Hey there,

No breakthroughs to report this cycle, but a couple of items worth your attention.

--- SECTION 2: NOTABLE DEVELOPMENTS ---

IEEE Spectrum profiles **X Square Robot's** effort to build an end-to-end foundation stack for general-purpose embodied AI, arguing that robotics lacks the "pretrain on broad data, get general capability" recipe that transformed NLP. The piece frames the central challenge as unifying perception, planning, and control into a transferable, task-agnostic system rather than stitching together narrow modules. If this approach matures, it could accelerate deployment of general-purpose robots in warehouses, field robotics, and potentially manipulation-heavy scientific automation. Watch for whether they release open benchmarks or model weights that let external labs validate the generality claims. IEEE Spectrum

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

- **IEEE Spectrum** -- A researcher demonstrated systemic jailbreak vulnerabilities across nearly all major LLMs, exposing an industry-wide safety gap with implications for any domain deploying LLM-based agents. link
- **Apple ML Research** -- Released **Pare**, a framework for simulating stateful, sequential user interactions to evaluate proactive AI assistants, addressing a key gap in agent evaluation methodology. link
- **MIT Technology Review** -- Analyzes Anthropic's latest interpretability research, noting what it does and doesn't establish about model internals and potential model "experience." link
- **Ars Technica** -- Reports on "context bombing," a defensive technique where defenders use prompt injection against malicious AI hacking agents to shut them down before they cause harm. link

Back in 48 hours.

P.S. 15 RSS feeds | last 48h | 34 entries fetched, 6 scored >= 5