



id: SI\_002

model: qwen/qwen3.6-plus

seeds:

04\_risiko: sårbarhet

03\_relasjon: makt

05\_handling: implementer

01\_perspektiv: strategisk

02\_tidhorisont: kortsiktig

## Instans SI\_002

**Seeds:** strategisk | kortsiktig | makt | sårbarhet | implementer

**1. Arkitektur:** Velg Azure OpenAI + Python (Semantic Kernel/AutoGen) på Azure Container Apps. Copilot Studio/Power Automate kollapser under 100 parallelle kall (state-mangling, hard rate-limits). Strategisk beholder du makt over token-ruting og unngår low-code lock-in. Implementer orkestrering via Azure Service Bus + Durable Functions.

**2. & 3. Spesialisering & Inbox:** 100 dupliserte prompts gir konvergens-kaos. Bruk 12-18 spesialiserte roller med dynamisk `context_injection`. For Case 2 router Service Bus-meldinger basert på pre-beregnet kompleksitet.

```
# Rollerspesifikk kontekst, ikke prompt-duplisering
prompt = SYS_TEMPLATE.format(role=role,
context=ai_search_retriever.query(metadata))
queue.send(message=case, properties={"route": "high" if complexity > 0.7 else
"batch"})
```

Aggregering: `meta-agent` veier svar etter autoritet og konfidens. Sårbarhet: Swarm amplifiserer hallusinasjoner ved dårlig grounding. Implementer strict threshold før konsolidering.

**4. Data & kontekst:** Helse = null kompromiss. Azure AI Search med managed identity og Entra ID RBAC. Data kopieres aldri; retrieval skjer inline via privat endpoint og token-scoped principals. Kritisk: Content filtering før context-window, full audit-logg.

**5. Observerbarhet & kortsiktig ROI:** Tag kall med `agent_id` + `use_case`. Track via Application Insights. Kortsiktig kostnadskutt: Fallback til `gpt-4o-mini` for low-complexity, cache embeddings, batch requests. Auto-throttle ved >12% feilrate. ROI målt i `cost_per_decision`.



**6. Transition & dissens:** 0-30d: PoC (5 agenter, batch). Quick win: Dokumentreview. 30-90d: Entra ID governance, observability live. 90-180d: Full drift. Dissens: Swarm er feilvalg for deterministiske, compliance-krevde oppgaver der audit må være 100% reproducerbar. Bruk kun der usikkerhet er høy og parallell perspektiv reduserer risiko.

**Neste steg:**

1. Definer 5 agent-roller med eksakte `context_injection` -boundaries; test på anonymisert helsedata.
2. Bygg Service Bus + Durable Functions routing-lag med dead-letter queue før LLM-integrasjon.
3. Sett opp App Insights custom metrics for `cost_per_case` og `consensus_score` som produksjonsgates.