



id: SI_003

model: anthropic/claude-sonnet-4.6

seeds:

04_risiko: mulighet

03_relasjon: tillit

05_handling: implementer

01_perspektiv: økonomisk

02_tidhorisont: kortsiktig

Instans SI_003

Seeds: økonomisk | kortsiktig | tillit | mulighet | implementer

Kjerneperspektiv: Bygg tillit billig, deretter skaler

I helsevesen med 200 ansatte og streng datasikkerhet er den største risikoen ikke teknisk — det er **tillitsgapet** mellom IT-governance og klinisk/operasjonell ledelse. Implementeringsveien må derfor gi økonomiske bevis tidlig, ikke arkitektonisk eleganse.

Arkitekturvalg: Unngå “big bang”

Min dissens fra mainstream: Ikke start med Azure OpenAI + Python-orkestreringslag. Det er riktig endepunkt, men feil startpunkt for en cloud-only, strengt regulert organisasjon.

Kortsiktig (0-3 mnd): Copilot Studio + Power Automate

- Lever innenfor eksisterende Entra ID-grenser — ingen ny dataflyt til eksterne API-er
- Power Automate kan simulere enkel swarm via parallelle “Apply to Each”-brancher med ulike system-prompts
- **Kostnad:** Inkludert i M365 E3/E5-lisenser dere sannsynligvis allerede betaler
- **Begrensning:** Maks ~10-15 reelle parallelle agenter, ingen token-granulert observerbarhet

Mellomsteg (3-9 mnd): Azure OpenAI + Managed Identity



```
# Eksempel: Agent-spesialisering via role injection
agent_roles = {
    "regulatory": "Du er GDPR/Datatilsynet-revisor. Vurder kun compliance-
    risiko.",
    "clinical_ops": "Du er klinisk driftsanalytiker. Vurder kun pasientflyt-
    implikasjoner.",
    "cost": "Du er helseøkonom. Kvantifiser kun direkte
    budsjettsekvenser."
}
# Samme case-dokument, ulik system_prompt per agent
# Output aggregeres av én "synthesizer"-agent med eksplisitt vektingsregel
```

Data forblir i Azure-tenant via **on-behalf-of flow** mot SharePoint/Dataverse — ingen ekstern kopiering.

Case 2: Inbox-modell

Service Bus Queue → Azure Function trigger → parallelle OpenAI-kall med delt base-prompt + case-spesifikk kontekst. Consistency-sikring: **konfidensterskel** i aggregator avviser output under 0.7 og re-router til manuell behandling.

Cost-realitet

100 GPT-4o-agenter × 4000 tokens = ~\$6–8 per sverm-kjøring. Ved 10 kjøring/dag = ~\$2 000/mnd. Bruk GPT-4o-mini for screening-agenter, GPT-4o kun for synthesizer. Reduserer kostnad 60–70 %.

Konkrete neste steg

1. **Uke 1-2:** Bygg én 5-agent parallell analyse i Copilot Studio — velg ett reelt problem, mål kvalitetsgevinst vs. enkeltinstans
2. **Uke 4-6:** Sett opp Azure OpenAI-ressurs med Managed Identity, valider at SharePoint-data kan injiseres uten å forlate tenant
3. **Mnd 2:** Lever ROI-rapport til ledelse med faktiske token-kostnader og tidsmåling — dette er tillitsbyggingen som låser opp budsjett for fase 2