



id: SI_004

model: qwen/qwen3.6-plus

seeds:

04_risiko: resiliens

03_relasjon: tillit

05_handling: implementer

01_perspektiv: operasjonell

02_tidhorisont: historisk

Instans SI_004

Seeds: operasjonell | historisk | tillit | resiliens | implementer

Arkitektur: Unngå Copilot Studio for sverm; den er optimalisert for UI-samtale, ikke batch-parallellitet. Velg Azure OpenAI + Python (FastAPI) + Azure Container Apps. Bruk CrewAI/LangGraph kun som logikk-bibliotek, ikke runtime. Dette bevarer MS-investering via Entra ID, men gir nødvendig operasjonell kontroll over throughput og nettverksisolasjon. Historisk lesson: Enterprise-orchestrering kollapser når abstraksjonen skjuler feilrøtter – hold runtime eksplisitt.

Spesialisering: 100 identiske prompts genererer støy. Definer fagdomener med sub-roller. Injiser kontekst via Azure AI Search med strikt scope per rolle. Aggregering skal ikke tvinge konsensus, men tillit-skape gjennom avviskartlegging. Syntese-agenten kvantifiserer uenighet. Pseudo: `dispatch(role, context_slice(role)); return resolve(conflicts)`.

Inbox-modell: Azure Service Bus styrer køen. Worker-pool allokterer dynamisk basert på kompleksitets-score. Del system-template med parametrisert kontekst. Variabel kompleksitet håndteres via modell-routing: enkle → `gpt-4o-mini`, komplekse → `gpt-4o`. Resiliens krever idempotent prosessering og eksponensiell backoff ved 429-feil.

Data & kontekst: Helsedata forlater aldri vNET. Bruk Azure OpenAI med Managed Identity og Private Link. Injeksjon skjer via vektor-referanser, ikke bulk-kopiering. Entra ID Conditional Access begrenser lesetilgang. Audit-logg av hver token-generasjon er ikke-valgfritt. Tillit bygges gjennom deterministisk datagrense.

Observasjon & kostnad: Application Insights med custom dimensions (`agent_id` , `case_type`). Track token-forbruk per rolle. Cost-optimering: Batch til 50 req/s, cache RAG i Redis, trim prompt-bloat (historisk viser 60% kostnad er unødvendig kontekst). Fallback til rimeligere modeller ved rate-limit.



Overgang: 12 uker. Uke 1-4: 10-agent prototype. Uke 5-8: Service Bus + kø-drift. Uke 9-12: Produksjonsbatch. Sverm feiler ved sterkt kausalt resonnement eller når datasiloer bryter tillitsmodellen. Da er enkel RAG overlegen.

Neste steg:

1. Etabler Private Link til Azure OpenAI og definer dataklassifisering for RAG-scoping.
2. Bygg 5-agent POC med Application Insights-telemetri og idempotent kø-håndtering.
3. Spesifiser avviks-aggregeringslogikk før du skalerer til 100.