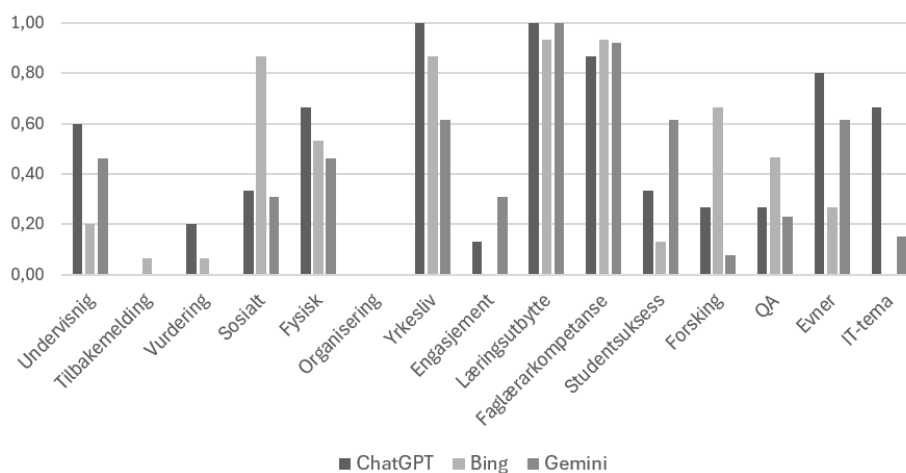
Tabell 3: Gjennomsnittleg semantisk likskap, P1 vs P2

Prompt	Snitt	Min, maks
ChatGPT	0.70	0.55-0.83
Bing Copilot	0.69	0.61-0.84
Gemini	0.68	0.45-0.82

Figur 1 indikerer kva tema tekstane var innom. Nesten alle tekstane understreka at pensum og læringsutbytte måtte vere relevante, og dei fleste la også vekt på tilknytning til yrkeslivet – om enn litt varierende mellom verktøy (alle svara til ChatGPT, 84% av

svara til Bing, men berre 62% av svara frå Gemini). Typiske fraser i tekstfragment som blei klassifiserte til å handle om tilknytning til næringslivet kunne vere «industry partnership» og «alignment with industry needs», og mange tekstar foreslo også spesifikke tiltak, som til dømes «internships», for å gjere studentar kjende med arbeidslivet.

Eit stort fleirtal av tekstane (80-90% avhengig av verktøy) hadde også med noko om at det var essensielt for kvaliteten at ein har faglærarar med høg kompetanse, eksemplifisert ved frasen «experienced faculty with industry and research expertise» i Sitat 1 øvst i Seksjon 4.



Figur 1. Resultat frå manuell analyse av tematikk i svara på P1. Score på 1 indikerer at alle tekstane var innom dette temaet, 0 at ingen var det.

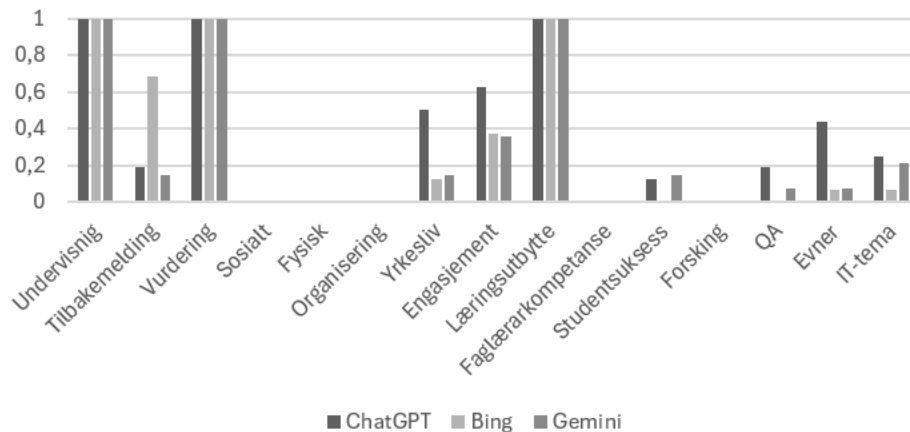
Ein del andre tema var med i rundt halvparten av tekstane, som i dette avsnittet blir gått gjennom frå venstre mot høgre i diagrammet. **Undervisning**, eksemplifisert ved «innovative teaching methods» i Sitat 1 – i andre tekstar til dømes «effective teaching», vart nemnt i 60% av tekstane til ChatGPT, men mykje sjeldnare av Bing (20%). Bing var derimot meir fokusert på **Sosialt læringsmiljø**, som var mindre framme hos dei andre. **Fysisk læringsmiljø** var noko meir uniformt fokusert mellom systema, med rundt halvparten av tekstane. Ofte handla dette om tilgjenge til moderne fasilitetar, laboratorium etc., som til dømes i Sitat 1. **Studentsuksess** undervegs og etter uteksaminering vart nemnt i over halvparten av tekstane frå Gemini, men godt under halvparten hos dei andre. For **Forsking** og **Kvalitetssikring** var det derimot Bing som nemnde dette oftast, medan dei andre nemnde det i mindre enn 30% av tekstane.

Som ein kan sjå lengst til høgre i Figur 1, varierte det i kva grad systema nemnde spesifikke **Evner** og **IT-tema** som studentane burde få gjennom utdanninga. ChatGPT hadde mest av dette i tekstane, medan Bing-tekstane hadde lite på Evner og ingen døme på spesifikke IT-tema. Alle tekstane sett under eitt var følgjande generiske læringsutbytte (Evner) nemnt i minst fem tekstar: kritisk tenking (17), problemløysing (14), evne til tilpassing («adaptability») (8), etiske vurderingar (6), kreativitet / innovasjon (6),

samarbeid (5). Av IT-spesifikke tema var dei klart oftast nemnde KI (11) og cybersikkerheit (11). Andre tema var nemnde berre få gonger, som «cloud» (4), «data science» (3), algoritmer (2), programmering (2), programmeringsspråk (1) og personvern (1).

Figur 2 viser tilsvarende fordeling av tematikk i svara på P2, der kvalitet i høgare IT-utdanning skulle forklarast i lys av samstemd undervisning («constructive alignment»). I desse tekstane var det slik at absolutt alle svara både frå ChatGPT, Bing og Gemini var innom tematikk rundt **Læringsutbytte**, **Undervisning** og **Vurdering**. Læringsutbytte var ein dominerande tematikk også i svara for P1, men det var litt skilnad i vinklinga. For P1 dreidde det seg mest om at pensum og læringsutbytte måtte vere oppdaterte og relevante i forhold til teknologisk utvikling og behova i arbeidslivet. For P2 dreidde det seg meir om at læringsutbytta måtte vere presist definerte.

Tilbakemelding til studentar vart nemnt i ein god del av svara på P2, særleg av Bing (nær 70%) – medan dette var nesten fråverande i svara på P1. Derimot var sosialt og fysisk læringsmiljø, samt faglærarkompetanse, ikkje eksplisitt nemnt i svara til P2, medan dette var mykje framme i svara på P1. Svara til P2 gjekk dessutan sjeldan inn på spesifikke IT-tema. Det var større tendens til å nemne spesifikke læringsmetodar og vurderingsformer, for å gje døme på samstemd undervisning.



Figur 2. Resultat frå manuell analyse av tematikk i svara på P2. Score på 1 indikerer at alle tekstane var innom dette temaet, 0 at ingen var det.

Omgrepet «constructive alignment» finst i alle svara på P2, primært i første setning der KI-systemet gjerne parafraserer promptet, men gjerne også for andre gong lengre nede i somme av tekstane. Derimot er det **ingen** av svara på P1 som inneheld omgrepet «constructive alignment» og heller ikkje ordet «constructive». Ordet «alignment» finst i om lag 15% av svara til P1, men da typisk for å uttrykkje at utdanninga må svare på behova i arbeidslivet, t.d. «industry alignment», «alignment with global standards», «alignment with emerging technologies».

4.1 Oppsummering av funn i lys av forskningsspørsmåla

Oppsummerte svar på dei fem forskningsspørsmåla blir som følgjer:

FS1: Kva perspektiv har KI-systema på kvalitet i høgare IT-utdanning, versus indikatorar i Studiebarometeret? Svara på P1 fokuserer mest på Læringsutbytte og Tilknytning til yrkeslivet – og dessutan med noko fokus på sosialt og fysisk læringsmiljø. I tillegg er det mykje fokus på Faglærarkompetanse, som ikkje er direkte adressert i Studiebarometeret. Svara på P2 fokuserer mest på Læringsutbytte, Undervisning og Vurdering, og i nokon grad på Tilbakemelding og Studentengasjement. Indikatoren Organisering frå Studiebarometeret er knapt nemnt.

FS2: I kva grad har svara innhald som er spesifikt for IT-utdanningar, versus innhald som kunne gjelde kvaliteten i all slags høgare utdanning? På P1 vart IT-spesifikke tema nemnde i mange av svara frå ChatGPT og somme frå Gemini. Tema som var oftast nemnt, var KI og cybersikkerheit. På P2 var dei aller fleste svara generiske med tanke på fagfelt.

FS3: I kva grad fører to ulike prompt – eitt meir ope, eitt meir spesifisert – til skilnader i svara som blir gitt? Som Figur 1 og 2 viser, er det tydelege skilnader mellom P1 og P2 sine svar. P2, med sitt fokus på samstemd undervisning, fører til at svara primært fokuserer på forholdet mellom klart definerte læringsutbytte, eigna undervisning og vurdering. Svara på P1 fokuserer derimot mindre på undervisning og vurdering, og meir på kvalitetar på utdanninga slik dei kan opplevast frå samfunnet utanfor, som at kandidatane får kompetanse som er relevant for yrkeslivet, at dei lukkast med å gjennomføre utdanninga og får gode jobbar, og at faglærarane har erfaring både frå undervisning, forskning og industri.

FS4: Kva skilnader er det mellom svara frå ulike KI-system? Ulike KI-system sine svar på same spørsmål er ganske like, jamfør semantisk likskap typisk rundt 0,85 for P1 og 0,9 for P2. Samstundes er det noko variasjon mellom dei. På P1 hadde ChatGPT klart større tendens enn andre system til å nemne IT-spesifikke fagtema som studentane burde lære om (særleg KI og cybersikkerheit), jamfør Figur 1. Bing nemnde aldri slike tema, men hadde derimot større tendens til å nemne sosialt læringsmiljø og studentar sine mogelegheiter til å bli involvert i forskning. Når gjennomsnittleg semantisk likskap likevel blir så høg, har det å gjere med at nesten alle tekstane bruker mesteparten av orda sine på å gå gjennom dei mest populære poenga, som er relevante læringsutbytte, tilknytning til yrkeslivet og kompetente faglærarar for P1 (Figur 1), og veldefinerte læringsutbytte med samsvarande undervisning og vurdering for P2 (Figur 2).

FS5: I kva grad viser svara utvikling over tid? Det var ikkje nokon klar tendens til utvikling over tid i svara – dei same sentrale tematikkane gjekk igjen, og det var ingen auke i semantisk distanse for svar som låg fleire månader unna kvarandre i tid. Spesifikt for ChatGPT sine døme på spesifikke IT-kunnskapar som studentane burde ha, kan det også seiast at KI og cybersikkerheit dukka opp allereie i forholdsvis tidlege svar, så det var ikkje slik at dette vart meir hyppig nemnt mot slutten av datainnsamlingsperioden.

5 Diskusjon

Resultata presentert i førre seksjon viser at KI-svara er innoom tematikk frå Studiebarometeret i ulik grad. Somme tema inngår i mange tekstar, som Læringsutbytte (både P1 og P2), Sosialt og Fysisk Læringsmiljø (særleg P1), Tilknytning til yrkeslivet (særleg P1), Undervisning (særleg P2) og Vurdering (særleg P2). Studentengasjement er også dekt i om lag halvparten av svara på P2, men lite i P1. Temaet Organisering finst knapt i tekstane, men dette kan vere forståeleg – da spørsmål om tilgjenge på og kvalitet på *informasjon om* studieprogrammet kan vere av interesse for studentar under studiet, men fort blir for internt og detaljert til å kome med i ein tekst om utdanningskvalitet der ein berre har 100 ord til disposisjon.

KI-svara inneheldt også tematikk som ikkje er lagt vekt på i Studiebarometeret. Det mest slåande er Faglærarkompetanse, som var nemnd i dei fleste svara på P1. Det kan likevel vere forståeleg at Studiebarometeret ikkje spør om dette, da ein nok har sett føre seg at det kan vere vanskeleg for studentane å vurdere direkte om lærarar er fagleg kompetente (iallfall for studentar på bachelornivå, som svarer på dette midt i sitt tredje semester som studentar) – men lettare for studenten å svare på korleis dei opplever undervisning, tilbakemelding og vurderingsformer. Faglærarkompetanse er også ein av dei mest markante skilnadene mellom svara på P1 – der det er med i nesten alle – og P2, der det slett ikkje er vektlagt. Grunnen til dette kan vere at fokus på «constructive alignment» dreier svara mykje i retning av samsvar mellom læringsutbytte, undervisningsmetodar og vurdering. Implisitt kan ein anta at det trengst ein kompetent fagstab for å greie å gjennomføre samstemd undervisning på ein god måte, men dette kjem ikkje fram i tekstane. Med berre 100 ord til disposisjon, bruker KI gjerne dei 20 første berre på å parafrasere spørsmålet og seie kva «constructive alignment» er, deretter 20 ord på kvart av dei tre hovudpunkta (klart definerte læringsutbytte, passande undervisningsmetodar, og vurderingsformer som samsvarer med utbytta som skal testast), og til slutt 20 ord på å runde av svaret. Det er dermed sjeldan plass til ei setning som eksplisitt seier at ein må ha kompetente fagfolk for å greie å gjennomføre samstemd undervisning.

Med tanke på FS2 (i kva grad tekstane nemner IT-spesifikke tema) var det for P1 særleg ChatGPT som gjorde dette, og mest hyppig nemnt var KI og cybersikkerheit (i 11 tekstar kvar). Programmering og algoritmer var kvar derimot berre nemnde i to av svara på P1 – sjølv om desse faga gjerne vil ta opp monaleg fleire studiepoeng i mange studieplanar. Ein spekulasjon med tanke på årsak til dette, kan vere at treningsmaterialet til KI-systema kan vere meir prega av dokument som skriv noko om fag som bør prioriterast opp i studia (til dømes at studentane bør lære meir om KI og cybersikkerheit) enn av dokument som skildrar kva som finst i studia allereie. At IT-kandidatar skal kunne programmere, er dessutan kanskje sett på som så innlysande at det er tatt for gitt, og dermed ikkje eksplisitt diskutert i så mange dokument om utdanningskvalitet.

Funna knytte til FS3 viser at tillegget av dei fem orda «with regards to constructive alignment» på slutten av P2 hadde stor effekt med tanke på svara i forhold til det opnare P1. Det er slåande at ingen av svara på P1 nemnde «constructive alignment». Jamvel om dette er sett på som eit viktig prinsipp for effektiv læring på universitetsnivå – jfr. t.d. at Biggs (1996) sin artikkel har over 6000 siteringar – er det altså ikkje noko som

KI-systema nemner «uoppfordra» i eit essay om kvalitet i høgare utdanning dersom dei berre har 100 ord til disposisjon. For P2 måtte svara derimot handle om dette, som gav større fokus på undervisning og vurderingsformer enn det som var tilfelle for P1. I nokon grad kan ein seie at P1 fekk svar med eit meir eksternt syn på kvalitet i høgare utdanning – at studentane lukkast, har relevant kompetanse og er i stand til å gjere nytte for seg i samfunnet når dei kjem ut, og at faglærarar har god kompetanse og industrikontakt. P2 fekk derimot svar med eit meir internt syn: korleis er læringsutbytte definert og korleis er undervisning og vurdering lagt opp for å nå desse læringsutbytta.

Relatert til FS4 var det større likskapar enn ulikskapar mellom svar på same prompt frå ChatGPT, Bing Copilot og Gemini – som ein kan sjå i Figur 1 og Figur 2 var dei same tematikkane dominerande hos alle tre. Visse skilnader var likevel tydelege, som at ChatGPT kom med meir IT-spesifikk terminologi på P1, og at Bing Copilot kom med meir om sosialt læringsmiljø. Det er vanskeleg å seie kva som er årsak til dette – om det til dømes er ulike dokumentsamlingar dei har blitt trena på, eller om det er ulikskapar i sjølve programkoden til systema som gjer det.

Relatert til FS5 var det ikkje nokon klar tendens til utvikling av svara over tid. Dette kan vere fordi kvalitet i høgare utdanning er eit tema det allereie er skrivne mykje om – og der mengda av eventuelt nytt materiale som måtte ha dukka opp akkurat i perioden juli 2024 – mars 2025 er monaleg mindre enn det som fantest frå før, pluss at det kanskje ikkje var så mykje radikalt nytt som skjedde heller. Utvikling over tid ville kanskje ha vore meir markant dersom promptet galdt eit tema som var mindre kjent på førehand, og/eller der det føregjekk store paradigmeskifte innan syn på utdanningskvalitet akkurat i den aktuelle perioden.

Ein annan observasjon som blei gjort, var tendensar til at svar varierte avhengig av kva nettlesar som vart brukt – dvs. same KI-system men ulik nettlesar (t.d., Microsoft Edge vs. AVG Secure Browser). Særleg ein deltakar prøvde å variere dette både for P1 og P2, og såg at svara vart ulike avhengig av nettlesar. Dette vart likevel ikkje undersøkt systematisk av alle deltakarane, og er dermed ikkje sett opp som eit forskingsspørsmål i artikkelen. Det er også uvisst kva som er årsaken til dette – ein mogleg spekulasjon kan vere at det er relatert til cookies i nettlesarane.

6 Konklusjon

Denne studien indikerer at gitt same spørsmål, vil dei tre generative KI-systema Bing Copilot, ChatGPT og Google Gemini også kome opp med ganske like svar – både frå gong til gong og systema imellom – iallfall på eit såpass allment og etablert tema som kvalitet i høgare utdanning. Derimot førte ei endring i spørsmålet – med tillegg om å sjå kvalitet i lys av samstemd undervisning – til ganske stor endring i kva svara fokuserte på. Svara blei også enda meir like både mellom systema og over tid med denne presiseringa av spørsmålet. Dette er for så vidt ikkje overraskande – di meir presist spørsmål, di meir presise svar kan ein forvente, så sant presiseringa retter seg mot noko som fanst i treningsmaterialet til KI-systema.

6.1 Veikskapar med studien

Berre 100 ord er ganske lite for å seie noko vettugt om kvalitet i IT-utdanning, og det kan tenkjast at resultatata ville ha blitt meir interessante med litt lengre tekstar. På den andre sida har eit kort format også fordelar, ved at det blir mindre tidkrevjande å analysere tekstane, og at ein tvinger KI-systema til å prioritere mellom mange ulike aspekt som kan trekkjast fram om utdanningskvalitet. Ein annan veikskap er at den manuelle analysen av tekstane vil vere prega av subjektive vurderingar, og særleg når det var berre éin person som gjorde analyse per tekst blir dette sårbart for variasjon i vurderinga, til dømes relatert til konsentrasjon og humør. Her kunne vi ha oppnådd større tryggleik for resultatet dersom fleire personar uavhengig hadde analysert kvar tekst, med samanlikning av resultat og registrering av usemje i etterkant. Sist men ikkje minst kan ein også spørje i kva grad det er interessant å finne ut kva tre KI-system svarer på spørsmål om kvalitet i høgare IT-utdanning. Kanskje ville det ha vore meir interessant om prompta var utforma på ein måte som inviterte til meir kontroversar, til dømes for å kunne avdekkje underliggjande veikskapar i KI-systema sitt syn på kvalitet. I tillegg kunne studien også testar på korleis KI-systema hadde “agert” om en hadde fylgd opp promptane med nye prompt for å be KI om å gje meir spesifikke svar på den fyste prompta, samt bruke ulike prompt-teknikkar for å lede KI-systemet i ei gitt retning. Det kan òg vere at funna er meir ein refleksjon av styrkane og veikskapane i dei valde KI-systema enn ein reell indikasjon på kvaliteten i utdanninga som heilskap

Referansar

- Bandi, A., & Zeng, R. (2024). Evaluation of the effectiveness of prompts and generative ai responses. In International Conference on Computer Applications in Industry and Engineering (pp. 56-69). Cham: Springer Nature Switzerland.
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher education*, 32(3), 347-364.
- Cohen, L., Manion, L and Morrison, K. (2018). Research methods in education, 8th ed. Routledge, London and New York.
- Green, D. (1994). What Is Quality in Higher Education?. Taylor & Francis, 1900 Frost Road, Bristol, PA 19007-1598.
- Hansen, G. og Sindre, G. (2023). IT-utdanningar i Noreg: Arbeidslivsrelevans og «Tilknytning til yrkeslivet» i Studiebarometeret. MNT-konferansen, Stavanger, april 2023.
- Karlsen, J. R., Müftüoğlu, I. B., & Slaattelid, R. T. (2017). Studiebarometeret som kunnskapsform og samfunnsfenomen. *Nytt norsk tidsskrift*, 34(1), 7-19.
- Krippendorff, K. (2018). Content Analysis: An Introduction to Its Methodology (4. utg.). SAGE.
- Munthe, E. (2020). Studentengasjement i høyere utdanning: Hvilken innsikt gir Studiebarometeret?. *Uniped*, 43(2), 104-116.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language.

Sindre, G. (2021). Studiebarometeret på programnivå: effekt av vurderingsformer og obligatoriske øvinger?. Norsk IKT-konferanse for forskning og utdanning (No. 4), UDIT'2021.

Slette, A. L., og Johansen, G. (2025). Notions of Quality and Excellence Within a Centre for Excellence Initiative in Higher Education. Higher Education Policy, 1-19.

Welzant, H., Schindler, L., Puls-Elvidge, S., & Crawford, L. (2015). Definitions of quality in higher education: A synthesis of the literature. Higher Learning Research Communications, 5 (3).

Wiers-Jenssen, J., og Hovdhaugen, E. (2019). Studieinnsats på lavere grad – hva kan Studiebarometeret fortelle oss? Uniped, 42(3), 274-289.