



id: SI_007

model: qwen/qwen3.6-plus

seeds:

04_risiko: mulighet

03_relasjon: tillit

05_handling: implementer

01_perspektiv: økonomisk

02_tidhorisont: kortsiktig

Instans SI_007

Seeds: økonomisk | kortsiktig | tillit | mulighet | implementer

Arkitektur: Unngå Copilot Studio for swarm (begrenset orkestrering, høy løpende kostnad). Velg **Azure AI Foundry + Azure Functions + Service Bus**. Serverless, Entra ID-native og optimalt for cloud-only. Eksterne rammeverk (CrewAI/LangGraph) krever eksponering av prompts/data utenfor tenant – bryter umiddelbart helse-tillit. Bevar Microsoft-investment, men flyt kontroll til Azure.

Case 1 - Perspektiv: Prompt-duplisering er økonomisk selvmord. Bruk kontekst-slicing: Del problemet i 10-12 spesialiserte domener. Hver agent får unik system-prompt + relevante Dataverse/SharePoint-felter via søkeindeks. Aggregering via en «synthesis-agent» som krever Pydantic-JSON. Pseudo: `results = await asyncio.gather([a.run(slice) for a in agents])` Kaos unngås med streng schema-validering, konfidens-vekting og avvisning av output uten struktur.

Case 2 - Inbox: Service Bus Queue som dispatcher. Metadata-merker cases til kompleksitetsnivå. Enhetlig system-prompt med dynamisk `{case_template}` sikrer consistency. Batch-kjøring (timer) gir kortsiktig økonomisk fordel: lavere throughput-kostnad, bedre token-caching, mindre retry-overhead.

Data & Sikkerhet: Bruk **Azure AI Search** med semantic indexing. Data forblir i tenant. Entra ID-managed identities styrer tilgang. Purview PII-maskering før prompt-injeksjon er ikke-negotiable. Ingen kopiering til eksterne API-er – tillit er hard currency i helse.

Obserbarhet & Kostnad: Application Insights med custom metrics (`token_cost` , `agent_latency` , `error_rate`). Cost-guards: auto-fallback til `gpt-4o-mini` ved lav kompleksitet, response-caching for repetitive queries. Mål: 40-60% token-reduksjon via strukturert output og batch.



Overgang & Failure: Fase 1 (0–4 uker): 10-agent pilot på anonymisert data. Fase 2 (4–12 uker): Entra/Service Bus-integrasjon + cost-tracking. Swarm feiler på lineære/oppgaver med <500 tokens eller ustrukturerte data. Over-engineering dreper både økonomi og operasjonell tillit.

Neste steg:

1. Map 5 brukstilfeller til Dataverse-skjemaer for kontekst-slicing.
2. Aktiver Purview DLP + Entra Managed Identity før første API-call.
3. Implementer cost-logging dag 1 – mulighet krever kontroll, ikke blind skalerbarhet.