

Build the system around the model

Recommended Reading from Tech Founders

2026-08-26

Build the system around the model

By Recommended Reading from Tech Founders • August 26, 2026

An organic recommendation list centered on the work between model capability and real outcomes: enterprise workflows, agent operations, chip/software co-design, founder education, and China-tech context.

A useful pattern runs through the strongest recommendations: the hard part is not access to a capable model but the design of the workflow, organization, or hardware that makes it useful. Aaron Levie's endorsement is the clearest anchor: he says the gap between AI models and enterprise workflows leaves room for applied-AI companies, and names context, change management, model routing, business-system integration, workflow UX, and evaluations as the work beyond model intelligence. [1]

Standout recommendation

Moats in the age of floods

- **Type / creator:** Article by @nayakkayak. The article frames intelligence as a utility that needs systems to distribute it into real-world outcomes, rather than assuming the model layer captures all the value. [2]
- **Link:** Read the article. [3]
- **Recommended by:** Aaron Levie, who calls it a strong account of applied-AI strategy at scale. [1]
- **Key takeaway:** The article argues that real work remains difficult because it carries context, exceptions, incentives, approvals, accountability, legacy systems, and human coordination; adoption can move faster than the institutional change needed to turn capability into productivity. [2]
- **Why it matters:** Use it as a filter for applied-AI companies: what do they own in context, change management, integration, workflow UX, routing, and evaluation—not merely which model they access? [1]

Infrastructure and inference

OpenAI Jalapeño

- **Type / creator:** SemiAnalysis article.
- **Link:** Read the article. Elad Gil shared it with the comment, “Impressive speed to tape out by OpenAI.” [4]
- **Recommended by:** Elad Gil.
- **Key takeaway:** SemiAnalysis describes Jalapeño as a generalized LLM-inference chip developed with Broadcom from a blank slate; design began in mid-2024 and reached manufacturing tape-out in roughly 16 months. The article attributes its reported performance to tight hardware/software co-design rather than narrow specialization for OpenAI models. [5]
- **Why it matters:** The useful lesson is the bring-up loop, not a simplistic “OpenAI beat Nvidia” claim: SemiAnalysis says the team produced strong results within nine months of tape-out and three months on real silicon, but also notes that the performance numbers came from OpenAI, the full benchmark suite was not run, and production was only scheduled to ramp gradually during 2027. [5]

DeepSeek’s published research on KV-cache compression

- **Type / creator:** Research; DeepSeek.
- **Recommended by:** Neil Movva, identified in the discussion as an ex-Nvidia GPU and kernel engineer who now runs Sail Research. [6]
- **Key takeaway:** Movva says the KV cache is still “quite uncompressed,” with storage potentially off by one or two orders of magnitude, and points to DeepSeek’s research as making roughly order-of-magnitude progress every year. [7]
- **Why it matters:** This is a concrete research direction for anyone studying inference cost: memory representation, not just raw compute, may be a major source of future efficiency gains. [7]

Agents and company design

Coding with agents

- **Type / creator:** A 50-minute podcast/video posted by @0xCodez, featuring a SpaceXAI engineer identified as ex-Cursor. [8]
- **Link:** Martin Casado’s recommendation.
- **Recommended by:** Martin Casado calls it the “Single best discussion ever recorded on coding with agents.” [9]
- **Key takeaway:** The featured engineer says he runs 10–20 agents that automate 90% of his routine work, with a “Chief of Staff” agent that knows about and manages the other bots. [8]
- **Why it matters:** It offers an operating example of delegation and coordination to examine, rather than another abstract claim that agents will

change software development. [8]

***Founder Mode* and Lee Robinson’s product-engineer post**

- **Type / creators:** Paul Graham essay; Lee Robinson blog post, which Guillermo Rauch refers to as “the rise of the product engineer.”
- **Link:** Rauch interview discussing both recommendations.
- **Recommended by:** Guillermo Rauch. He says *Founder Mode* resonated because founders must go deep and stay in the weeds, then describes scaling that behavior through autonomous product-level owners. [10]
- **Key takeaway:** Rauch calls Robinson’s post “great” and “spot-on” for combining product management with P&L, customers, expansion, retention, and engineering judgment. [10]
- **Why it matters:** Together, the two recommendations form an organizational-design pair: preserve founder-level depth while giving capable people end-to-end ownership of products.

How Universities Should Prepare Founders

- **Type / creator:** Essay by Paul Graham.
- **Link:** Read the essay. Ryan Hoover shared it while saying he has never been a fan of whiteboarding a startup. [11, 12]
- **Recommended by:** Ryan Hoover.
- **Key takeaway:** Graham argues that universities should not create an entrepreneurship curriculum; they should teach subjects that make students good at building things, make starting a company feel viable, and encourage students to work on their own projects. [13]
- **Why it matters:** The project prescription is unusually concrete: projects deepen subject knowledge, reveal potential cofounders, make self-direction natural, and generate ideas that would look implausible on paper. Graham contrasts that with business-plan competitions, arguing that founders should build prototypes for users rather than practice stories for investors. [13]

A China-tech reading stack

Sinocism, China Talk, Hello China Tech, and *Breakneck*

- **Type / creators:** Three China-focused blogs/newsletters—Sinocism, China Talk, and Hello China Tech—plus Dan Wang’s book *Breakneck: China’s Quest to Engineer the Future*.
- **Recommended by:** Packy McCormick, who calls the three publications “great China-focused blogs for western audiences” and the book “great.” [14]
- **Key takeaway:** Packy’s rationale is contextual: readers tracking energy, AI, hardware, and global competition need a better understanding

of China and should use sources that explain it from more than a cartoonish adversarial frame. [14]

- **Why it matters:** This is a practical starter stack for correcting a U.S.-only view of technology competition, rather than a recommendation to consume another general futurist essay.

A primary source with a founding consequence

Bitcoin white paper

- **Type / creator:** Technical paper; Satoshi Nakamoto.
- **Recommended by:** Brian Armstrong, whose endorsement comes as an origin story: after first reading the white paper while working at Airbnb, he spent nights and weekends building the prototype that eventually became Coinbase, then carried Airbnb's practices into the new company. [15]
- **Key takeaway:** Armstrong presents the paper as a trigger for moving from curiosity to a working prototype.
- **Why it matters:** It is a rare founder-level example of a primary technical document producing action, not just changing an opinion. [15]

Sources

1. X post by @levie
2. X article by @nayakkayak
3. X post by @nayakkayak
4. X post by @eladgil
5. OpenAI Jalapeño: Better Than Nvidia Blackwell
6. X post by @patrick_oshag
7. Ex-NVIDIA Engineer: Why AI Is About to Get 1000x Cheaper
8. X post by @0xCodez
9. X post by @martin_casado
10. Inside Vercel's AI-era playbook: "Recursive Founder Mode" | Guillermo Rauch (CEO)
11. X post by @rrhoover
12. X post by @rrhoover
13. How Universities Should Prepare Founders
14. An American Engineer in China
15. The Man Behind Coinbase: Interview with Brian Armstrong | KMP Ep.52