



# Debrief — anthropic/claude-sonnet-4.6

---

- Instanser: 9
  - Tokens inn: 21,070
  - Tokens ut: 8,415
  - Kostnad: \$0.1894
  - Kjørt: 2026-04-14 16:04:44.264011+00:00
- 

## Sverm-debrief

---

### Konsensus

1. **Azure OpenAI er eneste forsvarbare plattform** for helsedata. Alle 8 instanser avviste CrewAI/Anthropic-hosted og lignende. Norway East med Private Endpoint er ikke valgfritt.
  2. **Copilot Studio er feil verktøy for ekte sverm.** Power Platform-throttling (25 samtidige flows) og manglende token-granularitet gjør det uegnet for 100 parallelle agenter.
  3. **Rolle-differensiering krever ulik kontekst, ikke bare ulik prompt.** Regulatory reviewer får lovtekst-chunks; cost-analyser får budsjetttabeller. Samme dokument til alle 100 er bortkastet kapasitet.
  4. **Aggregering via strukturert JSON, ikke rå tekst.** Meta-agent leser strukturerte outputs — dette er revisjonsbart og operasjonelt håndterbart i helsevesen.
  5. **Observerbarhet fra dag én, ikke etterpå.** Application Insights med token-tracking per agent-rolle er forutsetning for produksjonsdrift, ikke nice-to-have.
- 

### Dissens

**Orkestreringsrammeverk:** SI\_002, SI\_008 og SI\_009 anbefaler **Semantic Kernel** (Microsoft-nativt, Python-basert). SI\_001, SI\_004, SI\_006 og SI\_007 anbefaler **Azure Durable Functions** (serverless fan-out/fan-in). Begge er gyldige, men Durable Functions er mer cloud-native og infrastrukturlett; Semantic Kernel gir mer programmatisk kontroll over agent-logikk.



**Startpunkt:** SI\_003 er eneste instans som anbefaler Copilot Studio som midlertidig første steg for tillitsbygging. Øvrige avviser dette. SI\_003 har et poeng om organisatorisk tillitsgap, men undervurderer throttling-problemet i praksis.

---

## Blindsoner avdekket

- **Forhandlingsmakt mot Microsoft** (SI\_007, SI\_008): Ingen enkelt-instans ville sannsynligvis fremhevet at 200 ansatte i helsevesen gir reell forhandlingsposisjon på EA-avtalen. Azure OpenAI private endpoint og data residency-garanti er noe du forhandler inn, ikke kjøper.
- **Sekvensiell avhengighet som sverm-killer** (SI\_008): Når output fra agent A endrer premiss for agent B, er sverm feil. Dette er en konkret failure mode som enkelt-instans-analyse typisk overser.
- **Complexity scoring som ressursallokering** (SI\_002, SI\_009): Automatisk routing av enkle cases til gpt-4o-mini og komplekse til gpt-4o er undervurdert som kostnadsoptimering — 60–70 % kostnadsreduksjon er realistisk.

---

## Anbefalinger

1. **Denne uken:** Forhandle Azure OpenAI private endpoint + Norway East data residency inn i eksisterende Microsoft EA-avtale før arkitekturmøtet.
2. **Uke 1-2:** Etabler Azure OpenAI-instans med Entra ID-rolleoppsett. Kartlegg hvilke Dataverse-tabeller og SharePoint-biblioteker som er lovlige å injisere — dette blokkerer alt annet.
3. **Uke 3-6:** Deploy minimal Durable Functions-orkestrator eller Semantic Kernel-pilot med **5 rolle-differensierte agenter** på én reell, lavrisiko case-type. Ikke 100 fra dag én.
4. **Innen 60 dager:** Implementer token-tracking per agent-rolle i Application Insights og lever ROI-rapport til ledelse med faktiske kostnader. Dette er tillitsbyggingen som låser opp budsjett for fase 2.
5. **Definer human-in-the-loop-terskel skriftlig** før første produksjonskjøring: hvilken menneskelig review kreves før sverm-output brukes i klinisk beslutning?