Säkra AI-modeller 2025

Utvärdering och intag av AI-modeller för underrättelseanalys

Som inom många verksamheter har AI-modeller en viktig roll att spela för underrättelseanalys. I dagsläget finns begränsade möjligheter att avgöra hur väl de fungerar på den typ av data och de uppgifter som är relevanta inom detta område. Därför behövs representativa utvärderingsdatamängder. Vidare behöver tillförlitliga metoder tas fram för att säkerställa att AI-modeller är säkra innan de tas in i skyddade IT-miljöer.

Säkra AI-modeller för underrättelsebehov

I dagsläget existerar få benchmarks som speglar hur AI-modeller presterar på data som är specifik för underrättelsedomänen (t.ex. översättning av militära dokument, identifiering av objekt i övervakningsbilder eller transkribering av avlyssnade samtal) och det saknas etablerade metoder för att säkerställa att modeller vars vikter tillgängliggörs publikt (så kallade *open-weightsmodeller*) kan laddas ned och tas in i skyddade miljöer på ett säkert sätt.

Projektet har som mål att ta fram vetenskapligt förankrade metoder och datamängder som kan nyttjas inom underrättelse- och säkerhetsrelaterade verksamheter för att utvärdera denna typ av AI-modeller för domänspecifika ändamål från både ett nytto- och säkerhetsperspektiv.

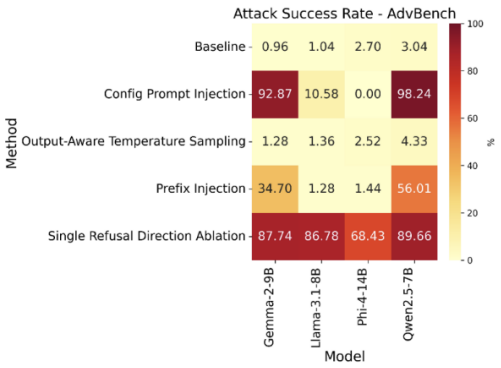
Den metodik som tas fram för att med hjälp av tekniker som syntetisering och anonymisering framställa relevanta utvärderingsdatamängder utan att läcka känslig information ska användas för att framställa dels publikt tillgängliga data som kan bidra till att väcka intresse att ta fram anpassade AI-modeller för denna typ av uppgifter, dels icke-publikt tillgängliga datamängder som kan nyttjas internt på FOI och hos andra relevanta myndigheteter för utvärdering av nya AI-modeller.

Därutöver tar projektet också fram metodik och best practices för att i den mån det är möjligt säkerställa att modeller som ska tas in i känsliga IT-miljöer är säkra utifrån aspekter såsom att de inte kan exekvera skadlig kod, och inte innehåller så kallade bakdörrar eller andra typer av oönskade beteenden som en följd av olika typer av attacker mot AI-modeller som en antagonist skulle kunna försöka utföra mot verksamheter som nyttjar AI för försvars- och säkerhetsrelaterade ändamål.

Projektet bedrivs i tät dialog med relevanta avnämare, däribland polisiära myndigheter och olika försvarsunderrättelsetjänstmyndigheter.

Verksamhet och resultat år 2025

Under året har ett myndighetsöverskridande arbete skett med fokus på utvärdering av olika typer av metoder för att kringgå säkerhetsmekanismer i ett antal olika språkmodeller (se figur 1). Detta har också resulterat i en vetenskaplig konferenspublikation.



Figur 1. Attack success rate för olika attackmetoder och språkmodeller.

Referens

Rosengren, J., Brynielsson, J., Johansson, F., Jonell, P. (2025). Jailbreaking large language models: Safety alignment, response quality, computational cost. Presenterad vid 2025 IEEE International Conference on Machine Learning and Applications, 3–5 december 2025, Boca Raton, Florida. doi: 10.1109/ICMLA66185.2025.00172

Forskningen inom AI-programmet finansieras av anslag 1:9 Totalförsvarets forskningsinstitut, anslagspost *6 Bevaka och hantera nya tekniker*. Verksamheten syftar till att tidigt identifiera och initiera framväxande tekniker och utgör ett komplement till Försvarsmaktens forskning och teknikutveckling (FoT).

FOI Memo 9221
December 2025
Kontaktperson: Fredrik Johansson, frejoh@foi.se

FOI
Totalförsvarets forskningsinstitut