



Debrief — openai/gpt-5.4-mini

- Instanser: 9
 - Tokens inn: 14,540
 - Tokens ut: 7,613
 - Kostnad: \$0.1578
 - Kjørt: 2026-04-16 00:48:12.685904+00:00
-

Sverm-debrief

Konsensus

- **Agentisk KI må behandles som en ny aktør med begrenset mandat**, ikke som en vanlig integrasjon. Alle instansene avviser bred, direkte tilgang til fagsystemer.
 - **Kontroll må ligge utenfor modellen**: policy-motor, orkestrator eller gateway skal avgjøre hva som kan utføres før kall mot systemer.
 - **Minste privilegium per handling** er gjennomgående: oppgavespesifikke rettigheter, korte levetider på tokens, egne tjenestekontoer per agent/oppdrag.
 - **Irreversible handlinger må sperres eller kreve menneskelig godkjenning**. Typiske røde soner er sletting, utbetaling, tilgangsendring, vedtak og eksterne utsendelser.
 - **Sporbarhet må bygges inn fra start**: sentral audit-logg, korrelasjons-ID på tvers av systemer, og logging av beslutning, policyutfall og effekt.
-

Dissens

- **Venting vs implementering**: Noen instanser anbefaler å vente og starte svært forsiktig med shadow mode og lesing først; andre vil implementere en defensiv pilot raskt for å bygge kapasitet og læring.
- **Grad av autonomi i piloten**: Enkelte foreslår at lavrisiko handlinger kan gå fullt automatisk tidlig, mens andre mener alt utover lesing/utkast bør kreve godkjenning i pilotfasen.
- **Hvordan rollback skal forstås**: Noen vektlegger teknisk rollback, andre understreker at man i legacy-miljøer må planlegge kompensasjon eller blokkering fordi faktisk reversering ofte ikke finnes.



Blindsoner avdekket

- **Legacy-systemer er større risiko enn modellene:** manglende idempotens, skjulte sideeffekter, batchprosesser og svak transaksjonsstøtte kan gjøre ett kall irreversibelt.
- **Fragmentert logging skaper falsk trygghet:** lokalt ser hvert system korrekt ut, men helheten blir uetterprøvbar.
- **Eksfiltrasjon skjer indirekte:** sensitiv info kan lekke via parametere, metadata, feilmeldinger, oppsummeringer eller eksterne API-er.
- **Agenten kan eskalere indirekte** ved å kombinere flere små tillatte handlinger til en større effekt.
- **Maktforskyvning til policy-laget:** den som kontrollerer policy-gatewayen, kontrollerer i praksis virksomhetens handlingsrom.

Anbefalinger

1. **Start med 2-3 lavrisiko use cases** og klassifiser alle handlinger i lese, foreslå, klargjøre, utføre, forby.
2. **Etabler policy-gateway før pilot** med allowlist, TTL-baserte tokens, rate limiting og hard stopp for irreversible handlinger.
3. **Lag én sentral audit-kjede** med korrelasjons-ID fra agentinput til systemeffekt, og test den mot ett legacy- og ett modernesystem.
4. **Definer røde linjer for data og handlinger:** ingen direkte tilgang til produksjonsdata der det ikke er nødvendig, og DLP/redaksjon før eksterne kall.
5. **Test feilscenarier eksplisitt:** dobbelkall, feilskriv, dataeksfiltrasjon, uautorisert eskalering og rollback/kompensasjon før utvidelse.