Hey there,

--- SECTION 2: NOTABLE DEVELOPMENTS ---

Anthropic -- A researcher shared results showing that automated systems improved model performance across 10 benchmarks targeting specific misaligned behaviors, with no degradation in overall performance. This is a meaningful signal that automated, targeted behavioral correction at scale is becoming practical -- a key prerequisite for self-improving alignment pipelines. If this approach generalizes beyond curated benchmark sets, it could shift how labs iterate on safety and capability simultaneously. TechCrunch

--- SECTION 3: ALSO WORTH NOTING ---

- **Apple ML Research** -- Introduces the "information processing gap" metric to quantify how LLMs deviate from Bayesian belief updating, with implications for deploying LLMs in medicine, law, and other domains requiring calibrated uncertainty. Apple ML Research

- **Apple ML Research** -- Proposes Agent Seer, a method to automatically synthesize realistic evaluation scenarios for tool-using AI agents from API specifications alone, addressing the scalability problem of hand-built agent benchmarks. Apple ML Research

Back in 48 hours.

P.S. 15 RSS feeds | last 48h | 15 entries fetched, 3 scored >= 5