

Teknologi og digitalisering

I denne spalten vil Lars Erlend Leganger og noen av hans kollegaer skrive om aktuelle temaer innen teknologi/digitalisering som direkte eller indirekte også vil påvirke revisors hverdag. Lars Erlend er AI-ekspert og direktør i PwC. Han har en PhD i teoretisk fysikk fra NTNU.

Erfaringer med bruken av AI

I denne artikkelen diskuterer vi forskjellige strategiske tilnærminger til å hente ut verdi fra kunstig intelligens, og deler PwC Norges erfaringer med bruk av ChatPwC, en internutviklet chatbot basert på generativ kunstig intelligens.



PhD
Lars Erlend Leganger
Direktør i PwC



MSc
Tuva Martin
Data Scientist i PwC

Hvorfor bygge selv når det finnes hylleware?

For å ta det åpenbare først – hvorfor bygge en egen chat-løsning når det finnes gode hylleware-alternativer? For oss i PwC var det særlig tre faktorer som spilte inn:

1. Gi folk tilgang...

For oss var det et mål å tilgjengeliggjøre generelt anvendbar kunstig intelligens (heretter «GenAI», fra engelsk «General Purpose AI»¹ for alle i bedriften: Vi vet ikke på forhånd hvor stor verdi forskjellige virksomhetsområder og roller ville finne i GenAI chat etter hvert som de blir kjent med verktøyet, og vi har stor tro på å «gi folk tilgang på nye verktøy og se hva som skjer»-tilnærmingen til teknologidrevet innovasjon. Det skader heller ikke at en med en bred utrulling

slipper å bruke tid og krefter på lisens-administrasjon og behovsvurderinger. Ulempen med en slik bred utrulling er at den kan bli kostbar. Hylleware-GenAI-lisenser koster typisk mellom 3000 og 6000 kroner i året pr. hode – det er sure penger å betale for eventuelle brukere som ikke finner verdi i verktøyet.

GenAI-modeller er også tilgjengelige som skytjenester med enkle integrasjonsgrensesnitt (heretter «API», fra engelsk «Application Programming Interface»). Når en bruker GenAI-APIer betaler man ikke for antall brukere, men for bruk, f.eks. antall chat-meldinger sendt. Med en internutviklet chat-løsning bygget på GenAI-APIer kan vi derfor anvende «gi folk tilgang og se hva som skjer»-innovasjonsstrategien uten å bekymre oss for å betale for lisenser som ikke blir brukt: En inaktiv bruker som knapt sender meldinger til GenAI-APIet, genererer ingen signifikante ekstrakostnader.

2. ... og se hva som skjer!

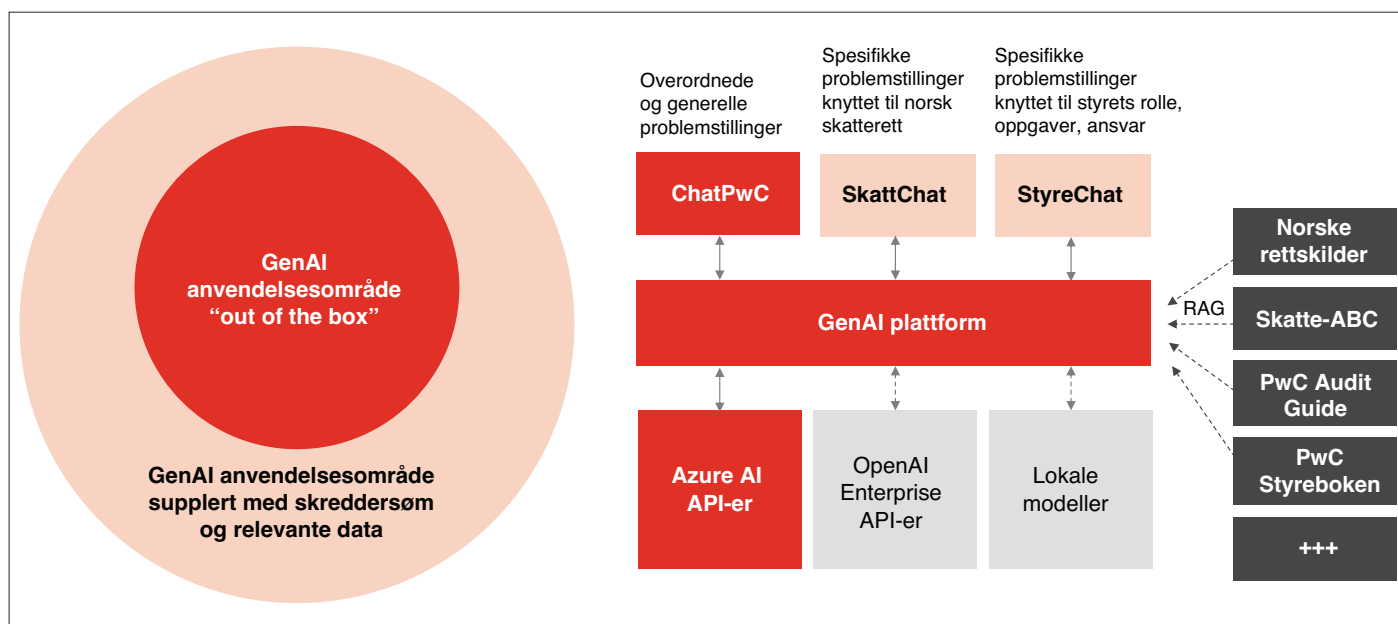
Når de ansatte først har fått tilgang, hvordan kan vi best samle erfaringer

om hvor GenAI fungerer bra, og hvor det fungerer dårlig? Hvordan finner vi hvilke typer brukere som kommer raskt i gang, og hvilke som trenger mer støtte for å realisere potensialet i verktøyet? En kan alltid gjøre brukerintervjuer og sende ut spørreskjemaer, men enda bedre er det om vi kan analysere den faktiske bruken: Vi mener Jakob Niensens «basic rules of usability»² er vel så relevante når en skal utforske anvendelsesområdet for ny teknologi: «*Watch what people actually do. Do not believe what people say they do. Definitely 'on't believe what people predict they may do in the future.*»

Med en internutviklet chat-løsning kan vi hente ut anonymiserte meldingslogger og løpende monitorere faktisk GenAI-bruk på et helt annet detaljnivå enn hva de fleste hylleware-løsninger legger opp til. Analysene vi deler i denne artikkelen hadde ikke vært mulige å gjennomføre uten nærheten til bruksdataene en internutviklet løsning gir.

¹ Det norske Språkrådet har, i motsetning til våre naboland, gått inn for forkortelsen «KI» for kunstig intelligens. Som et slag for språklig internkonsistens og vern mot anglisering av norsk er forsøket på å innføre «KI» forbilledlig. Samtidig er mye av litteraturen som skrives om kunstig intelligens og annen teknologi på engelsk, og for enkelthets skyld (og litt fordi gammel vane er vond å vende) holder vi enn så lenge fast på enkelte innarbeidede tekniske engelske forkortelser som «AI» og «API».

² Jakob Nielsen (2001). First Rule of Usability? Don't Listen to Users. <https://www.nngroup.com/articles/first-rule-of-usability-dont-listen-to-users/>



Figur 1: Mange prosesser og beslutninger krever kjennskap til interne data som ikke er tilgjengelig ut av boksen for GenAI-modellene. En plattform som bl.a. støtter RAG (retrieval augmented generation) gjør det enklere og tryggere å nyttiggjøre seg av nye data i skreddersydde løsninger og prosjektleveranser.

3. Erfaringsbygging med – og infrastruktur for – skreddersydde løsninger

Sist, men ikke minst: GenAIs anvendelsesområder kan økes drastisk ved å kombinere grunnleggende GenAI-modeller med prosess-spesifikke skreddersydde løsninger og domenespesifikke interne datasett. På sikt forventer vi at frittstående generelle GenAI-verktøy gradvis vil erstattes av løsninger med domenespesifikke tilpasninger og tilgang på bedriftsinterne data. ChatPwC-infrastrukturen er designet for å gjøre det enkelt og sikkert å bygge slike skreddersydde GenAI-løsninger for både interne prosesser og eksterne prosjektleveranser, se figur 1.

Hvordan brukes GenAI i PwC (så langt)?

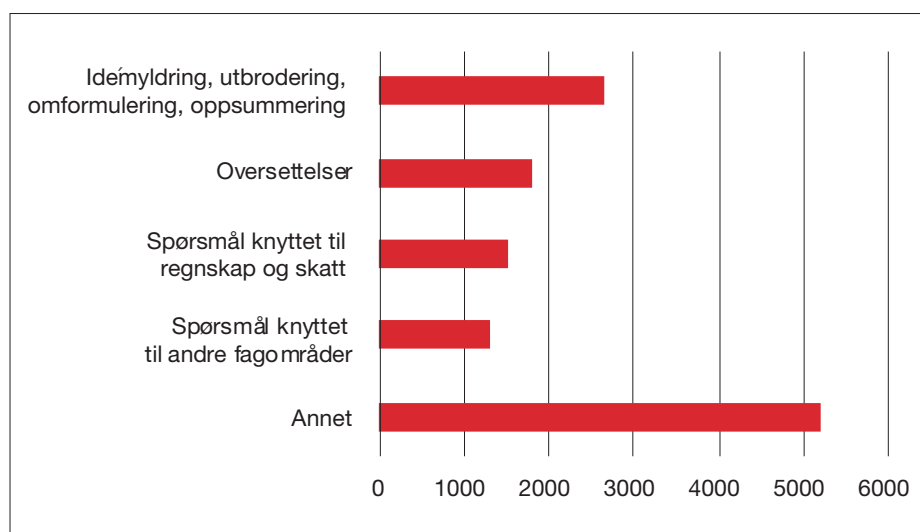
Et par måneder etter bred utrulling av ChatPwC gjorde vi en analyse av bruken så langt, der vi blant annet brukte AI til å kategorisere meldingene brukerne hadde sendt inn,³ se figur 2. De tre mest utbredte kategoriene er inn-

holdsproduksjon (idémyldring, førsteutkast, omskrivning), oversettelse, og spørsmål vedrørende regnskap og skatt – disse utgjorde til sammen over halvparten av alle forespørsler til ChatPwC i perioden vi analyserte.

1. Idémyldring/førsteutkast og utbrodering/omformulering er oppgaver ChatPwC er velegnet for. Evnen til å skrive presise forespørsler («prompts») for denne typen oppgaver er noe av det som

utmerker seg ved GenAI-«superbrukere», men også enkle instruksjoner gir ofte gode resultater for denne typen oppgaver:

- «Rewrite: We are already familiar with the management, which ensures good cooperation»
- «Kan du lage et lengre notat, i rapportstil»
- «skal ringe et selskap ifra sveits og purre på en bekreftelse, kan du skrive et manus på engelsk til meg?»
- «Kan du korte ned teksten til ca. halvparten?»



Figur 2: ChatPwC brukes mest til idémyldring/førsteutkast, utbrodering/omformulering, oversettelser, oppsummeringer og spørsmål knyttet til regnskap og skatt.

³ Vi overlater til leseren å vurdere om det å bruke AI for å studere bruk av AI utgjør elegant bruk av ny teknologi, eller om det er nok et endetidstegn, jf. Revisjon & Regnskap nr. 7 2023 AI-etikk vs. AI-sikkerhet.

2. Enkle oversettelser klarer ChatPwC fint «ut av boksen», og den gir også ofte brukbare resultater på mer krevende oversettelser dersom man i forespørselen inkluderer eksempler og informasjon om fagspesifikk terminologi:

- «Hva kan «general ledger» oversettes til på norsk innen regnskap og revisjon?»
- «Hva er «international development aid» på norsk?»
- «Kan du gi meg eksempler på engelsk hvordan man avslutter en e-post og ønsker god helg når den tid kommer?»
- «Translate to norsk: We believe that we are an ideal partner to guide [client] through the process and to overcome potential obstacles.»

Fagspesifikke oversettelser er et godt eksempel på områder der spissede løsninger kan være aktuelle, beriket med relevante eksempler, definisjoner, m.m.

3. Spørsmål knyttet til regnskap og skatt forekommer (overraskende?) ofte. Faktaspørsmål er generelt et anvendelsesområde som innebærer høy risiko for AI-hallusinasjoner⁴, så her er det viktig at brukerne er årvåkne, spesielt ved spørsmål om særnorske og perifere temaer som de grunnleggende GenAI-modellene ikke kan forventes å ha sett mye til i treningsdataene sine:

- «Er kampanjekostnader skattesensitive kontoer?»
- «Hei! Er transaksjonskostnader fradragserrettiget i henhold til NGAAP?» «hvordan skal parkering behandles iht. IFRS 16»
- «elaborate on the IFRS 3 part»

At ChatPwC tilsynelatende brukes i såpass stor grad som oppslagsverk for regnskap og skatt fordrer å gjøre

⁴ Den herostratisk berømte Steven Schwartz, som refererte til ChatGPT-hallusinert rettspraksis i et rettsdokument, er et godt eksempel på hvor ille det kan gå når en forveksler et GenAI-verktøy med en søkemotor eller oppslagsverk: <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/>

videre analyser i presisjonen i svarene på slike spørsmål, og – om nødvendig – anskaffe (eller bygge internt) GenAI-verktøy der de grunnleggende modellene er supplert med tilstrekkelig detaljkjennskap til norske regnskapsregler og skatterett.

Hvem er brukerne?

Etter to måneder ser vi konturene av tre typer brukerreiser for ChatPwC: Halvparten av de som har prøvd ChatPwC ender opp som «ikke-brukere» – de er innom en gang eller to før de mister interessen. En tredjedel blir «vanlige brukere» som benytter ChatPwC en håndfull ganger i uka, og en sjettedel av de som har testet verktøyet, har blitt «superbrukere» som benytter ChatPwC flere ganger hver dag.

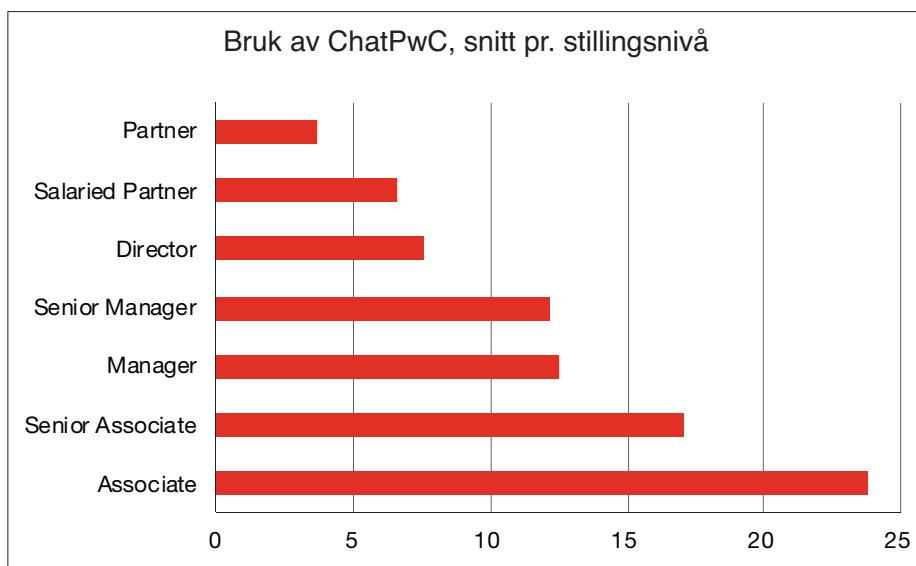
Et viktig spørsmål å grave videre i her er årsaken(e) til at «ikke-brukerne» gir opp: Er GenAI lite anvendelig for deres arbeidsoppgaver, går det på manglende kjennskap, eller frykt for å gjøre noe galt? Foreløpige analyser indikerer at brukerne som kun prøver GenAI en håndfull ganger før de gir opp, ofte har stilt spørsmål som ligger utenfor kompetanseområdet til den underliggende GenAI-modellen, f.eks. knyttet til data som modellen ikke hadde tilgang på i trening (bedriftsinterne dokumen-

ter, hendelser skjedd etter at modellen ble sist trent, osv.):

- «Hvor benter man ut bankbrev?»
- «Hva er motivasjonen til å ta MRR?»
- «Gjennomsnittlig valutakurs for 2023 euro mot russiske rubler»
- «Hvordan lagrer PwC data?»

Den som forventer å få svar på slike spørsmål fra et generelt ikke-skreddersydd chatverktøy, vil fort bli skuffet. En lærdom vi leser ut av dette er viktigheten av å tydeliggjøre for brukerne – gjennom opplæring og i verktøyets brukergrensesnitt – hvilke data det har/ikke har tilgang på, og at ChatPwC, som andre GenAI-løsninger, ofte hallusinerer i stedet for å melde pass når de er på tynn is.

Demografisk sett er trenden klar: Aktivitetsnivået på ChatPwC er omvendt proporsjonalt med stillingsnivået. De yngste ansatte, Associates, bruker ChatPwC mest, mens partnerne bruker det minst, se figur 3. Høyere bruk pr. ansatt på juniornivå enn på seniornivå kan nok delvis forklares med at juniorer typisk har flere GenAI-kompatible oppgaver (innholdsproduksjon, oversettelse, osv.), men også her vil det være interessant å grave videre ned i dataene fremover.



Figur 3: Junior-ansatte bruker ChatPwC mer enn senior-ansatte – Associates skriver i snitt seks ganger så mange chat-meldinger som partnere.

Hva betyr GenAI-utviklingen for din AI-strategi?

Tradisjonelt har verdiskapning med kunstig intelligens handlet om å bygge skreddersydde AI-løsninger for spesifikke, snevert definerte problemstillinger – fra musikk anbefalinger til kredittvurdering. Beslutningsmodellene i slike løsninger bygges ofte (men ikke alltid) ved å la maskinlæringsalgoritmer lete etter mønstre og sammenhenger i data som anses relevante for problemstillingen.

En musikk-anbefalings-AI «lærer» fra data om hvilke sanger du, og brukere som ligner på deg, har likt og lyttet til tidligere. En kredittscore-AI kan ta utgangspunkt i historiske data om hvilke kunder som har betalt renter og avdrag på forbrukslånene sine, og hvilke som ikke gjorde det. Slike systemer kan ofte overgå mennesker i treffsikkerhet (og ikke minst effektivitet) i vurderingene de gjør, men kun innenfor den snevert definerte oppgaven de er bygget for. En skreddersydd musikk-anbefaling-AI er (sannsynligvis⁵) ubrukelig for å vurdere kredittrisiko, og omvendt.

Tradisjonelle spesifikke-AI-systemer, med modeller trent fra bunnen av på problem-

spesifikke data, kan være kostbare å utvikle og drifte: Maskinlæringsingeniører trengs for å bygge og produksjonssette modellene på en forsvarlig måte, og bygging og vedlikehold av tilhørende dataflyter og systemintegrasjoner er komplekse fag i seg selv. Dette begrenser anvendelsesområdet for tradisjonell AI: Kun prosesser med stor potensiell effektiviseringsgevinst (f.eks. høyt volum, kostbar arbeidskraft) og/eller stor potensiell kvalitetsgevinst (f.eks. prisberegning i pris-sensitive markeder, vurderinger med høye gebyrer/straff/omdømmeskader ved feil) har tradisjonelt kunnet forsvare de nødvendige investeringene for å utvikle og produksjonssette AI-løsninger.

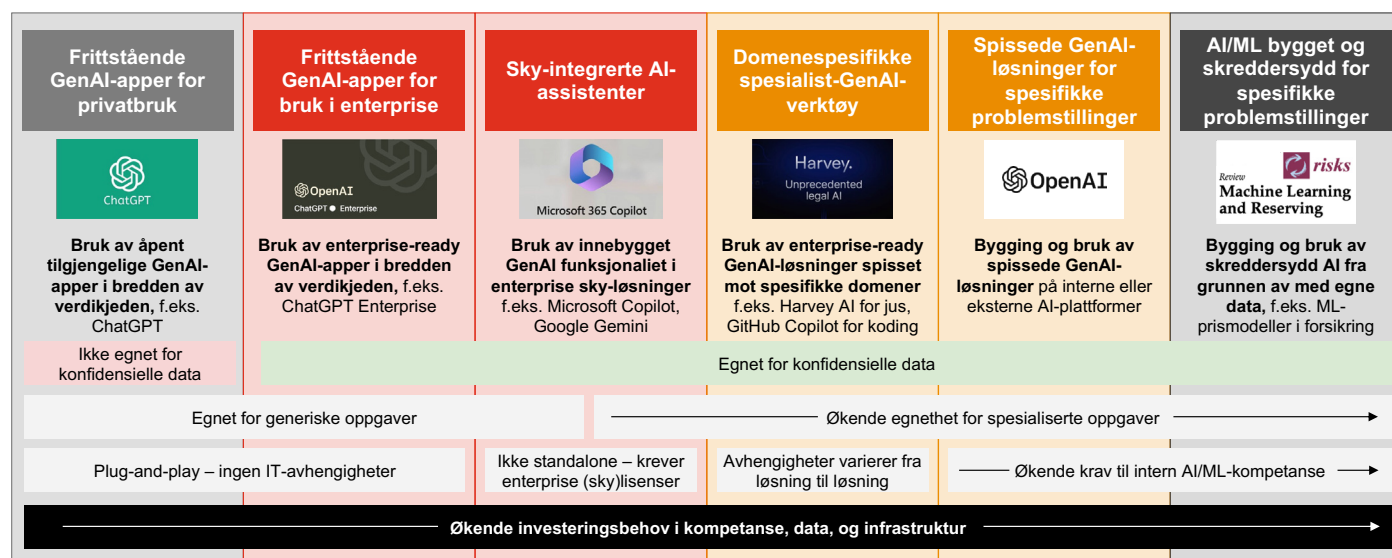
Det fremvoksende økosystemet av GenAI-verktøy og -APIer stiller nye krav til selskapers AI-strategi. Etter hvert som GenAI-verktøyenes kapabiliteter øker, og implementasjonskostnadene for skreddersydde løsninger synker, kan AI-anvendelser som tidligere ikke lot seg forsvare økonomisk, fort likevel bli aktuelle kandidater for AI-automatisering og -beslutningsstøtte. Den som skal utforme en AI-strategi nå, står foran et bredt spekter av implementeringsstrategier, fra frittstående GenAI-apper som ChatGPT og Harvey AI, sky-integrerte AI-assistentter som Microsofts Copilot og Googles Gemini, til skreddersydde løsninger, med

eller uten GenAI-APIer, for spesifikke problemstillinger – se figur 4.

Et første skritt for mange vil være å undersøke hvilken nytte generelle GenAI-verktøy som ChatGPT og Copilot gir i forskjellige deler av verdikjeden. For selskap med mange ansatte og sterke teknologimiljøer kan skalerbare internutviklede løsninger utgjøre kostnadseffektive og innovasjonsvennlige alternativer til hyllevere i utforskningsfasen, lik de vi i denne artikkelen har skissert med ChatPwC.

For å ta skrittet videre fra frittstående GenAI-verktøy og utnytte GenAI fullt ut må selskaper bygge erfaringer med hvordan AI-løsninger kan integreres med eksisterende IT-infrastruktur, hvordan interne og eksterne data kan brukes til å tilpasse AI-modeller for spesifikke behov, og ikke minst hvordan man sikrer at AI-løsninger opererer innenfor etiske og juridiske rammer, i tråd med personvernforordningen, EUs kommende AI act (som etter all formodning også blir en del av norsk rett), samt øvrig generell og sektorspesifikk lovgivning. Det krever en strategisk tilnærming som aktivt og kryssfunksjonelt undersøker nye muligheter for å anvende AI til å skape verdi, enten det er gjennom bruk av GenAI-hyllevere, eller integrasjon av GenAI i selskapets egne produkter og tjenester.

5 Hvorvidt det foreligger statistisk signifikante korrelasjoner mellom musikksmak og kredittverdighet er for så vidt et interessant spørsmål, som vi ikke går videre inn på her.



Figur 4: Generiske oppgaver kan ofte løses greit av GenAI-apper «ut av boksen» uten tilpasninger. Mer spesifikke oppgaver krever fremdeles spissing/skreddersydd opp mot det enkelte domene og den enkelte problemstilling, men ofte med utgangspunkt i forbåndstrengte GenAI-modeller.