



id: SI_009

model: deepseek/deepseek-v3.2

seeds:

04_risiko: mulighet

03_relasjon: konkurranse

05_handling: utforsk

01_perspektiv: økonomisk

02_tidhorisont: kortsiktig

Instans SI_009

Seeds: økonomisk | kortsiktig | konkurranse | mulighet | utforsk

Analysen din er her.

- Arkitektur-valg:** Bygg kortsiktig på Azure OpenAI + Python (FastAPI/Flask) i isolert VNet. Det er den **økonomisk** sikreste måten å **utforske** swarm mens du preserverer Microsoft-investering og helsedata. Proprietære rammeverk (CrewAI) gir rask start, men låser deg til deres stack. Copilot Studio er for enkelt for 100 agenter. Trade-off: Du må orkestrere selv (Kubernetes/Container Apps), men har full kontroll.
- Agent-spesialisering:** Ikke 100 duplikater. Design 8-12 distinkte rolleprofiler (f.eks. «GDPR-compliance-filter», «kostnadsoptimering», «konkurranse-risikovurdering»). Hver profil får sin egen, datamatet systemprompt og tilgang til spesifikke datakilder (f.eks. kun policy-dokumenter). Deretter instansieres flere kopier av hver rolle for parallell ytelse. Aggregering skjer via et «sammendragsagent»-mønster som trekker ut konsensus og dissens – dette håndterer «meningers kaos».
- Inbox-modell:** Bruk Azure Service Bus (queue). En «dispatcher»-funksjon leser cases, vurderer metadata (kompleksitet), og plasserer i dedikerte køer for hver agent-type. Hver agent-type har en delt systemprompt-mal, men med dynamisk kontekst-injection per case. Consistency sikres ved at alle resultater valideres mot et grunnsett av business rules før lagring i Dataverse.
- Data-injection:** Kritisk punkt. Aldri send rå helsedata til API. Bruk Azure AI Search (forhåndsindeksert) med strenge access-kontroller. Agenter henter kun relevante dokumentfragmenter via søk med sikret identitet (Entra ID). Alle kall logges for etterlevelse. Dette minimerer eksponering.



5. **Cost-optimering:** Batch alt mulig. Bruk billigere modeller (gpt-3.5-turbo) for forarbeid, reservér gpt-4 for final analyse. Implementer caching av vanlige spørsmål/svar. Overvåk med Azure Monitor – lag dashboards for token-forbruk per agent-profil og feilrate. Stopp automatisk ved uvanlige kostnadsspiker.
6. **Transition:** Ikke en «big bang». Fase 1 (6 uker): Bygg én spesialisert agent-gruppe (f.eks. 5 regulatoriske agenter) for en konkret analyse-case. Demonstrer **mulighet** og verdi. Fase 2: Skaler til flere roller og introduser inbox-modellen. Realistisk tidslinje til delvis produksjon: 4-6 måneder.

Neste steg:

1. Start et proof-of-concept i Azure med én agent-profil (f.eks. «regulatory reviewer») koblet til Azure AI Search på et begrenset dokument-sett.
2. Etabler omfattende logging og cost-tracking fra dag én for å kvantifisere **økonomisk** påvirkning.
3. Test «dispatcher»-mønsteret med 10 test-cases i Service Bus for å validere orkestrering før full skalering.