



id: SI_005

model: qwen/qwen3.6-plus

seeds:

04_risiko: mulighet

03_relasjon: makt

05_handling: utforsk

01_perspektiv: strategisk

02_tidhorisont: langsiktig

Instans SI_005

Seeds: strategisk | langsiktig | makt | mulighet | utforsk

1. Arkitektur & makt: Unngå Copilot Studio for sverm – den er UI-basert og begrenser orkestrering til lineære flows. Velg Azure OpenAI + Azure Functions + Azure Service Bus. Dette gir programmatisk makt over routing, feilhåndtering og dataflyt, mens det bevarer Microsoft-investeringen. Proprietære rammeverk (CrewAI/AutoGen) legger til unødvendig overhead og bryter med din strategiske sky-only-policy.

2. Spesialisering & utforsk: Duplisering skaper støy. Bruk en «Router-Dispatcher» med dynamisk parameter-injisering. Hver agent får en base-prompt + kontekst-lens:

```
role_lens = {"perspektiv": "regulatorisk", "data_sone": "policy_only",  
"utdata_fmt": "json_schema_v3"}  
prompt = f"{base_prompt}\n\nAGENS-ROLL: {role_lens}"
```

Aggregering skjer via én «Synthesizer»-agent som vakter svar mot en faglig rubrikk (ikke rå concatenation). Dette utforsker kompleksitet uten å miste kontroll.

3. Inbox-modell & langsiktig: Azure Service Bus Topics med subscriptions per agent-type. Agenter puller fra kø, validerer output mot Pydantic-skjema, og skriver til Azure Cosmos DB. Variable cases rutes til «fallback»-kø. Skema-validering sikrer langsiktig consistency automatisk.

4. Data-injisering: Ingen data forlater Azure. Agenter bruker Azure AI Search med Entra ID RBAC og «on-your-data» grounding. Kritisk: Private Endpoints, Microsoft Purview DLP, og audited managed-identity-passing. Sikkerhet er arkitekturens grunnmur, ikke et tillegg.

5. Observerbarhet & kostnad: Application Insights + OpenTelemetry. Track token/agent via custom dimensions. ROI-optimalisering: `gpt-4o-mini` for routing/syntese,



gpt-4o kun for kjerneanalyse. Cache embeddings i Azure Cache for Redis. Batch-kjøring utnytter Functions consumption plan uten latency-kompromiss.

6. Overgang & mulighet: Fase 1 (uker 1-4): 5-10 agenter, manuell godkjenning. Fase 2 (mnd 2-3): Auto-routing, streng skjema-validering. Fase 3 (mnd 4-6): Full sverm med automatisk fallback. Quick wins: (1) Parallell review av 50 historiske helse-saker, (2) Auto-triage av internt sakskø. Muligheten ligger i parallelisering av ekspertise, ikke i volum.

Når svermen feiler: Deterministiske prosesser, sanntidskrav (<2s), eller air-gap-miljøer. Sverm er designet for tolkning, ikke transaksjon.

Neste steg:

1. Provision Azure Service Bus + Functions med Entra ID managed identity og private endpoints.
2. Definer 3 agent-roller med eksakte JSON-skjemaer; test på 20 anonymiserte cases.
3. Implementer Application Insights-dashbord før skala økes. Strategisk kontroll må etablere muligheten.