FFI Forsvarets forskningsinstitutt

UTVIDET SAMMENDRAG

Store språkmodeller og forsvarlig sikkerhet

En studie av sårbarheter, trusler og mulige tiltak for bruk i egen infrastuktur

Utvidet sammendrag av FFI-rapport 26/034

Det vil alltid være risiko forbundet med å bruke store språkmodeller. Samtidig er nytteverdien så stor at vi ikke kan unnlate å bruke dem. Denne rapporten gir grunnlag for å forstå risikoen og hva man kan gjøre for å redusere den.

Store språkmodeller (LLM-er) gir nye muligheter til å løse oppgaver som tidligere krevde betydelig menneskelig innsats. Når modellene kobles til verktøy, data og andre systemer, kan de også brukes som agenter som ikke bare produserer innhold, men utfører handlinger.

Samtidig introduserer teknologien nye sårbarheter og trusler. For virksomheter innen nasjonal sikkerhet, inkludert Forsvaret, må gevinstene ved å ta teknologien i bruk veies mot risikoen den introduserer.

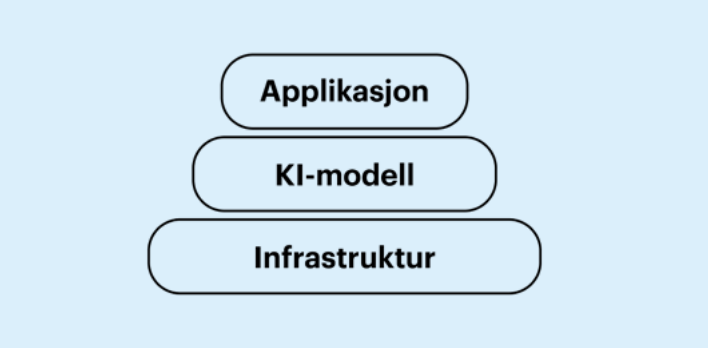
Denne rapporten bidrar til et bedre grunnlag for en slik avveining. Målet er å forklare sentrale sikkerhetsutfordringer og beskrive tiltak som kan redusere risikoen. Rapporten tar utgangspunkt i systemer der virksomheten kontrollerer modellen og infrastrukturen.

Store språkmodeller og nytteverdi

En stor språkmodell er i bunn og grunn en statistisk modell som er trent på svært store mengder tekst. Den genererer tekst ved å beregne sannsynligheten for neste tekstdel (token) basert på teksten som allerede finnes i konteksten.

Gjennom førtrening lærer modellen språk, mønstre og kunnskap fra store datamengder. Gjennom ettertrening lærer den å anvende kunnskapen ved å følge instruksjoner og løse oppgaver.

Dette gjør at språkmodeller kan brukes til blant annet innholdsgenerering, analyse, og programmering. Koblet til datakilder og verktøy kan de også hente informasjon og utføre handlinger. Det øker nytteverdien, men også sikkerhetsrisikoen.



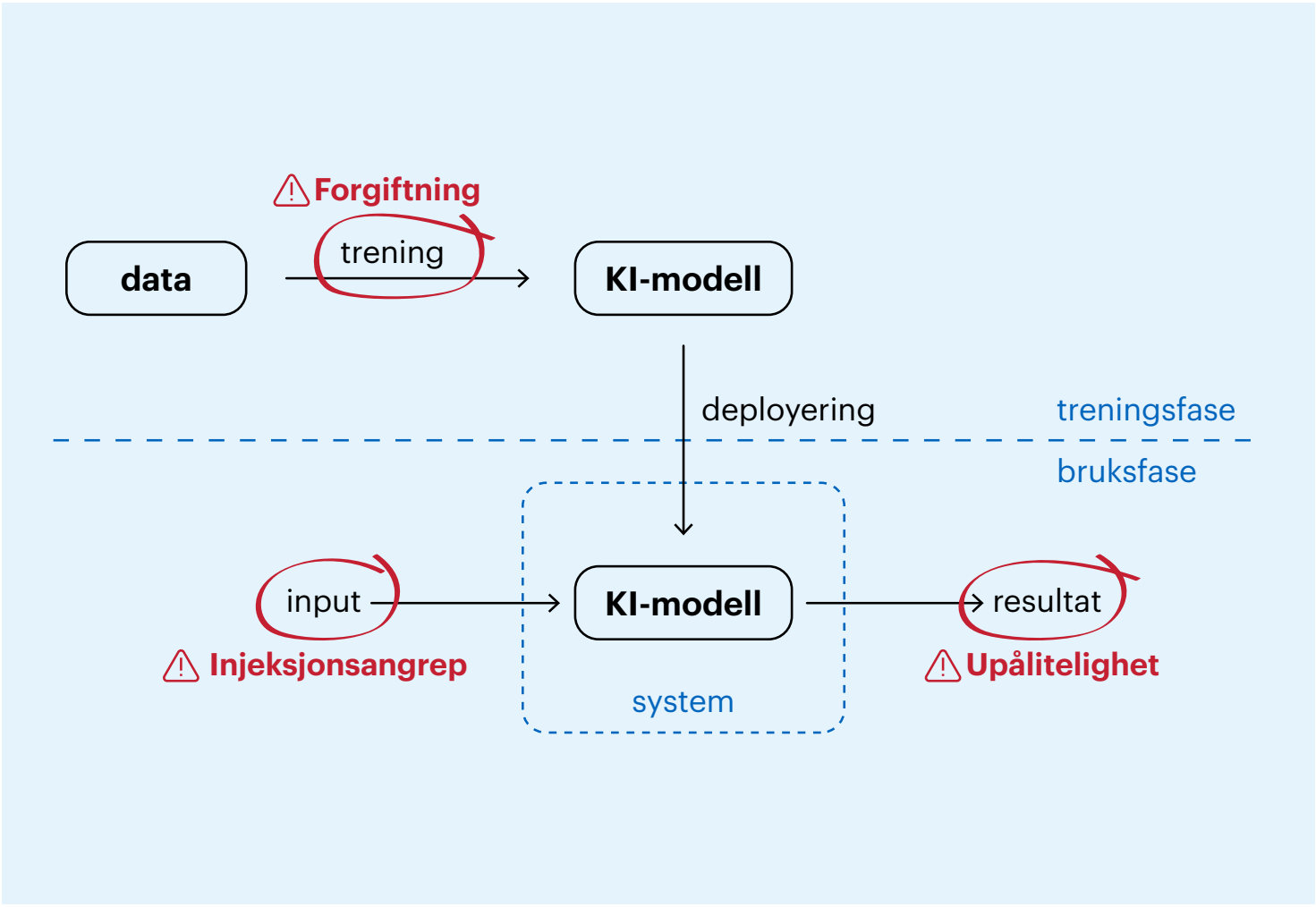
Språkmodeller og agentisk KI brukes i applikasjoner (programmer) som igjen er en del av en virksomhets IKT-infrastruktur. Sårbarhetene i modellene gjør at vi i praksis må behandle dem som en insider-risiko.

Sårbarheter og trusler i språkmodeller

Det finnes tre grunnleggende utfordringer ved store språkmodeller:

- **Vi kan ikke fullt ut vite hva modellen har lært**
- **Modellen kan påvirkes og overstyres gjennom input**
- **Svarene vi får er ikke fullt ut pålitelige**

Disse problemene følger i stor grad av hvordan språkmodellene er utviklet og trent, og kan derfor ikke fjernes uten at man samtidig påvirker egenskapene som gjør modellene nyttige.



Språkmodeller er sårbare både under trening, gjennom inputen de får fra brukere og fordi vi aldri kan stole helt på resultatene og svarene de gir.

Forgiftningsangrep kan skje ved at treningsdata eller modellen manipuleres for å innføre uønsket funksjonalitet. Risikoen er særlig relevant ved bruk av ferdigtrente modeller der virksomheten har begrenset innsyn i treningen og leverandørkjeden.

Injeksjonsangrep utnytter at modellen behandler instruksjoner og data gjennom den samme tekstlige kanalen. Instruksjoner skjult i dokumenter, nettsider eller andre eksterne kilder kan påvirke hvordan modellen oppfører seg. Problemet blir særlig viktig i systemer som automatisk henter informasjon fra kilder virksomheten ikke kontrollerer.

Upålitelighet er en utfordring også uten en angriper. Modeller kan hallusinere, misforstå informasjon eller finne uønskede snarveier for å løse en oppgave.

Et viktig poeng er at konsekvensene av alle utfordringene viser seg i modellens resultat. Fra utsiden kan det derfor være vanskelig å vite om uønsket oppførsel skyldes et angrep, påvirkning av modellen under trening eller at modellen ganske enkelt har gjort en feil.

Sårbarheter i LLM-applikasjoner og agenter

En språkmodell brukes alltid som del av en applikasjon – det vil si en programvare som gjør modellen nyttig for en bestemt oppgave. Sikkerheten kan derfor ikke vurderes ut fra modellen alene. I applikasjonslaget kombineres sårbarhetene i språkmodellen med tradisjonelle programvaresårbarheter og nye angrepsflater som oppstår når modellen kobles til data, verktøy og andre systemer.

Språkmodeller finnes i alt fra enkle chatteprogrammer, til agenter med stor handlingsevne og autonomi. I et chatteprogram er det mennesket som vurderer resultatet og eventuelt handler på bakgrunn av det. I et fullt agentisk system bruker språkmodellen selv verktøy, planlegger oppgaver og velger hvordan de skal gjennomføres. Jo større handlingsevne og autonomi systemet har, desto større kan konsekvensene av feil og angrep bli.

Denne kombinasjonen med andre komponenter kan medføre at sårbarhetene i språkmodellen reduseres men også at de forsterkes. Applikasjonslaget innfører også nye sårbarheter og trusler: kommunikasjon mellom menneske og agent; leverandørkjeder; misbruk av verktøy og uautoriserte handlinger. Tilgang til sensitiv informasjon, upålitelige eksterne data og mulighet til å kommunisere ut av systemet utgjør en spesielt krevende kombinasjon. Konsekvensen kan være brudd på både konfidensialitet (f.eks. datalekkasje), integritet (f.eks. uønsket sletting av filer) og tilgjengelighet (f.eks. ved at tunge prosesser kjøres av agenten), og kan både være en tilsiktet handling, via for eksempel et injeksjonsangrep, eller utilsiktet grunnet iboende upålitelighet til språkmodellen. Agentsikkerhet må derfor behandles som et systemproblem, ikke bare som et spørsmål om å gjøre språkmodellen sikrere.

> *Det finnes ikke ett enkelt tiltak som løser disse problemene. Sikkerheten må bygges i flere lag med tiltak etter prinsippet om dybdeforsvar. Det bør baseres på en risikovurdering, hvor man må forstå både verdier som skal beskyttes, hvilke trusler som er aktuelle og hvilke sårbarheter de kan utnytte for å skade verdiene.*

Tiltak for å redusere risiko

Det finnes ikke ett enkelt tiltak som løser disse problemene. Sikkerheten må bygges i flere lag med tiltak etter prinsippet om **dybdeforsvar**. Det bør baseres på en risikovurdering, hvor man må forstå både verdier som skal beskyttes, hvilke trusler som er aktuelle og hvilke sårbarheter de kan utnytte for å skade verdiene.

Sikkerhetstesting er nødvendig både for å kartlegge modellens kapabiliteter og for å undersøke relevante sårbarheter. Testing kan imidlertid vise at en svakhet finnes, ikke garantere at systemet er sikkert. Evalueringen må derfor knyttes til den konkrete modellen, applikasjonen og måten den skal brukes på, og gjentas når modell eller konfigurasjon endres.

Menneskelig kontroll kan redusere risiko, men må brukes målrettet og med omhu. Dersom brukeren stadig må godkjenne handlinger, kan resultatet bli godkjenningstretthet og mekanisk aksept. Mennesker gjør både feil og kan bli lurt. Kritiske og irreversible handlinger bør derfor prioriteres for menneskelig kontroll, men det bør ikke baseres alene på godkjenning av et menneske. Systemet ellers bør begrense hvilke handlinger agenten kan utføre.

Andre tiltak omfatter **isolasjon og sandkasser** som begrenser skadeomfanget; **identitets- og tilgangshåndtering** som gir agenter minst mulig privilegier for å kunne løse oppgaven; **informasjonsflytkontroll** som

SEAL

FFI har forsket på KI over lang tid, med flere aktiviteter på både sikkerhet og evaluering av KI-modeller og applikasjoner. Dette arbeidet har nå blitt samlet under en felles lab kalt **SEAL «Sikkerhet, Evaluering og Analyse (av KI) Lab»**. Målet med SEAL er både å samle kreftene, og å ha et felles fokus og en satsing på dette området. Her jobbes det tett med spesielt Forsvarets senter for data og KI. Det fokuseres spesielt på evalueringer av kapabiliteter og sikkerhetsaspekter av store språkmodeller for relevante bruksområder for nasjonal sikkerhet og spesielt Forsvaret. Dette for å kunne gi råd om tekniske krav og rammeverk, noe rapporten underbygger. Det jobbes også med andre problemstillinger som hvordan lage sikre agentiske systemer.

En språkmodell bør ikke behandles som en komponent man uten videre kan stole på, men som en upålitelig del av systemet som kan gjøre feil, manipuleres eller handle på uventede måter.

begrenser hvor data og kontroll kan flyte; **guardrails** som kontrollerer input og output; og **overvåking og deteksjon** for å oppdage uønsket oppførsel i drift. Man bør også vurdere bruk av **nulltillitsprinsippet** *(zero trust)* og på sikt vurdere mer **formelle metoder** for enda sterkere sikkerhetsgarantier.

Anbefalinger

En språkmodell bør ikke behandles som en komponent man uten videre kan stole på, men som en upålitelig del av systemet som kan gjøre feil, manipuleres eller handle på uventede måter. Sikkerheten må derfor i stor grad bygges rundt modellen.

Virksomheter som tar i bruk LLM-baserte løsninger bør:

1. gjennomføre en konkret risiko- og trusselvurdering for bruksområdet og gjøre vurdering av tiltak basert på det
2. teste modell og applikasjon mot relevante sårbarheter før produksjonssetting og ved vesentlige endringer
3. behandle agenten som en innsidetrussel og legge til grunn at feil eller kompromittering kan skje
4. følge prinsippet om minste handlerom og gi agenter bare de dataene, verktøyene og rettighetene som er nødvendige
5. begrense autonomi og kreve menneskelig kontroll der konsekvensene av feil eller uønskede handlinger er store. Dette må kombineres med andre tiltak.
6. benytte flere uavhengige sikkerhetsmekanismer, blant annet tilgangskontroll, isolasjon, informasjonsflytkontroll og guardrails
7. vurdere sikkerheten på nytt når bruksområdet utvides, slik at en løsning ikke gradvis tas i bruk til nye formål uten at risikoen vurderes på nytt

Målet er et forsvarlig sikkerhetsnivå, der sikkerhetsrisikoen ved bruk av språkmodeller er forstått og håndtert, og veid mot den operative effekten teknologien kan gi. Høyere risiko kan være forsvarlig dersom gevinsten er stor nok og risikoen er tilstrekkelig forstått og håndtert.

Hvor kan du lese rapporten?
Rapporten vil bli tilgjengelig på FFI sine nettsider. Meld deg på FFI sitt nyhetsbrev for å få mail når rapporten er tilgjengelig.

Forfattere:
Johan Vik Mathisen, Gudmund Grov

Har du spørsmål, ta kontakt med:

Gudmund Grov, sjefsforsker
gudmund.grov@ffi.no

Kommunikasjonsenheten ved FFI
info@ffi.no

Mer informasjon om FFI og forskningen vår finner du på ffi.no

Oktober 2026