



Debrief — qwen/qwen3.6-plus

- Instanser: 9
 - Tokens inn: 19,636
 - Tokens ut: 34,637
 - Kostnad: \$0.0739
 - Kjørt: 2026-04-14 16:07:22.807235+00:00
-

Sverm-debrief

Konsensus

- **Avvisning av low-code orkestrering:** Alle instanser enes om at Copilot Studio/Power Automate kollapser under 100 parallelle kall pga. state-mangling og harde rate-limits. Azure-nativ stack (Azure OpenAI + Python i Container Apps/ Functions + Azure Service Bus) er eneste akseptable valg for helsevesenets sikkerhetsnivå.
 - **Ingen prompt-duplisering:** Spesialisering krever dynamisk kontekst-injeksjon via rolle-spesifikke RAG-filtre og streng JSON-schema-validering. Aggregering må skje via én «Arbiter»-agent med lav temperatur (0.1-0.2) og vektet konsensus.
 - **Null dataeksport:** All grounding skjer innenfor tenant via Azure AI Search med Private Endpoints, Entra ID Managed Identity og Purview DLP. Agenter injiseres kun med tokeniserte referanser eller maskert metadata.
 - **Innebygd økonomi & observabilitet:** Application Insights med custom dimensions (`agent_id` , `tokens` , `error_rate`) er obligatorisk. Kostnadskontroll krever fallback til `gpt-4o-mini` , embedding-caching, batch-kjøring og hard budsjettgrense (auto-stop ved 85 %).
-

Dissens

- **Runtime-miljø:** Flere foretrekker Azure Functions for ren serverless-skalerbarhet, mens andre argumenterer for Azure Container Apps for bedre nettverksisolasjon og kontroll over parallellitet. SI_006 foreslår Copilot Studio som frontend, noe majoriteten avviser som en «operasjonell blindgate» for swarm.



- **Rammeverk-bruk:** Noen (SI_004, SI_009) åpner for LangGraph/CrewAI som rene logikk-bibliotek, mens flertallet (SI_001, SI_003, SI_007) krever fullt Azure-native kode for å unngå vendor-lock-in og utilsiktet dataeksponering.
- **Aggregeringsmål:** Uenighet om output-filosofi: vektet flertall/konsensus (SI_001, SI_006) vs. eksplisitt avvikskartlegging og konfliktkvantifisering (SI_004). Sistnevnte prioriterer «tillit gjennom dokumentert uenighet» fremfor tvungen enighet.

Blindsoner avdekket

- **Prompt-bloat:** Opptil 60 % av token-kostnaden skyldes unødvendig kontekst i system-prompts. Swarm uten streng kontekst-slicing blir økonomisk uholdbar.
- **Arbiter-bottleneck:** Syntese-agenten kan bli et enkelt feilpunkt og kostnadslekkasje hvis den ikke har egen rate-limit, fallback-logikk og isolert cost-tracking.
- **Juridisk ansvars plassering:** Swarm flytter beslutningsvekt til algoritmen. Uten eksplisitt mapping av «human-in-the-loop»-signaturer og audit-trail per agent-call, bryter løsningen helsevesenets compliance-krav.
- **Hallusinasjons-amplifisering:** Parallell kjøring uten streng confidence-threshold (>0.8) og schema-validering forsterker feil før aggregering.

Anbefalinger

1. **Bygg minimal orchestrator (uke 1-4):** Deploy Azure Service Bus + Python-workers i Container Apps. Test 5 spesialiserte agenter med anonymisert helsedata, streng Pydantic-validering og Application Insights-telemetri.
2. **Implementer zero-trust dataflyt:** Konfigurer Azure AI Search med Private Endpoints, Entra ID RBAC og Purview DLP. Sikre at agenter kun henter indeksreferanser, aldri rå PHI.
3. **Definer governance-gates:** Sett opp auto-stop ved 85 % budsjett, kill-switch for feilrate $>12\%$, og eksplisitt «human sign-off»-punkt før Arbiter-output commit-es til Dataverse.
4. **Fase ut Copilot Studio for orkestrering:** Bruk den kun til sluttbrukergrensesnitt. All routing, kø-håndtering og agent-livssyklus må kodes i Azure-native tjenester for å sikre skalerbarhet, auditability og compliance.