KI-lekse 3: Sjekk hvor presis KI-en er

For barn på 5.–10. trinn og en voksen · ca. 30 minutter · én skjerm, som den voksne styrer

> **KI-leksene henger sammen (3 av 4):** I lekse 2 så dere hvordan modellen gjetter neste ord. Her undersøker dere hvor presis den er på et område dere selv kan noe om — og hva som skjer når dere er uenige med den.

Til deg som er voksen

Én ting først: I denne leksa er det du som skriver til KI-en, ikke barnet. De nasjonale rådene sier at elever på barnetrinnet i all hovedsak ikke skal bruke generativ KI selv, og at bruken på ungdomstrinnet innføres gradvis. Å lære *om* KI — undersøke den, prøve den, vurdere svarene — er ikke omfattet av den begrensningen. Barnet bestemmer spørsmålene og vurderer svarene; du betjener verktøyet.

Hva dere skal se etter har endret seg. De første årene med chatboter handlet dette om åpenbart oppdiktede svar. Slike forekommer fortsatt, men i dagens verktøy er de sjeldnere: modellene er blitt bedre, og de fleste søker nå på nettet før de svarer. Det som står igjen, er vanskeligere å oppdage. Svaret er i hovedsak riktig, mens detaljene er feil — årstall, tall, navn, hvem som gjorde hva, hva som gjelder nå. Det er også dette som gjør feilene praktisk viktige: et helt oppdiktet svar avsløres av seg selv, mens en gal detalj i et ellers korrekt svar går rett igjennom.

To andre trekk er verdt å kjenne til, og begge kan dere undersøke selv:

- **Modellen svarer heller enn å la være.** En studie publisert i *Nature* i april 2026 viser hvorfor: måten modeller måles på, gir uttelling for å svare og teller «vet ikke» som feil. Da blir det å gjette den beste strategien, også når grunnlaget er tynt. Nyere modeller sier oftere fra om usikkerhet enn før, men tendensen til å svare ligger i systemet.
- **Modellen har lett for å gi brukeren rett.** En studie i *Science* i mars 2026 fant at modeller bekreftet brukerens standpunkt 49 % oftere enn mennesker gjorde. En annen studie, i *Nature* samme måned, fant at jo vennligere og varmere en modell er innstilt, jo oftere bekrefter den feil oppfatninger hos brukeren — omtrent 40 % oftere. Det betyr at «den var enig med meg» ikke er noe kvalitetstegn.

Dette trenger dere

En chatbot på den voksnes enhet (for eksempel ChatGPT, Copilot eller Gemini — gratisversjon er tilstrekkelig). Et ark med fire kolonner: **Spørsmål · Riktig · Detaljfeil · Vet ikke**. Og et tema familien kan bedre enn de fleste.

Én regel før dere starter: Ikke skriv inn navn, adresse eller andre opplysninger om dere selv eller folk dere kjenner. Dere skal undersøke verktøyet, ikke gi det informasjon.

Slik gjør dere det

1. **Velg området deres.** Fotballklubben, hesteinteressen, bygda eller bydelen, et yrke i familien, en hobby. Det beste temaet er noe dere kan mye om, men som det er skrevet lite om — der er forskjellen mellom dere og modellen størst.
2. **Lag fem spørsmål sammen, fra generelt til spesifikt.** Begynn bredt («Hva kjennetegner en islandshest?») og snevre inn mot det etterprøvbare («Når ble klubben vår stiftet, og hvilken divisjon spiller A-laget i nå?»). Spørsmål med årstall, tall,

navn og «hva gjelder nå» er de mest nyttige, fordi svaret enten stemmer eller ikke.

3. **Vurder svarene detalj for detalj.** Les hvert svar høyt og gå gjennom opplysningene én og én. Er helheten riktig, men året feil? Er navnet nesten riktig? Er noe utdatert? Noter i kolonnen for detaljfeil. Merk også når modellen sier at den ikke vet, eller tar forbehold — det er et bedre svar enn en presis påstand uten dekning, og verdt å si høyt.
4. **Undersøk kildene, ikke bare om de finnes.** Skriv: «Hvor har du dette fra? Oppgi lenker.» Åpne så lenkene. To ting kan gå galt, og begge er dokumentert i en gjennomgang av kommersielle KI-verktøy fra april 2026: mellom 3 og 13 % av lenkene modellene oppgir, viser til nettsider som aldri har eksistert, og mellom 5 og 18 % av alle oppgitte kilder virker ikke. Vel så vanlig er det at kilden finnes, men ikke sier det modellen påsto den sa. Sjekk begge deler: finnes siden, og står det der?
5. **Prøv å få modellen til å skifte mening.** Når dere har fått et svar dere vet er *riktig*, si imot: «Det stemmer ikke, jeg mener det var i 1994.» Se hva som skjer. Holder den på svaret og begrunner det, eller gir den dere rett? Dette er den mest nyttige testen i hele leksa, fordi den viser noe man ikke ser i et enkeltsvar: at enighet fra en KI ikke betyr at man har rett.
6. **Oppsummer.** Hvor mange svar var riktige i alle detaljer? Hvor mange var riktige i hovedsak, men gale i en detalj? Sa den fra når den var usikker? Ga den seg da dere sa imot? Snakk om hva dette betyr for hvordan man bruker slike svar.

Snakk om det etterpå

- **Var detaljfeilene lette å se?** Nei, og det er poenget. De står i samme rolige, korrekte språk som resten. Forskjellen er ikke synlig i teksten, bare i sammenligningen med noe man vet.
- **Hva må man kunne for å oppdage en KI-feil?** Noe om temaet. Uten egen kunnskap har man ingenting å kontrollere mot. Det gjelder også når svaret har lenker: lenker må leses, ikke telles.
- **Hva betyr det at den ga seg da vi sa imot?** At den er laget for å være til nytte og lett å ha med å gjøre, ikke for å ha rett. Enighet er ikke bekreftelse.
- **Når betyr det noe om svaret er riktig?** Ideer og formuleringer: mindre kritisk. Helse, regelverk, skolearbeid, nyheter, penger: kontroller.

Godt å vite (for den voksne)

Hvorfor forsvinner ikke feilene? Fordi de ikke er en feil i programmet, men en følge av framgangsmåten. Modellen beregner sannsynlig tekst, ikke sanne påstander. Nettsøk gjør svarene bedre forankret, men flytter feilen: nå kan modellen hente en kilde og gjengi den upresist i stedet for å finne på alt selv.

Hvorfor sier den ikke oftere «jeg vet ikke»? Fordi den måles på om den svarer riktig, og et «vet ikke» teller som feil i de fleste målingene. Det gjør gjetting til en lønnsom strategi. Dette er hovedpoenget i *Nature*-artikkelen fra april 2026 — problemet ligger i hvordan modellene evalueres, ikke bare i modellene.

Hvorfor er den svakere på norske forhold? Fordi modellene er trent på det det finnes mye av, og det er i hovedsak engelskspråklig materiale fra amerikansk kontekst. Norsk regelverk, norsk arbeidsliv, lokalhistorie og dialekt er tynnere dekket. Nynorsk er et konkret eksempel: da Språkrådet testet ChatGPT i 2024, fant de om lag tre ganger så mange språkfeil per side på nynorsk som på bokmål. Vil dere teste dette, kan dere be om samme svar på nynorsk og se hva som skjer.

Hvorfor svarer den ulikt hvis vi spør igjen? Fordi den beregner svaret på nytt hver gang. Samme spørsmål kan gi ulike svar, og det er en grunn til at «KI-en sa det» ikke er en kilde.

Kilder

- Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. (22.04.2026): *Evaluating large language models for accuracy incentivizes hallucinations*. Nature 653. Om hvorfor evalueringsmåten belønner gjetting framfor «vet ikke».
- Cheng, M., Lee, C., Yu, S., Han, D., Khadpe, P. & Jurafsky, D. (26.03.2026), Science, omtalt i Stanford Report. Elleve modeller bekreftet brukerens standpunkt 49 % oftere enn mennesker.
- Ibrahim, L., Hafner, F. S. & Rocher, L. (29.04.2026): *Training language models to be warm can reduce accuracy and increase sycophancy*. Nature 652. Varmere modeller bekreftet feil brukeroppfatninger om lag 40 % oftere.
- Rao, D., Wong, E. & Callison-Burch, C. (03.04.2026): *Detecting and Correcting Reference Hallucinations in Commercial LLMs and Deep Research Agents*. 3–13 % av oppgitte lenker var oppdiktet; 5–18 % av kildene virket ikke.
- Språkrådet (2024): språktest av ChatGPT på bokmål og nynorsk — 2,6 feil per side på bokmål mot 8 per side på nynorsk.

Vurderingsskjema

Tema: _____

Sett kryss i riktig rute for hvert spørsmål.

Spørsmål	Riktig	Feil	Finner den på?
_____	☐	☐	☐
_____	☐	☐	☐
_____	☐	☐	☐
_____	☐	☐	☐
_____	☐	☐	☐

Oppsummering: _____ riktige av fem