|  |
| --- |
| **Rapport**Risiko- og sårbarhetsvurdering av gpt.uio.no |
|  |

**Dato for gjennomføring: 29.09.2023**

|  |
| --- |
|  Til den det måtte angå  IT-avdelingen, UiODato: 29.09.2023Gjelder: GPT UiO |

|  |
| --- |
|  |

Innholdsfortegnelse

[Innledning 3](#_Toc147226340)

[Bakgrunn 3](#_Toc147226341)

[Gjennomføring 4](#_Toc147226342)

[Del 1: Metode 4](#_Toc147226343)

[Arbeidsmøte 4](#_Toc147226344)

[Risikorapport 5](#_Toc147226345)

[Scoring 5](#_Toc147226346)

[Del 2: Hovedfunn – oversikt 7](#_Toc147226347)

[Alle hendelser 7](#_Toc147226348)

[Hovedinntrykk og overordnede risikoområder 7](#_Toc147226349)

[Del 3: Hendelser med medium risiko og anbefalte tiltak 9](#_Toc147226350)

[Vedlegg 1: Risikomatrise UiO GPT 10](#_Toc147226351)

[Vedlegg 2: Utfylt regneark som ble benyttet i risikovurderingen 11](#_Toc147226352)

[Vedlegg 3: Innledende ROS i forkant av utvikling 12](#_Toc147226353)

[Vedlegg 4: Eksempel sannsynlighet og konsekvens 13](#_Toc147226354)

|  |
| --- |
| **Deltagere** |
| **Navn** | **Rolle/funksjon** | **Institusjon** |
|  |  |  |
| Dagfinn Bergsager  | Seksjonssjef - Webutvikling |  IT-avdelingen, UiO |
| Lars Lauvstad Sættem | Senioringeniør - Apputvikling |  IT-avdelingen, UiO |
| Magnus Alderslyst Nygaard | Senioringeniør - Apputvikling |  IT-avdelingen, UiO |
| Pål Fugelli  | Seksjonssjef - Apputvikling |  IT-avdelingen, UiO |
| Isak Falch Alsos | Rådgiver - IT-sikkerhet og -jus | IT-avdelingen, UiO |
| Ernesto A. G. Benitez | Senioringeniør - Digitale kommunikasjonsplattformer | IT-avdelingen, UiO |

# Innledning

## Bakgrunn

Risiko- og sårbarhetsanalysen er tilrettelagt av IT-avdelingen, Universitetet i Oslo.

Rapport og vedlagt regneark er utformet etter mal fra UNINETT.

GPT UiO er en spesialtilpasset tjeneste ved Universitetet i Oslo som gjør OpenAIs ChatGPT tilgjengelig for studenter og ansatte via FEIDE-innlogging. Tjenesten brukes til ulike formål, inkludert akademiske prosjekter og forskning som involverer persondata opp til det som er klassifisert som gule data. Teknologisk er denne tjenesten basert på Generative Pre-trained Transformer (GPT), en avansert form for kunstig intelligens utviklet av OpenAI. GPT UiO er designet for samtale med et naturlig språk og har evnen til å generere dynamiske svar med informasjon over et vidt spekter av fagfelt. Det er også mulig for brukere å laste ned hele samtalen i en tekstfil.

Web-applikasjonen kjører på universitetets egen infrastruktur, med fysiske servere plassert i UiOs maskinpark på Blindern. Data lagres kun lokalt ved UiO, og Feide-brukernavnene er ikke eksponert for modellen. Tjenesten integrerer med servere i Europa via Microsoft Azure.

##

## Gjennomføring

Risiko- og sårbarhetsvurderingen av GPT UiO ble gjennomført som et arbeidsmøte (workshop) med deltakere fra IT-avdelingen.

Det ble også gjennomført en innledende ROS på systemnivå i forkant av utviklingen av web-applikasjonen (vedlegg 3).

Arbeidsmøtet ble avgrenset til supplerende ROS av web-applikasjonen gpt.uio.no og integrasjonen med Microsoft Azure

# Del 1: Metode

#

## Arbeidsmøte

Arbeidsmøte ble gjennomført på følgende måte:

1. Kort innføring og drøfting av risikovurderingsmetodikken.
2. Felles diskusjon av mulige uønskede hendelser som kan føre til at personopplysninger som behandles i UiO GPT-applikasjonen utsettes for uautorisert eksponering/tilgang, endring, sletting, tap eller utilgjengelighet.
3. Registrering og gjennomgang av slike uønskede hendelser i regneark, se vedlagt dokument til rapporten.
4. Registrering av informasjon om hvorfor hendelsene kan oppstå (sårbarheter/svakheter) og eventuelle sikringstiltak som allerede er etablert for å redusere risikoen for hendelsene (eksisterende beskyttelses- og kontrolltiltak).
5. Felles vurdering av risikoverdien – sannsynlighet + konsekvens – for hver enkelt uønsket hendelse.
6. Verdiene for sannsynlighet og konsekvens ble vurdert med utgangspunkt i en skala fra 1 (svært lite sannsynlig/alvorlig) til 4 (svært sannsynlig/alvorlig).
7. Utarbeidelse av forslag til sikringstiltak

Punktene 3-7 er oppsummert i vedlagte regneark til rapporten.

Etter arbeidsmøtet ble regneark sendt ut til deltakerne for kommentarer eller justeringer.

## Risikorapport

Den videre fremstillingen er delt inn i to hoveddeler. Første del gir en oversikt over hovedfunn i risiko- og sårbarhetsvurderingen. Deretter følger en nærmere gjennomgang over hendelser og anbefalte tiltak.

For hver av disse hendelsene spesifiseres ulike forbedringstiltak som anbefales iverksatt. Noen av disse ble foreslått og diskutert under arbeidsmøtene. Det er opp til institusjonen ved risikoeier å beslutte hvilke anbefalte tiltak som eventuelt skal iverksettes.

## Scoring

I denne rapporten scores risiko på følgende måte:

Sannsynlighet + (Konsekvens x 1,1)
Eksempel:

1s + (1k x 1,1) = 2 risikoverdi (lavest)
4s + (4k x 1,1) = 8 risikoverdi (maksimum)

Begge scores på en skala fra 1 til 4 hvor følgende tilnærming benyttes
Sannsynlighet:
1 er «lav» mens 4 er «svært høy» sannsynlig for at det kan skje

Konsekvens:
1 er «liten» negativ konsekvens for enkeltpersoner og/eller institusjon mens 4 er «svært høy» med store konsekvenser for enkeltpersoner og/eller institusjon

# Del 2: Hovedfunn – oversikt

#

## Alle hendelser

Risiko- og sårbarhetsvurderingen av GPT UiO avdekket totalt 8 uønskede hendelser. Dette er hendelser som kan føre til brudd på informasjonssikkerheten i opplysningene hos institusjonen (Se Vedlegg 1 regneark og Vedlegg 2: Risikomatrise)

**Hendelser med høy risiko (risikoverdi 6 eller høyere)**

3 av de 8 uønskede hendelser ble vurdert å ha høy risiko (hendelser som deltakerne ga risikoverdi 6 eller høyere) under arbeidsmøtet.

**Hendelser med middels høy risiko (risikoverdi 4-5)**

4 av 8 uønskede hendelser ble vurdert til å ha risikoverdien 4-5.

Det anbefales at institusjonen aksepterer risikoen for de øvrige hendelsene som ble avdekket og diskutert på arbeidsmøtene, fordi risikoen for disse hendelsene ble vurdert som lav eller relativt lav. Det innebærer at det ikke foreslås at tiltak iverksettes for å redusere sannsynligheten for eller konsekvensene av disse hendelsene.

## Hovedinntrykk og overordnede risikoområder

Etter å ha gjennomført arbeidsmøtene og evaluert hendelsene observerer vi at 3 av hendelsene som ble vurdert er utenfor akseptabelt risikonivå (rød, risikoverdi 5 eller høyere). Vi vil nedenfor presentere tiltak for på rødt nivå. Hovedfunn er som følger

* **Teknologien bak generativ KI i form av en språkmodell kan være vanskelig å forstå;** Det er risiko for at GPT UiO blir anvendt som en faktamodell i feil kontekst
* **Data kommer på avveie som følge av bruker-feil;** Det er enkelt å laste ned og dele samtaler.
* **Brukere velger alternative tjenester;** Manglende funksjonalitet og strenge innholds-filter i GPT UiO kan føre til bruk av alternative systemer hvor personvern ikke er tilstrekkelig ivaretatt

Tiltak presenteres nedenfor

# Del 3: Hendelser med medium risiko og anbefalte tiltak

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Risikoelement/hendelse 6** | **Sårbarhet** | **Eksisterende tiltak** | **S** | **K** | **Risiko** |
| Gule eller røde data kommer på avveie.  | Bruker kan laster ned hele samtalen som fil. Lagres og deles på ikke-godkjente systemer | Generell informasjon |  3 | 3 | 6 |
| **Tiltak** | **Tiltakskategori: Teknisk / administrativt** |
| Tiltak anbefales for å forebygge hendelser:* Tilstrekkelig informasjon presenteres i forbindelse med nedlasting av samtale. For eksempel en dialog eller popup når laster ned fil
 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Risikoelement/hendelse 7** | **Sårbarhet** | **Eksisterende tiltak** | **S** | **K** | **Risiko** |
| Studenter og ansatte bruker ekstern tjeneste for samtalebasert generativ AI og legger inn personopplysninger der | Det er et for strengt filter på UiOs versjon. Ønsker å bruke funksjonalitet i annen løsning Bruker velger andre GPT-løsninger hvor personvern ikke er tilstrekkelig ivaretatt | Ingen | 4 | 2 | 6 |
| **Tiltak** | **Tiltakskategori: Teknisk / administrativt** |
| Tiltak anbefales for å forebygge hendelser:* Justere og optimalisere innholdsfiltere i tjenesten
* Opprettholde høy kvalitet på funksjoner i tjenesten
 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Risikoelement/hendelse 8** | **Sårbarhet** | **Eksisterende tiltak** | **S** | **K** | **Risiko** |
| Brukere forholder seg til GPT som en fakta-modell og anvender den i feil kontekst | Teknisk og konseptuelt komplisert teknologi. | Ingen | 4 | 2 | 6 |
| **Tiltak** | **Tiltakskategori: Teknisk / administrativt** |
| Tiltak anbefales for å forebygge hendelser:* Brukeropplæring som endel av onboardingen til tjenesten. Formidle lett forståelig / god informasjon om begrensninger.
* Justere bunntekst i applikasjonen med lenke til informasjon som er synlig / tilgjengelig hele tiden
 |

# Vedlegg 1: Risikomatrise UiO GPT

Hver av hendelsene som ble identifisert på arbeidsmøtet er nummerert og puttet inn i risikomatrisen under. (se vedlegg 2 for detaljer)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  **Konsekvens** | 4Sværthøy | 2 |  |  |  |
| 3Høy | 1 |  | 6 | 8 |
| 2Moderat |  |  | 3 | 7 |
| 1Liten |  | 5 |  | 4 |
|  | 1Lav | 2Moderat | 3Høy | 4Svært høy |
|  |  **Sannsynlighet** |

# Vedlegg 2: Utfylt regneark som ble benyttet i risikovurderingen

# Vedlegg 3: Innledende ROS i forkant av utvikling

# Vedlegg 4: Eksempel sannsynlighet og konsekvens

|  |  |
| --- | --- |
| **Konsekvens** | **Sannsynlighet** |
| **1 – Svært lite alvorlig**Har ubetydelige skadevirkninger for enkeltpersoner eller institusjonen.  Eksempel: Mangelfull logging av tilgang til personopplysninger; uklart når ikke-kritisk informasjon sist ble oppdatert; noe tapt arbeidstid. | **1 – Svært lite sannsynlig**Hendelsen kan skje sjeldnere enn hvert femte år. |
| **2 – Mindre alvorlig**Har visse skadevirkninger for enkeltpersoner eller institusjonen.  Eksempel: Uautorisert eksponering av noen få personopplysninger; ufullstendig registrering av ikke-kritisk informasjon; informasjon er utilgjengelig opp til tre dager. | **2 – Lite sannsynlighet**Hendelsen kan skje sjeldnere enn hvert år, men oftere enn hvert femte år |
| **3 – Alvorlig**Har merkbare skadevirkninger for enkeltpersoner eller institusjonen. Eksempel: Uautorisert eksponering av sensitive eller større mengder alminnelige personopplysninger; viktig informasjon mangler eller er feil; informasjon er utilgjengelig opp til to uker.  | **3 – Sannsynlig**Hendelsen kan skje hvert år. |
| **4 – Svært alvorlig**Har store skadevirkninger for enkeltpersoner eller institusjonen.  Eksempel: Uautorisert eksponering av større mengder sensitive personopplysninger; kritisk informasjon mangler eller er feil; informasjon er utilgjengelig lengre enn to uker. | **4 – Svært sannsynlig**Hendelsen kan skje mange ganger i året. |