



id: SI_009

model: qwen/qwen3.6-plus

seeds:

04_risiko: mulighet

03_relasjon: makt

05_handling: vent

01_perspektiv: økonomisk

02_tidhorisont: kortsiktig

Instans SI_009

Seeds: økonomisk | kortsiktig | makt | mulighet | vent

Arkitektur: Velg Azure Container Apps + Python (LangGraph) framfor Copilot Studio/PA. CS er sekvensiell og låser deg til dialogmønstre; proprietære rammeverk (CrewAI) skaper vendor-lock og data-lekkasje-risk. Makt beholdes ved Azure-native orkestrering med Entra ID-godkjenning. Økonomisk gevinst: Container Apps skalerer til null, perfekt for batch.

Rolle-differensiering (Case 1): Ikke kopi-prompt. Bruk Azure AI Search med semantisk chunking. Hver agent får unikt `system_prompt` + kontekst via vektet RAG. Syntese: En meta-orchestrator (gpt-4o-mini) kjører konsensus-clustering.

```
results = await asyncio.gather(*[a.run(problem, ctx[i]) for i in range(100)])
synthesis = meta.run(results, schema=ConsensusModel)
```

Vent: Valider på 10 agenter før full skala for å unngå token-sløsing.

Inbox-modell (Case 2): Azure Service Bus queue + Python Workers. Hver worker trekker 1 case, injiserer standard systemprompt med Dataverse-payload. Konsistens: JSON Schema enforcement + temp 0.2. Kortsiktig effektivitet: Batch på timer muliggjør spot-prising og automatisk retry.

Data & Sikkerhet: Ingenting forlater tenant. RAG via Private Link. Tilgang arves fra Entra ID RLS. Kritisk: Slå av OpenAI-logging, bruk Purview for lineage, krypter alt. Makt over data > modellstørrelse i helsevesen.

Observerbarhet & Kostnad: App Insights + OpenTelemetry. Spor `session_id`, token, latency, feilrate. Strategi: Fallback til gpt-4o-mini ved timeout, cache RAG i Cosmos DB, aggreger API-kall. Økonomisk ROI: Mål NOK/case vs manuell tid.



Overgang & Fallgruver: Fase 1 (4 uker): 5 agenter, triage. Fase 2 (8 uker): Swarm på Functions. Swarm er feil hvis volum <50/dag eller krav til <2s latency. Da skaper det operasjonell kompleksitet uten mulighet for skalering.

Neste steg:

1. Implementer Entra ID RLS på Dataverse for sikker RAG-injection.
2. Deploy 10-agent batch-pipeline på Azure Container Apps med Service Bus queue.
3. Sett opp cost-per-case dashboard før du skalerer til 100.