



Debrief — anthropic/claude-sonnet-4.6

- Instanser: 9
 - Tokens inn: 10,631
 - Tokens ut: 8,039
 - Kostnad: \$0.1525
 - Kjørt: 2026-04-14 00:08:46.862885+00:00
-

Sverm-debrief

Konsensus

1. **Premisset er feil:** Alle 9 instanser avviser at det finnes en koordinert “illusjon”. Mekanismen er strukturell treghet, ikke aktiv suppresjon.
 2. **LLM-er har ikke løst autonome systemer:** Formell verifikasjon, robusthet under distribusjonsskifte, og sikkerhetsgarantier er genuint åpne problemer.
 3. **Incentivstrukturen er reell:** Tenure, PhD-pipelines, og NFR/Horizon-finansiering skaper konservativt seleksjonspress uten at noen trenger å lyve.
 4. **20 års ekspertise er en fordel, ikke en byrde:** Domenekunnskapen gjør deg til den best posisjonerte til å kartlegge LLM-enes faktiske grenser.
 5. **Peer review er selvforsterkende:** Reviewerne er feltet — disruptive papers får strukturelt lavere score uten ondsinnet intensjon.
-

Dissens

Alvorlighetsgraden av disrupsjon splittet instansene. SI_006 og SI_007 lente mot at deler av feltet er reelt foreldet og at maktspeilet er “brutalt transparent”. SI_001, SI_003 og SI_009 argumenterte mer aktivt for at forskerens premiss er feil og at LLM-trusselen er overvurdert. Ingen instans tok premisset for gitt ukritisk.

Blindsoner avdekket

Epistemisk segregering (SI_009): Den viktigste mekanismen er ikke intern akademisk beskyttelse, men at debatten allerede foregår — bare utenfor kanalene professorer kontrollerer. NeurIPS er halvt industriell. Diskusjonen skjer på arXiv og i industrilaboratorier, ikke i instituttmøter.



Waymo-argumentet: Ingen enkeltinstans ville sannsynligvis ha fremhevet at sikkerhetskritisk industri allerede har svart på spørsmålet i praksis — de bruker ikke LLM-er i kjernefunksjoner. Det er sterkere empirisk evidens enn akademisk debatt.

Selvdiagnoseproblemet: Svermen avdekket at spørsmålet selv er et symptom — en forsker med 20 år i feltet som spør om andre opprettholder illusjoner, bør først spørre om sin egen vurdering av “løst” er presis.

Anbefalinger

1. **Publiser én presisjonsstudie:** Velg ett konkret problem fra din forskning, test systematisk mot GPT-4o/Claude, dokumenter nøyaktig hvor de feiler. Dette er uangripelig og høy-impact akkurat nå.
2. **Skriv posisjonspaperet, ikke klagen:** “What LLMs cannot guarantee in autonomous systems” i Communications of the ACM — ikke et internt notat. Tving debatten ut av instituttkorridorene.
3. **Ring Waymo Research eller DeepMind Safety:** Spør direkte hva de ikke kan bruke LLM-er til. Industriens faktiske praksis er sterkere argument enn akademisk polemikk.
4. **Kartlegg NFR-porteføljen:** Hvilke aktive prosjekter i ditt felt ville miste legitimitet ved full åpenhet? Det gir deg det konkrete maktkartet — og potensielle allierte som også ser problemet.
5. **Slutt å vente på institusjonell tillatelse:** Åpen diskusjon finnes allerede — du er bare ikke i rommet der den skjer.