

GPT UiO for agentisk KI med kontroll på kostnader - inkludert

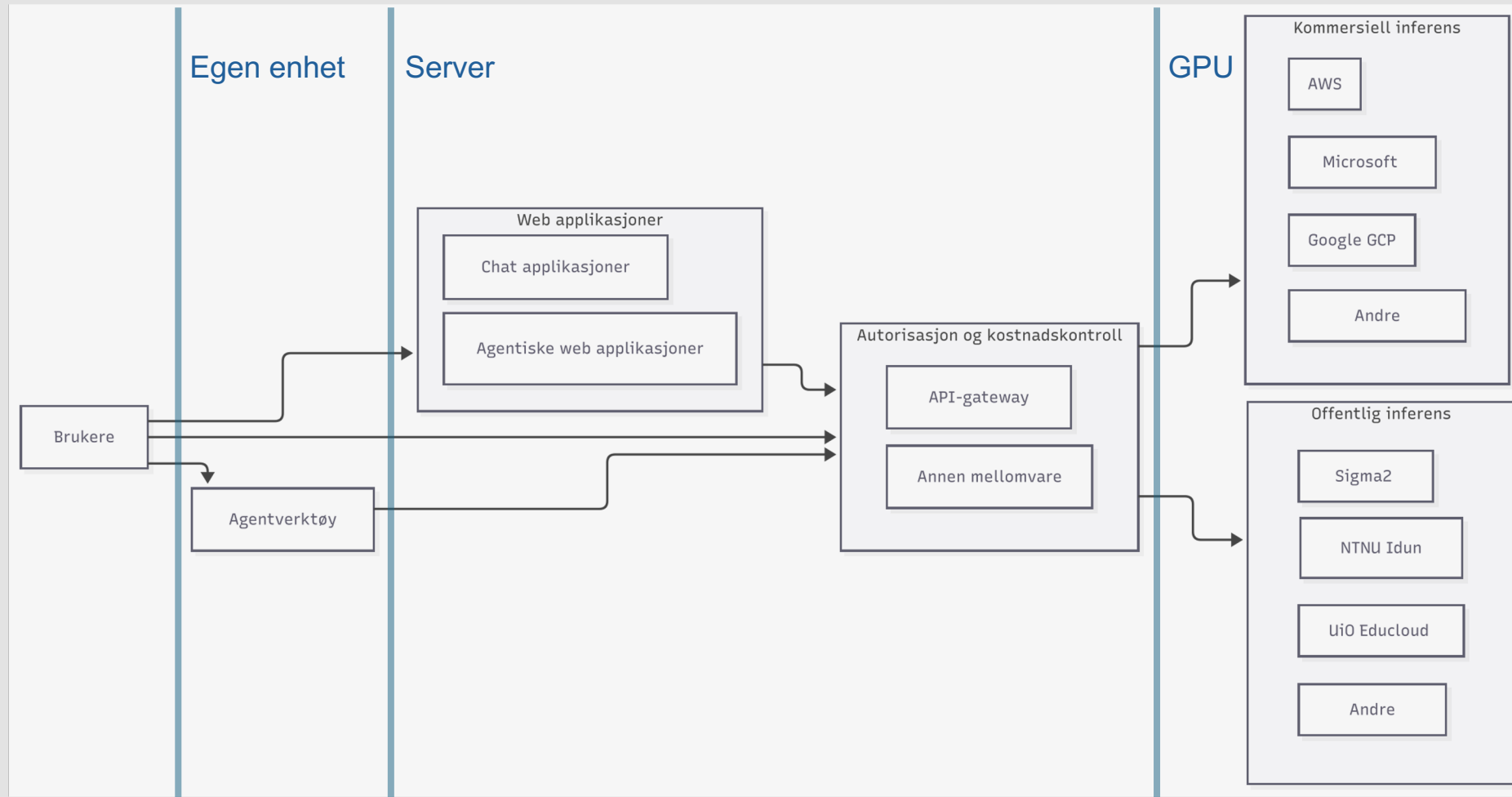


Oddmund Møgedal, UiO
Idémøte høsten 2026

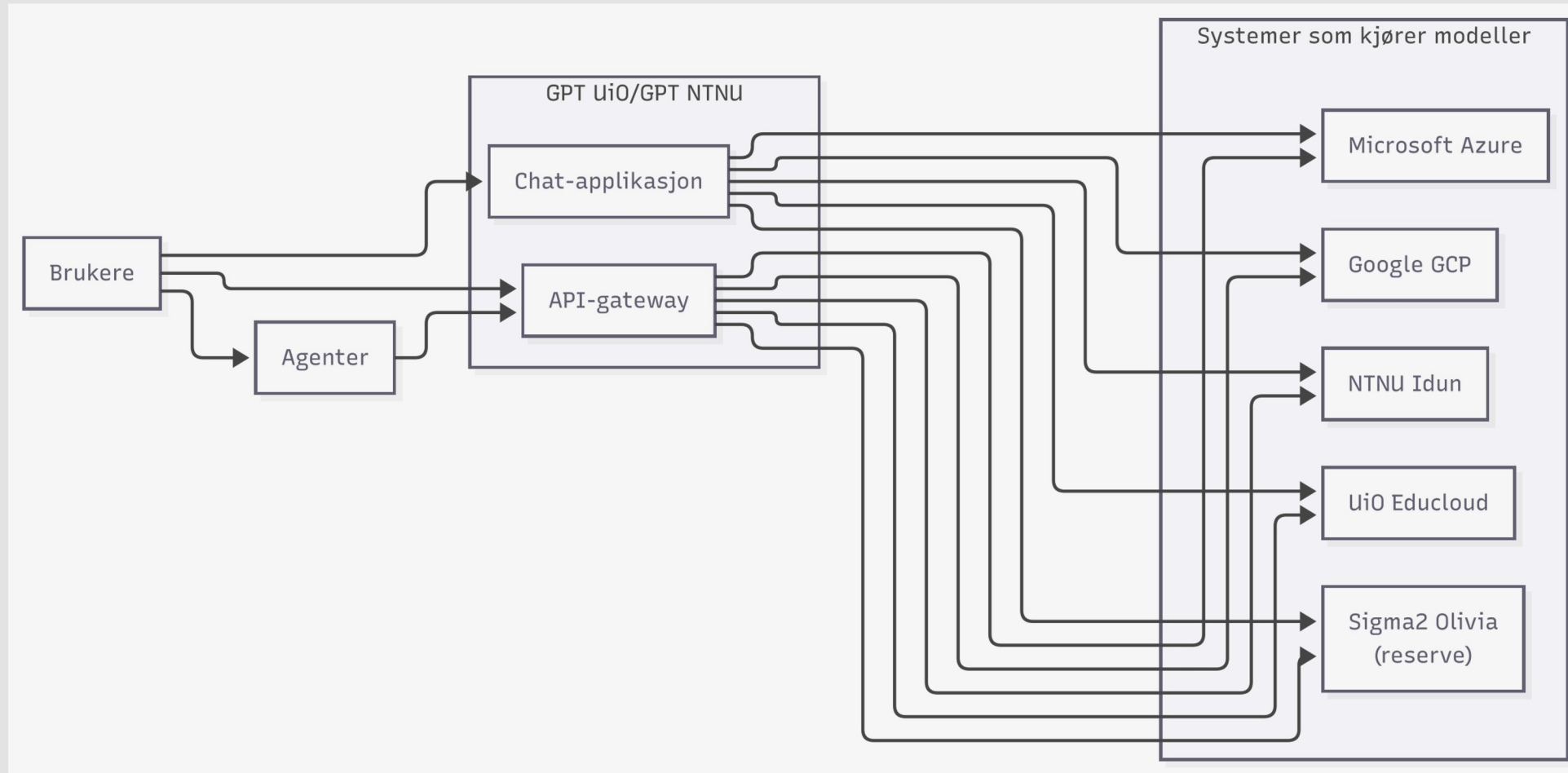


UNIVERSITETET
I OSLO

Forenklet: Overordnede KI-komponenter for LLM-bruk



GPT UiO: Samlet løsning for chat og API mot kommersielle og lokale LLM-er



Bakgrunn: GPT UiO og personvern

- Universitetet i Oslo (UiO) utviklet gpt.uio.no som en personvernsikker chat løsning for studenter og ansatte i 2023
 - Basert på GPT modeller fra OpenAI (hver institusjon bruker egen Azure)
 - Tilbydere (Microsoft, OpenAI) vet ikke hvem den enkelte sluttbruker er
 - Brukes av ca. 50 institusjoner i offentlig sektor, de fleste innen høyere utdanning
- Funksjonalitet
 - Først chat
 - Deling av KI-assistenter med egne instruksjoner, f.eks. i undervisning
 - Filopplasting og RAG

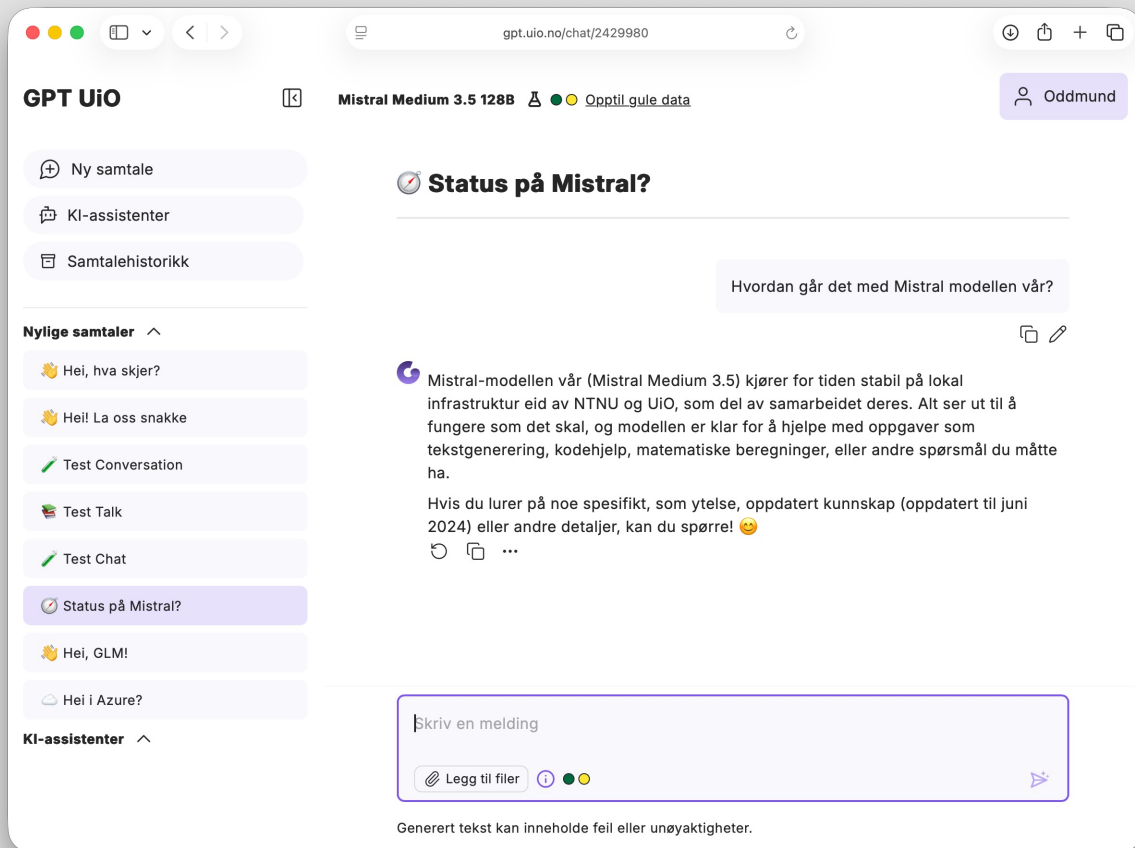


Samarbeid fra 2025: GPT UiO/GPT NTNU

- Lokale LLM-er med åpne vektorer (NTNU Idun og UiO Educloud Fox), i tillegg til kommersielle fra OpenAI
- Etter hvert et godt utvalg av modeller med åpne vektorer
 - De beste modellene med åpne vektorer er gode nok til det aller meste
 - Vi trenger å kommunisere bedre hvilke modeller som er gode nok til ulike formål
- Støtte for røde data for utvalgte modeller i egen infrastruktur
- Midlertidige samtaler
 - Ikke synlige når man går ut av samtalen
 - Innhold lagres likevel en stund blant annet pga. krav fra tilbydere



GPT UiO og GPT NTNU



GPT UiO

Mistral Medium 3.5 128B Opptil gule data Oddmund

Status på Mistral?

Hvordan går det med Mistral modellen vår?

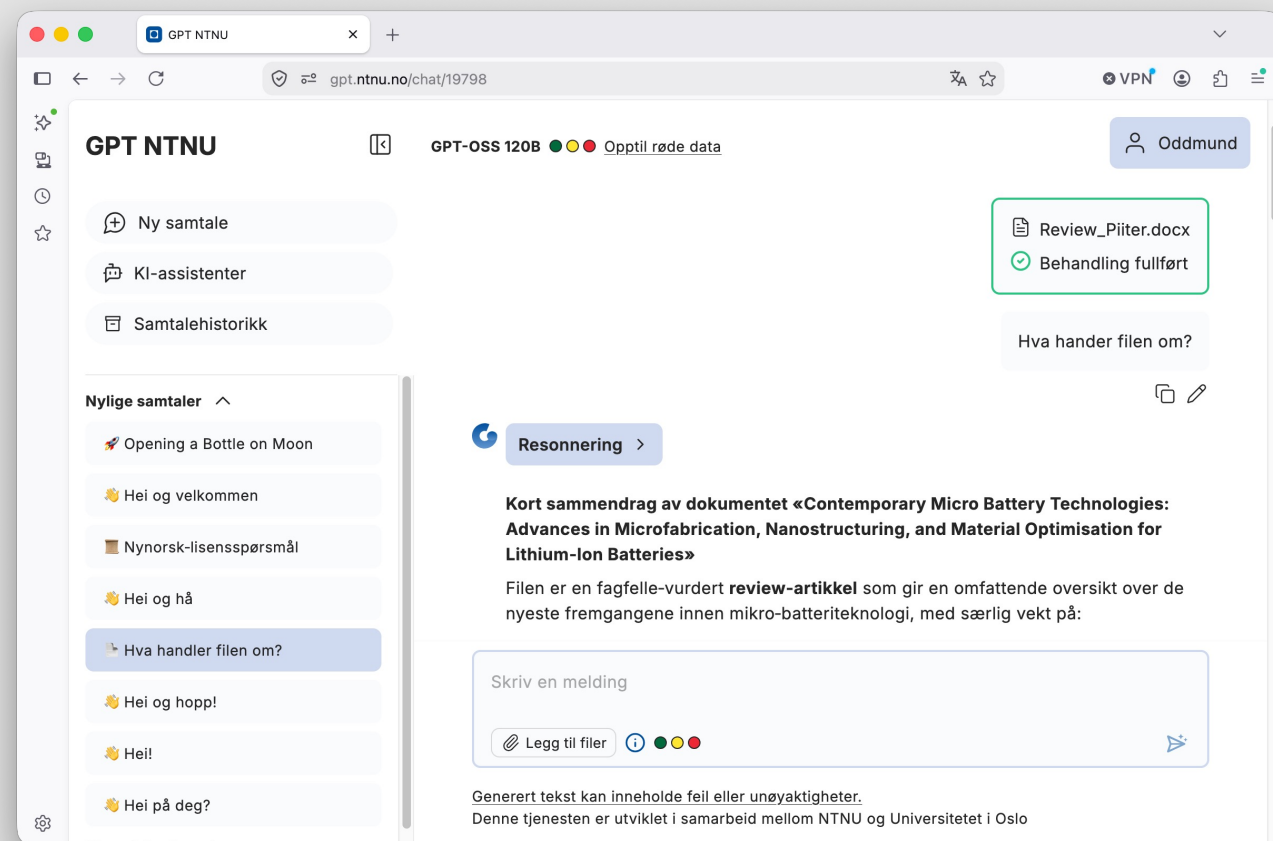
Mistral-modellen vår (Mistral Medium 3.5) kjører for tiden stabil på lokal infrastruktur eid av NTNU og UiO, som del av samarbeidet deres. Alt ser ut til å fungere som det skal, og modellen er klar for å hjelpe med oppgaver som tekstgenerering, kodehjelp, matematiske beregninger, eller andre spørsmål du måtte ha.

Hvis du lurar på noe spesifikt, som ytelse, oppdatert kunnskap (oppdatert til juni 2024) eller andre detaljer, kan du spørre! 😊

Skriv en melding

Legg til filer

Generert tekst kan inneholde feil eller unøyaktigheter.



GPT NTNU

GPT-OSS 120B Opptil røde data Oddmund

Review_Piiter.docx
Behandling fullført

Hva handler filen om?

Resonnering

Kort sammendrag av dokumentet «Contemporary Micro Battery Technologies: Advances in Microfabrication, Nanostructuring, and Material Optimisation for Lithium-Ion Batteries»

Filen er en fagfelle-vurdert **review-artikkel** som gir en omfattende oversikt over de nyeste fremgangene innen mikro-batteriteknologi, med særlig vekt på:

Skriv en melding

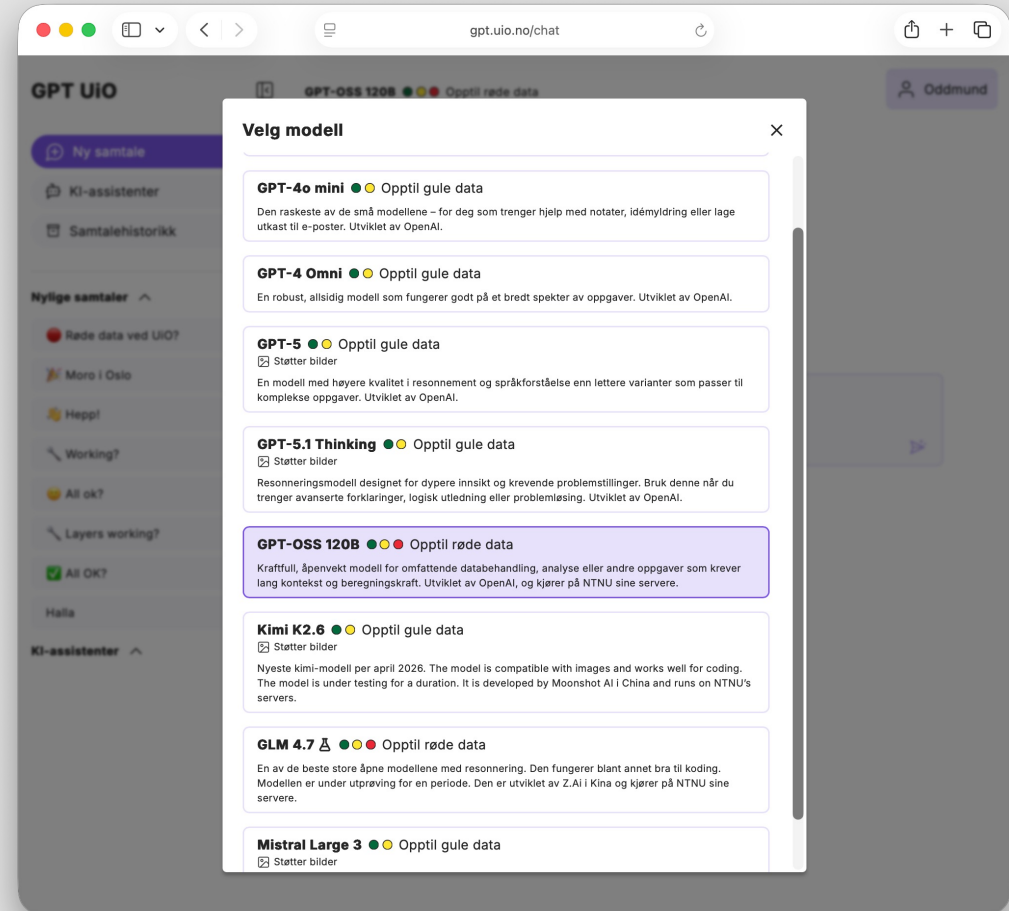
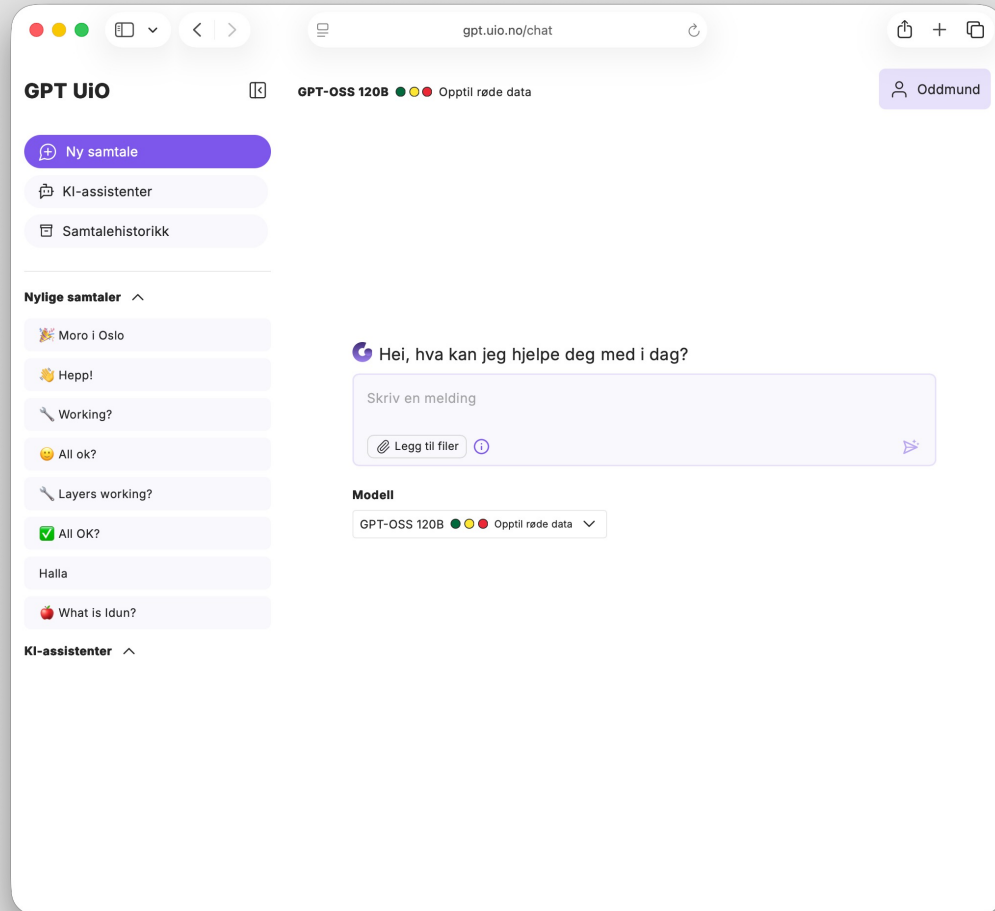
Legg til filer

Generert tekst kan inneholde feil eller unøyaktigheter.
Denne tjenesten er utviklet i samarbeid mellom NTNU og Universitetet i Oslo



UNIVERSITETET
I OSLO

Dataklassifisering og røde data



UNIVERSITETET
I OSLO

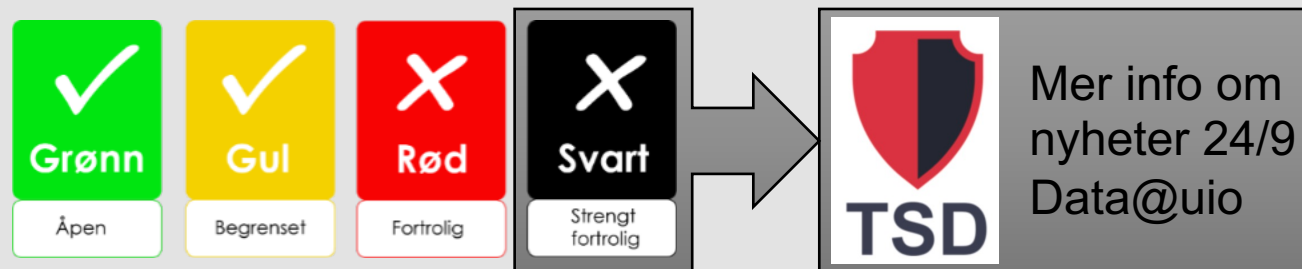
API-støtte for mange og etter hvert alle

- KI med API er ikke bare for utviklere
 - Støtte for ulike klientverktøy som kan bruke API-kall (OpenCode, Claude Code, Claude Desktop, Codex m.fl.)
 - Trenger regler og installasjonspakker på Windows for bruk av Claude Desktop og ChatGPT for Desktop
- API-støtte i GPT UiO legger grunnlaget for å tilby verktøy som krever API-tilgang til alle ansatte og studenter etter hvert
 - Nå tilgjengelig for ansatte som ber om det (trenger en gradvis økning)
 - Studenter trenger en begrunnelse knyttet til studier eller avtalt for emner
 - Bør forbedre behandling av ønsker om tilgang (leveringstid opp til en uke)
 - Lenke fra [Infoside om API-nøkler](#) (for øyeblikket e-post)

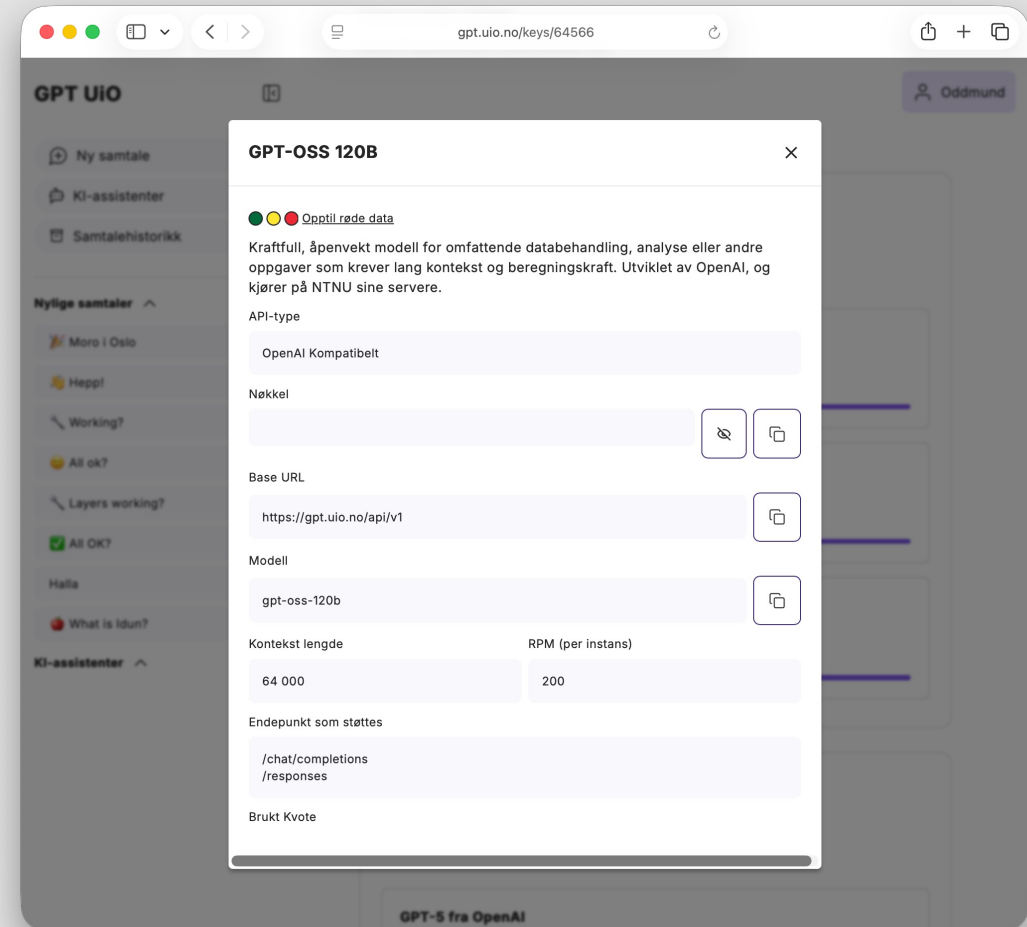
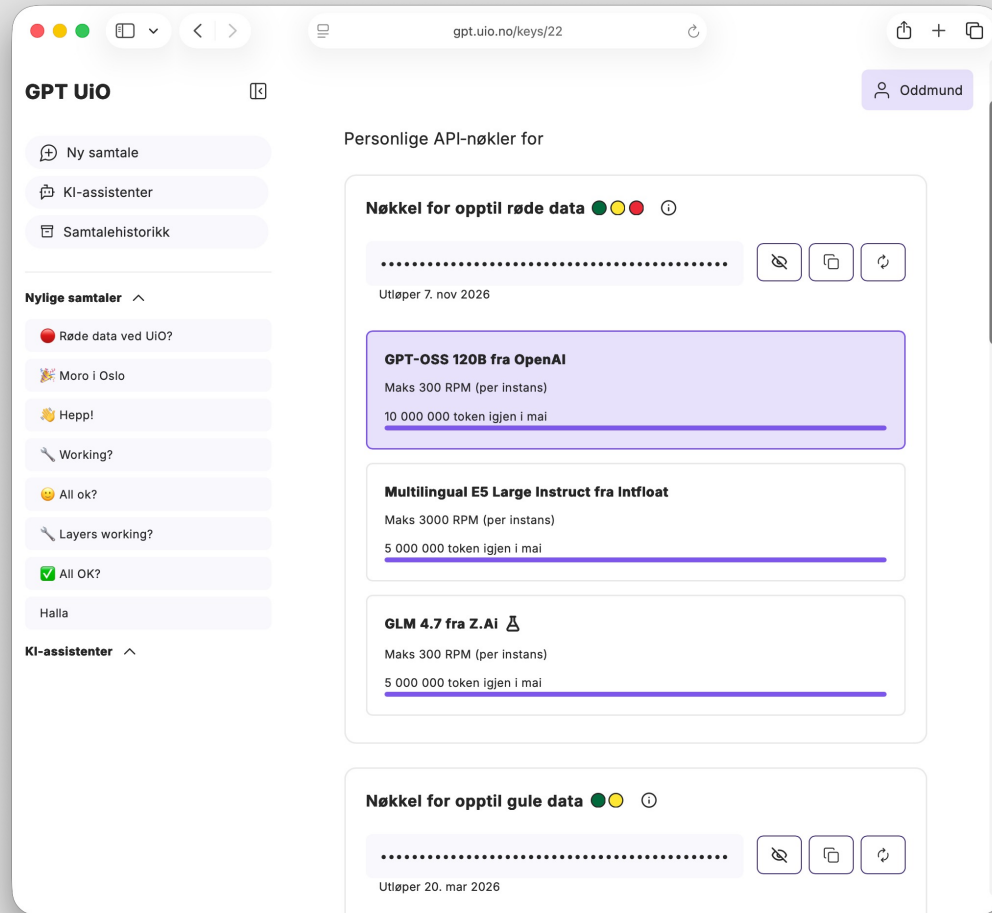


API-nøkler basert på dataklasse opp til rødt

- API-tilgang med personlige API-nøkler
 - Nøkler er personlige og uavhengige av den enkelte LLM-en
 - Separate API-nøkler basert på klassifisering av data (rød, gul og snart grønn nøkkel for hver bruker)
 - Brukere må selv velge riktig nøkkel som samsvarer med data som skal brukes/sendes
 - Månedlige kvoter for bruk per modell
- Mulighet for API-tjenestebrukere med større kvoter
 - Samme fargesystem som personlige API-nøkler



Personlige API-nøkler for hver dataklasse



UNIVERSITETET
I OSLO

Kommer snart:

Claude og det beste fra OpenAI via API med kvoter og kostnadskontroll

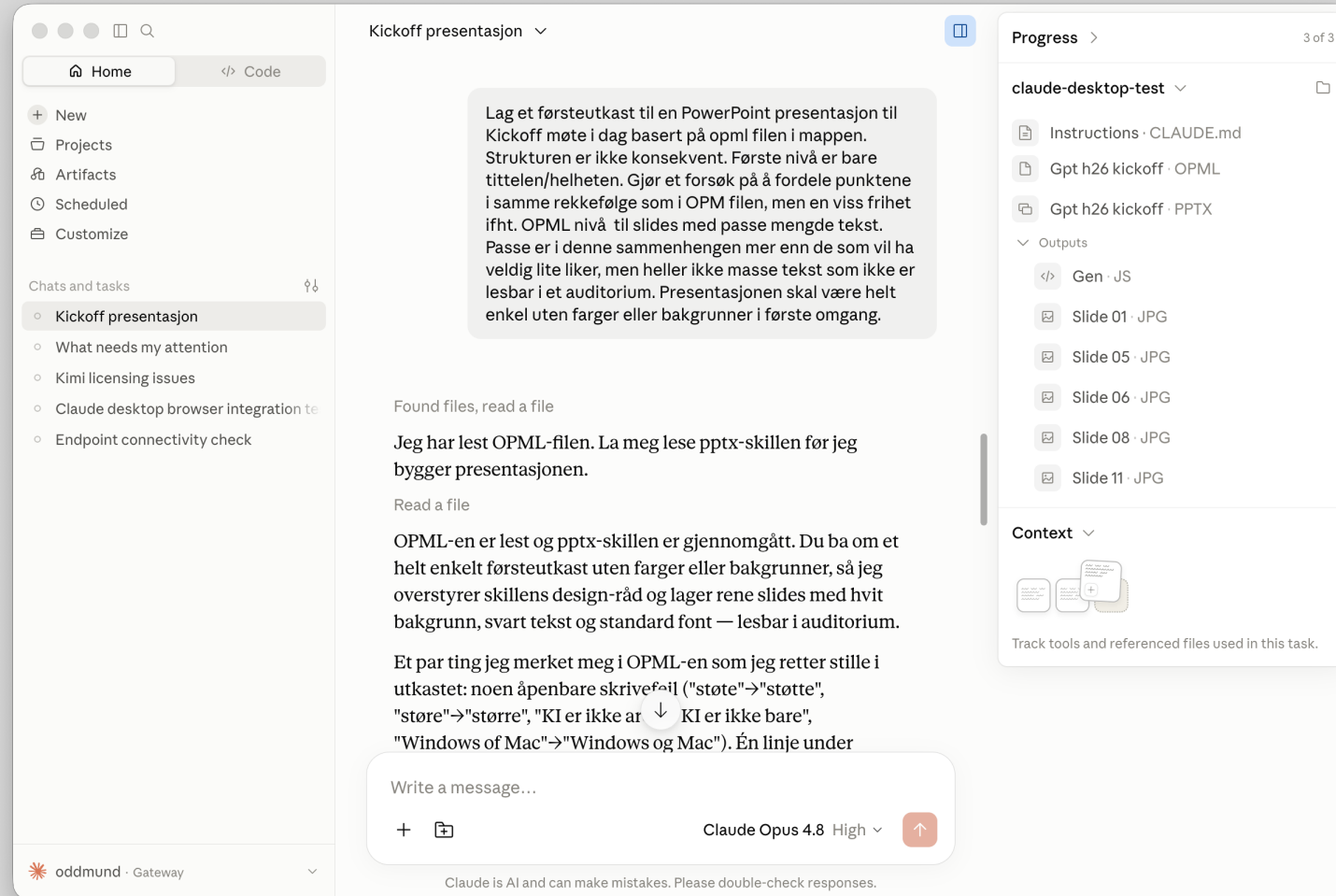


- En **organisasjonsenhet** kan autorisere at en bruker får en kvote, og **betale for faktisk bruk** senere
- **På tvers av leverandører** – med minst mulig binding til enkeltleverandører
 - Modeller i egen infrastruktur når det er hensiktsmessig
 - Claude og dyrere bruk av OpenAI-modeller når det trengs uten abonnement for hver bruker
- Kort pilot som starter i løpet av september.
- **Plan er å gjøre dette tilgjengelig med Claude/OpenAI (for de som noen vil betale for) i løpet av oktober** for modeller som er tilgjengelige i Europa



UNIVERSITETET
I OSLO

Claude Desktop som agentisk verktøy ...



UNIVERSITETET
I OSLO

koblet gjennom GPT UiO for kostnadskontroll ...

Claude

Configure third-party inference

Default Export

Search settings

Connection

Workspace

Connectors

Telemetry & updates

Limits

Appearance

Plugins

Egress

Source

Connection

Choose where Claude Desktop sends inference requests.

Gateway

GATEWAY CREDENTIALS

Test connection

Credential kind

Selects the credential source. When set, only that source is used (no fallback).

Static API key

Gateway base URL *

Full URL of the inference gateway endpoint.

https://gpt-dev.uio.no/api/claude/

Gateway API key *

.....

Gateway auth scheme

How the gateway credential is sent on the wire (Authorization: Bearer vs x-api-key header).

bearer

Artifact preview iframe origin

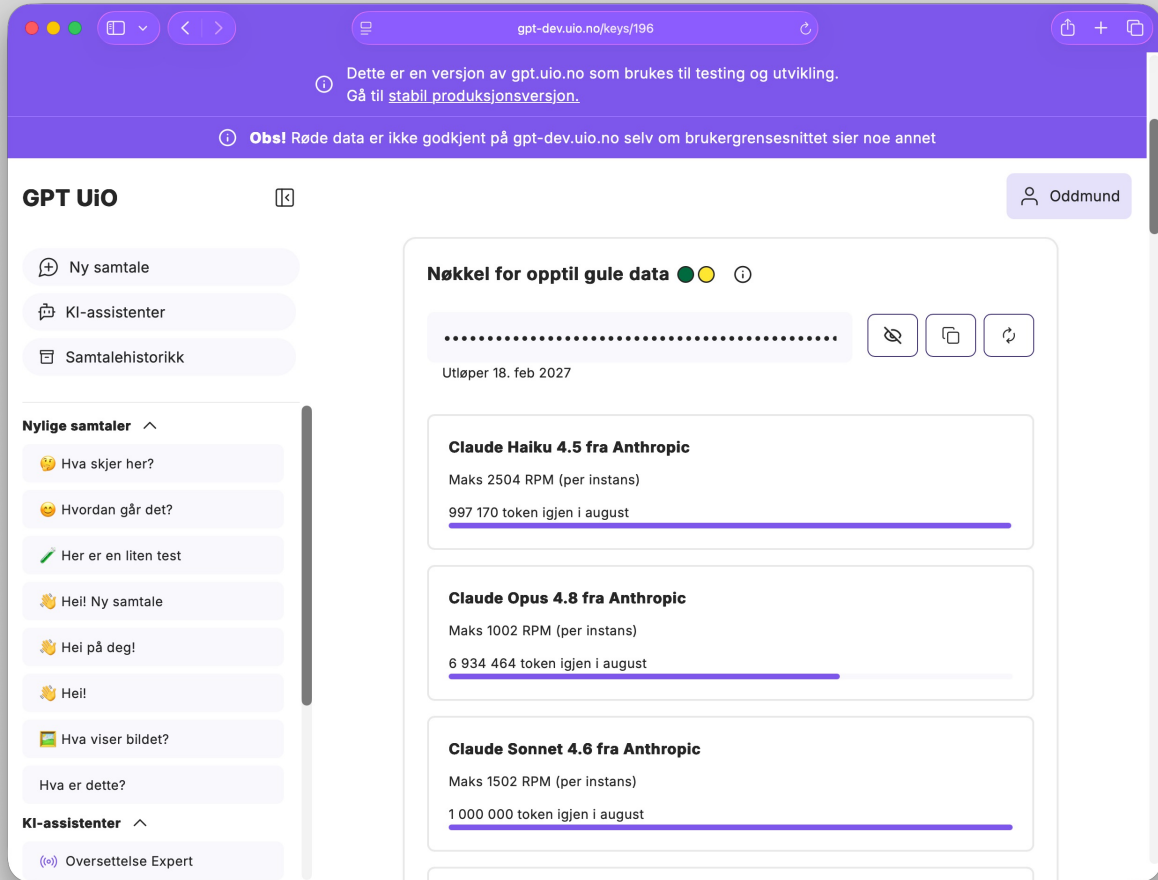
HTTPS origin of the user-content-renderer deployment used for artifact and file previews. Defaults to the workspace URL.

Apply Changes



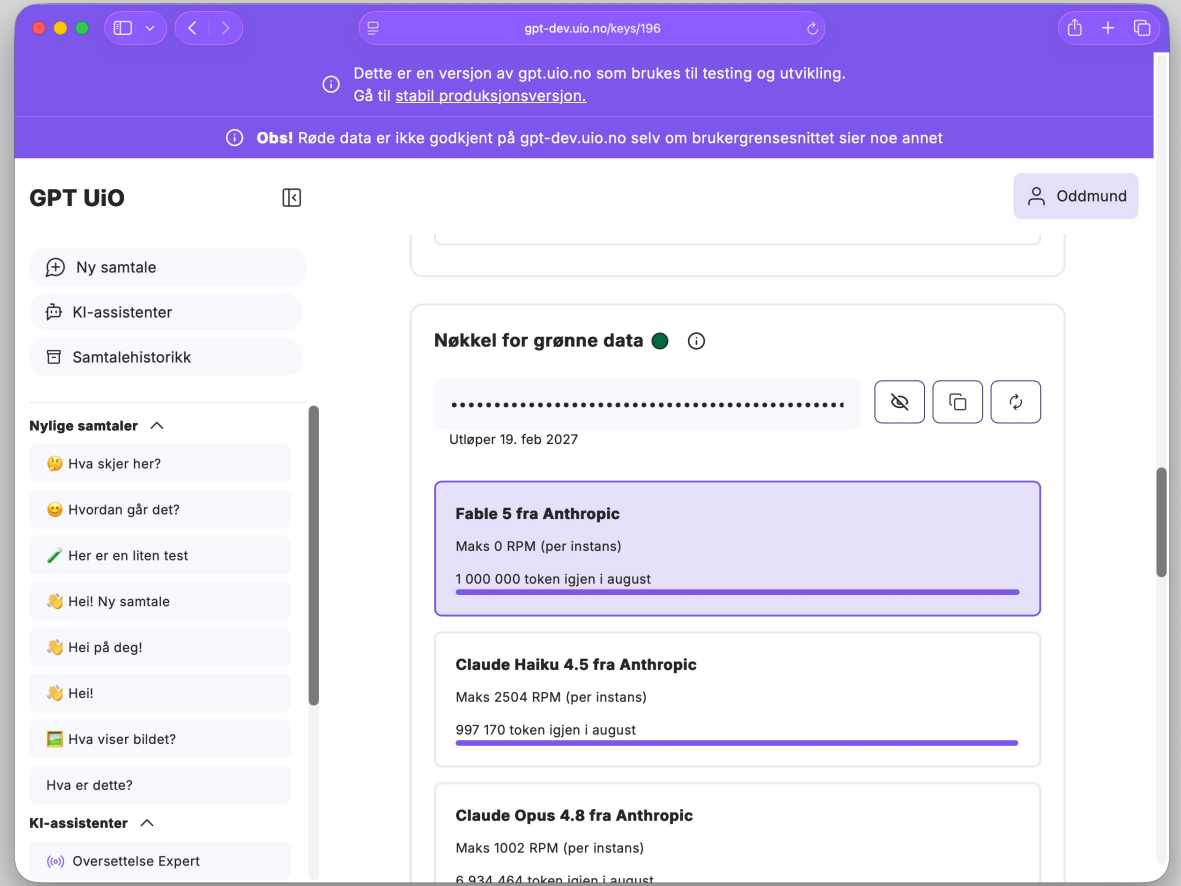
UNIVERSITETET
I OSLO

med personlige nøkler basert på dataklasse



The screenshot shows the GPT Uio web interface. The browser address bar displays `gpt-dev.uio.no/keys/196`. A purple banner at the top contains the text: "Dette er en versjon av gpt.uio.no som brukes til testing og utvikling. Gå til [stabil produksjonsversjon.](#)" and "Obs! Røde data er ikke godkjent på gpt-dev.uio.no selv om brukergrensesnittet sier noe annet". The user profile "Oddmund" is in the top right. On the left, there are buttons for "Ny samtale", "KI-assistenter", and "Samtalehistorikk", along with a list of "Nylige samtaler". The main content area is titled "Nøkkel for opptil gule data" with a red status indicator. It shows a redacted key, its expiration date "Utløper 18. feb 2027", and a list of models with their token usage:

Model	Max RPM	Usage
Claude Haiku 4.5 fra Anthropic	Maks 2504 RPM (per instans)	997 170 token igjen i august
Claude Opus 4.8 fra Anthropic	Maks 1002 RPM (per instans)	6 934 464 token igjen i august
Claude Sonnet 4.6 fra Anthropic	Maks 1502 RPM (per instans)	1 000 000 token igjen i august



The screenshot shows the GPT Uio web interface with a green key status. The browser address bar displays `gpt-dev.uio.no/keys/196`. The same purple banner and user profile "Oddmund" are present. The main content area is titled "Nøkkel for grønne data" with a green status indicator. It shows a green key, its expiration date "Utløper 19. feb 2027", and a list of models with their token usage:

Model	Max RPM	Usage
Fable 5 fra Anthropic	Maks 0 RPM (per instans)	1 000 000 token igjen i august
Claude Haiku 4.5 fra Anthropic	Maks 2504 RPM (per instans)	997 170 token igjen i august
Claude Opus 4.8 fra Anthropic	Maks 1002 RPM (per instans)	6 934 464 token igjen i august



Bruke LLM-er som er gode nok

- Kvoter er begrensende, men nødvendige
 - For å ha kontroll på kostnader med kommersielle modeller
 - For å kunne dele på modeller med åpne vektorer i egen infrastruktur
- Det blir nødvendig å bruke modeller som er gode nok, ikke alltid de beste
 - De beste modellene med åpne vektorer er bra nok til det meste, og ligger ikke så langt bak grensen for de beste ("frontier")
 - Tilbys nå på egen infrastruktur (UiO og NTNU)
 - Kan bli aktuelt å bruke disse også gjennom kommersielle tilbydere for å få nok kapasitet med god kvalitet til en fornuftig pris
 - Mindre modeller ligger bak de største men er bra nok til mye, og har ofte bedre forhold pris/ytelse
 - Modeller på egne enheter blir mer nyttige etter hvert som de blir gode nok til flere oppgaver



Fremtiden: Kostnader og kontroll

- LLM-er (tokens) vil være en begrenset ressurs i overskuelig fremtid
 - Kostnadene for bruk av kommersielle modeller er kanskje ikke reelle i forhold til investeringene som leverandørene gjør
 - Det vil aldri bli ubegrenset tilgang til de beste modellene
- Organisasjonen må forberede seg på kostnadene for Claude og lignende modeller
 - GPT UiO med API vil skalere etter hvert som bruk av KI øker og kostnadene blir (kanskje veldig) store
 - Pengene for bruk av de kommersielle modellene må komme fra et sted
 - Administrasjonen kommer til å følge med på kostnader
 - Det snakkes allerede i offentlig sektor om gevinstrealisering
 - Vesentlig økt bruk av modeller med åpne vektor må også finansieres



Hva kan det komme til å koste?



UNIVERSITETET
I OSLO

Takk for oppmerksomheten

- Veldig mange spennende muligheter med KI
- Kostnadene avhenger av bruksområder og modeller, og må håndteres
 - Kloge valg kan gi mer KI for mindre kostnad

