



# Debrief — openai/gpt-5.4-mini

---

- Instanser: 9
  - Tokens inn: 18,866
  - Tokens ut: 7,311
  - Kostnad: \$0.1663
  - Kjørt: 2026-04-15 17:24:35.314284+00:00
- 

## Sverm-debrief

---

### Konsensus

1. **Start med begrenset autonomi, ikke full produksjonsmakt.** Alle instanser anbefaler lese-/foreslå-først, med svært smale, reverserbare handlinger i pilot.
  2. **Runtime-kontroll må være obligatorisk.** Policy-as-code, deny-by-default og en policy-gate/proxy foran agenthandlinger ble sett som nødvendig, særlig siden dere mangler sentral policy-motor og helhetlig Zero Trust.
  3. **Agenten må ha kortlevde og smalt scoped tilganger.** Ephemeral credentials, tidsbegrensede tokens, task-scoped service accounts og hyppig rotasjon er et klart fellestrekk.
  4. **Execution graph og audit er kritisk.** Alle var enige om at dere må modellere plan → policy check → kall → respons → sideeffekt, med korrelasjons-ID, og ha en append-only audit-pipeline.
  5. **Legacy/API-risiko er høy.** Manglende idempotens, uklare kontrakter, delvise sideeffekter og latent coupling ble løftet som de største operasjonelle farene.
- 

### Dissens

- **Tempo og ambisjonsnivå:** Noen perspektiver er mer “vent og bygg kontroll først”, mens andre er mer “implementer en smal pilot nå”. Uenigheten handler ikke om målet, men om hvor raskt man bør gi agenten reell utførelse.
- **Styringsmodell:** Noen fremhever en sentral policy-motor som ideal, andre aksepterer en pragmatisk overgangsløsning med policy-gateway, sidecars og server-side enforcement.



- **Autonominivå:** Enkelte legger vekt på eksplisitte nivåer N0–N3 og kun read-only i pilot, mens andre åpner for begrensede writes dersom blast radius og rollback er godt nok kontrollert.

---

## Blindsoner avdekket

- **Syklisk feilforsterkning:** Agent → API-kall → sideeffekt → ny observasjon → nytt kall. Dette kan skape runaway-loops selv uten “feil” i enkelthandlinger.
- **Latent coupling:** Små, tilsynelatende trygge handlinger kan utløse batchjobber, varsler, økonomiske prosesser eller integrasjoner andre steder.
- **Reverserbarhet er ofte illusorisk:** I legacy-systemer kan feil oppdages sent, og rollback kan være delvis umulig.
- **Datafeil som maktproblem:** Dårlig datakvalitet kan få agenten til å handle “korrekt” på feil premisser.
- **Mangel på revisjonsspor:** Distribuerte logger alene gir ikke nok kontroll eller etterprøvbarehet.

---

## Anbefalinger

1. **Definer 3 autonominiivåer og en no-go-liste** for irreversible handlinger før pilot.
2. **Sett opp en minimal policy-gateway + audit-pipeline** rundt én avgrenset pilotprosess.
3. **Bruk ephemeral credentials og task-scoped tilgang** med TTL i minutter, ikke stående nøkler.
4. **Velg én modern og én legacy-integrasjon** og test blast radius, idempotens, rollback og kill-switch i sandbox.
5. **Lag stoppkriterier for runaway-agenter:** loopdeteksjon, policybrudd, ukjent sideeffekt, manglende sporbarhet.