

Treat Every LLM Change as a Hypothesis Test

Recommended Reading from Tech Founders

2026-09-22

Treat Every LLM Change as a Hypothesis Test

By Recommended Reading from Tech Founders • September 22, 2026

Martin Casado's standout talk recommendation turns stochastic AI work into a concrete measurement discipline; an open robotics corpus and two lighter organic picks round out the list.

The standout recommendation

An age of experimentation — BlueHat Asia 2026 slides

- **Content type / creator:** Talk slides. The sharing post is by Halvar Flake, who describes them as slides from his Microsoft BlueHat Singapore talk; the deck itself is titled *An age of experimentation* and credits Thomas Dullien, Cyber Security Research Team, OpenAI. [1, 2]
- **Link:** Slides
- **Recommended by:** Martin Casado, who calls it his favorite talk of the year and praises its clarity in an era of “AI shorthand.” [3]
- **Key takeaway:** The deck treats reasoning LLMs as more like fire than an internal-combustion engine: discovered more than constructed, still poorly understood, and improved mainly through empirical work. [2] Its practical rule is that every change in a probabilistic system is a hypothesis test; teams should ask what they want to show, measure rerun variation, identify randomness, and check whether changing conditions invalidate an IID assumption. [2]
- **Why it matters:** This is a useful operating manual for evaluating AI-enabled systems without mistaking activity for progress. The deck also recommends a benchmark hierarchy—small tests for rapid iteration and larger tests for periodic validation—while warning that the small tests must remain representative. [2]

Eidon AI’s egocentric robotics corpus

- **Content type / creator:** Open robotics dataset/resource, published by Solidic Labs Inc. (Eidon AI). The company has wound down, but the release remains available under CC-BY-4.0, with commercial use allowed when attribution is provided. [4]
- **Link:** Eidon AI on Hugging Face, including the video corpus and paired motion data
- **Recommended by:** Clem Delangue, who praised the decision to open-source 1,274 hours of egocentric robotics data across 13,451 recordings and called it a gift to the robotics community. He explicitly argued that open source lets work outlive the organization that created it. [5]
- **Key takeaway:** The release combines 13,451 egocentric videos with 779 million rows of 24 Hz orientation and motion data covering the same recordings; the two pieces join on `recording_id`. [4] The paired set contains 1,274 hours of video and arm tracking from 27 contributors, with automated quality-control scores. [4]
- **Why it matters:** This is an unusually actionable recommendation for embodied-AI work, as well as a concrete example of preserving research assets after a startup shuts down. The dataset is not a neutral snapshot of daily activity: the release warns that tasks are heavily weighted toward folding laundry, contributors are few and unevenly represented, and splits should be made by contributor rather than randomly. [4]

Product and idea-generation picks

“I design teams like terrorist organizations” — Peter Ellis product-advice video

- **Content type / creator:** A 90-minute YouTube interview featuring Peter Ellis, published through the @lennysan interview series. The associated post links to the video. [6]
- **Recommended by:** Keith Rabois, with the unqualified endorsement “Highly recommend.” [7]
- **Key takeaway:** The episode is organized around concrete product-leadership questions: managing Nikita Bier, why growth usually comes from the core, knowing when to stop, the shortcomings of the median product manager, and team design. [6]
- **Why it matters:** It is a focused listening assignment for leaders working through the trade-offs between product taste, growth, and organizational design—not a generic startup interview.

Sharif Shameem’s posts archive

- **Content type / creator:** Personal blog/archive by Sharif Shameem. Posts [8]

- **Recommended by:** Shaan Puri, who says he had just discovered the blog, calls it “very good,” and tells readers to read a couple of posts. [9]
- **Key takeaway:** Puri does not identify a particular post or lesson, so this is a browse recommendation rather than a prescribed syllabus. The archive’s visible material ranges from *The 28 AI tools I wish existed*—framed by the question “where are the tools?”—to proposals about improving large video calls and modeling programming as a reinforcement-learning task. [8]
- **Why it matters:** Treat it as an idea-generation archive: useful for finding sharp product and technical questions, but not evidence that the listed concepts were built, tested, or validated.

Sources

1. X post by @halvarflake
2. An age of experimentation — BlueHat Asia 2026
3. X post by @martin_casado
4. eidon-ai (Eidon AI)
5. X post by @ClementDelangue
6. X post by @lennysan
7. X post by @rabois
8. All Posts - Sharif Shameem
9. X post by @ShaanVP