



id: SI_003

model: qwen/qwen3.6-plus

seeds:

04_risiko: sårbarhet

03_relasjon: makt

05_handling: vent

01_perspektiv: juridisk

02_tidhorisont: kortsiktig

Instans SI_003

Seeds: juridisk | kortsiktig | makt | sårbarhet | vent

1. Arkitektur: Unngå Copilot Studio/Power Automate for 100 parallelle noder (rate-limits, tilstandsstyring svak). Proprietære rammeverk (CrewAI, Autogen) bryter med helsevesenets sårbarhet-krav om data residency og leverandørkontroll. Velg Azure AI Agent Service + Python i Azure Container Apps. Dette keeper makt over infra, gir Entra ID-styring, og muliggjør ekte parallell scaling uten skygge-IT.

2. Spesialisering (Case 1): 100 agenter \neq 100 kopier. Bruk dynamisk prompt-routing via rolle-vektmatrise. Pseudo:

```
sys_prompt = base + f"Du analyserer kun {role_focus}. Vektlegg {reg_clause}"
```

Tving JSON-schema output. Aggregering: Én «arbiter»-agent med konsensus-logikk (vektet flertall + confidence-threshold >0.8). Kaos unngås ved å droppe output som bryter schema og kjøre syntese først etter validering.

3. Inbox-modell (Case 2): Azure Service Bus som kø. Python-workers i Container Apps henter meldinger (prefetch=1). System-prompt: Én template med kontekst-injeksjon via placeholders. Konsistens sikres med `temperature=0.2` + Pydantic-validering før Dataverse-commit. Variabel kompleksitet løses med auto-routing til agent-pool (T1/T2/T3) basert på case-metadata og estimated token-cost.

4. Data & sikkerhet: Juridisk bindende: Aldri eksporter rå PII. Bruk Azure AI Search med Private Endpoint + RAG med PII-redaksjon-proxy foran modell. Data forblir i EU. Kritisk: Entra ID Managed Identity for alle kall, Purview for data lineage, og full audit-logg per agent-call. Injeksjon skjer via tokenisert referanse, ikke bulk-kopiering.

5. Observabilitet & kost: Application Insights med custom dimensions per agent. Spor `tokens`, `latency`, `error_rate`. Kostnadsstyring: Kjør batch om natten (vent til lav belastning), cache embeddings, og bruk fallback til `gpt-4o-mini` ved lav kompleksitet. ROI oppnås ved å reservere tunge modeller kun for arbiter/final scoring.



6. Overgang: Kortsiktig fokus: 8–10 uker til POC med 10 agenter på anonymiserte data. Full drift: 4–6 mnd. Quick win: Automatisert triage av henvisninger med «human-in-the-loop» sign-off.

Feilmodus: Swarm er feil valg hvis regelbasert flow dekker >70%, eller hvis juridisk ansvar ikke kan plasseres entydig. Uten streng governance blir svermen en compliance-bombe.

Neste steg:

1. Kartlegg datagrenser med Databehandleransvarlig – definer eksakt hvilke felter som kan RAG-gis uten nytt samtykke.
2. Bygg en 5-agent orchestrator i Container Apps med mock-Service Bus for å teste feilhåndtering og token-tracking.
3. Implementer Entra ID RBAC og Pydantic-skjemaer før første produksjons-call.