AI - LLMs, Scaling and Compute

AI / LLMs / Scaling Laws / Compute — Compiled Node

> **What this is.** A compiled synthesis of K's AI learning across the vault — not a market view and not Claude's own knowledge. It organizes what K has captured and written, separates the competing intellectual camps, flags where sources directly contradict each other, and lists what the vault does *not* yet contain. Closed-book over the vault: no web, no model-training facts injected. Where a source is K's own synthesis doc, that is marked.

Provenance key

Tag	Meaning	Date
`[PRIMER-K]`	K's own synthesis doc — Neuro-Symbolic AI Investor Primer	2026-06-16
`[CANON-K]`	K's own survey — "Machines Cannot Equal Humans" / Skeptical Canon	2026-06-16
`[WG-K]`	inPractise AI working-group synthesis — K was a participant (semi-primary)	2026-06-20
`[SUB]`	Substack synthesis doc in the vault (Vibhav Viswanathan, "Alpha & Agents")	2026-06
`[ART]`	Readwise highlight of an external article	dated inline
`[VID]`	YouTube video summary (post-reorg: per-video note under `Notes/`)	dated inline
`[TXT]`	YouTube transcript (verbatim)	undated/recent
`[STUB]`	Vault page exists but is empty / unreadable	—

> **Reliability ladder (unchanged).** A figure from a formal document outranks a number a founder/CEO said on a podcast. Almost all of the 2026 compute-economics datapoints below are `[VID]` — verbal claims on interviews, transcribed by Gemini. Treat them as *directional management/industry claims*, not audited figures. They are compiled faithfully and tagged so K can see the register a year later.

---

1. The central debate: does scaling get us there?

This is the spine of the whole corpus. The vault holds both sides explicitly.

The scaling/compute-maximalist pole — "The Bitter Lesson." Rich Sutton's thesis, captured in *The Bitter Religion* `[ART, Michael Dempsey / The Generalist, Nov 2024]`: "*general methods that leverage computation are ultimately the most effective, and by a large margin.*" Models improve because compute gets exponentially cheaper (Moore's Law); effort spent adding human knowledge or narrow tooling is, in Sutton's view, wasted — "fiddling with the fins on your surfboard as a surge gathers." The article frames AI optimism as quasi-religious ("theogenesis"), hence "The Bitter Religion."

The scaling-skeptic pole — "we need new research." Same article: Ilya Sutskever (and, per leaks, others inside OpenAI) believe scaling is "approaching a ceiling" and that *novel science* is required to reach AGI. This is corroborated first-hand in the Sutskever transcript `[TXT, "We're moving from the age of scaling to the age of research"]`: pre-training was "*a particular scaling recipe*"; "scaling" became powerful precisely as a *word* that told everyone what to do — and he now frames the field as moving past it toward research.

The bridge both camps point to — test-time compute. Rather than scaling *training*, dedicate more compute at *execution* ("the equivalent of bringing a calculator and an extra hour to the exam"). The frontier shifts from pre-training toward post-training / inference / RL. `[ART, Bitter Religion]`

2026 update — an insider says scaling is "alive and well." Anthropic CFO Krishna Rao, mid-2026: "*scaling laws are alive and well — no plateau in sight,*" citing recent releases; he concedes skepticism is natural for any research lab but says the data shows scaling isn't slowing `[VID, Krishna Rao, 2026-05-13]`. Jensen Huang implies the demand-side corollary: agentic AI takes **~1,000× more compute** than early generative LLMs `[VID, Jensen Huang, 2026-05-05]`. On the *supply* side of scaling, Dylan Patel notes the frontier model "Mythos/Methos" was "*a materially larger model… it's the scale*" and the biggest capability step in ~2 years `[VID, Dylan Patel, 2026-04-27]`. ⚠️ Note the register: these are builder/insider claims made while raising and deploying capital — they sit opposite Sutskever's "age of research" and Mercor's "hitting scaling-law constraints." Both live in the vault; see Conflict Log #4.

The open verdict (K highlighted this line). "*I sometimes wonder what will happen if scaling laws don't hold … I also sometimes wonder if scaling laws perfectly hold and if that will look similar to commoditization curves for first movers. Either way we're going to light a ton of money on fire finding out.*" `[ART, Bitter Religion]`

2. Emergence — and the measurement caveat

From *Emergent Abilities of LLMs* `[ART, assemblyai, May 2023]`: as LLMs scale they cross critical thresholds where new abilities (arithmetic, Q&A, summarization) appear "suddenly, as if out of thin air" — not directly trained. The note pegs a critical regime around **10^{10}–10^{11} effective parameters**, analogizing to phase transitions in physics.

The caveat the same note records — important, and it pre-stages the later "is emergence a mirage?" debate: **emergence is metric-dependent.** Swap the evaluation from exact-string-match to multiple-choice and the discontinuity smooths into gradual improvement — "the emergent behavior disappears." So some emergence is a property of the *measurement*, not the model.

Data ceilings (same source): a 100T-parameter model would need ~190× GPT-3's data per OpenAI's scaling law; Common Crawl in its entirety is ~12 PB; only a fraction is stored, textual, and trainable. Altman dismissed 100T-param claims for this reason.

3. The limits critique — neuro-symbolic and the Causal Ladder

K's own primer `[PRIMER-K]` is the most developed node here.

- **System 1 / System 2 (Kahneman).** Neural nets = fast pattern-matching (System 1); symbolic AI = slow, traceable logic (System 2). Neuro-symbolic AI wires both together for systems that perceive *and* reason auditably.
- **Frontier LLMs are NOT symbolic.** GPT-5 / Claude Opus / Gemini are "purely neural… no logic engine, no rule system." Chain-of-thought is *"simulated inference, not inference itself"* — statistically-likely-looking reasoning, which can flip under unusual framing. Supporting datapoint K cites: a small LM + **Z3 formal solver** matched large CoT models *at a fraction of the compute* — evidence that LLM "reasoning" is inefficient mimicry.
- **Pearl's Causal Ladder.** Rung 1 Association (all of deep learning, however large), Rung 2 Intervention (do-calculus), Rung 3 Counterfactual. Pearl's "devastating critique": DL is *permanently stuck on Rung 1* — association can't, in principle, yield causation. Bengio's program (Causal Representation Learning, disentangled representations) tries to *climb* the ladder through learning; Pearl is skeptical that observational data alone ever can.
- **Pearl in his own words (new primary source).** The vault now holds Pearl's own talk `[VID, Judea Pearl – The Science of Cause and Effect, 2026-04-13]`, which grounds the ladder the primer describes: causal science uniquely fuses *data + partial model + a specific query*; the equality sign must be replaced by an **asymmetrical assignment operator** to capture cause; **counterfactuals** ("would Kennedy be alive if Oswald hadn't fired"; "will I regret quitting") are the atomic unit of Rung 3; **do-calculus** lets you predict intervention effects from observational data *when identifiable*; and **selection diagrams** enable transfer learning across environments. Pearl frames the live-2026 relevance directly: LLM-generated text raises *fairness/bias* questions that are fundamentally about **direct vs. indirect effects** (e.g. "is a hiring disparity the direct effect of gender, or mediated by qualification?") — exactly the mediation questions association-only systems cannot answer.
- **K's investment framing** `[PRIMER-K]`: enterprise value flows to systems that reach Rung 2; the accuracy-vs-auditability gap is the business opportunity in regulated industries (health,

finance, law, insurance). "Every sector that cannot accept a black box is a potential customer for neuro-symbolic infrastructure."

4. The skeptical canon (K's survey) `[CANON-K]`

K's survey of "machines cannot equal humans" thinkers, with the explicit finding that **most skeptics have not recanted** in the LLM era:

- **Dreyfus** — embodiment/tacit know-how; vindicated on GOFAI's collapse, conceded connectionism sidesteps part of his critique, held firm on embodiment. (Died 2017, never saw LLMs.)
- **Searle** — Chinese Room; syntax ≠ semantics; never abandoned it. "LLMs are the Chinese Room made real at scale."
- **Penrose / Lucas** — Gödelian: human insight is non-computable, hence mind ≠ algorithm; Penrose's further leap to quantum (Orch-OR) is "the most contested item in the survey." Reaffirmed in a 2026 interview.
- **Marcus** — the contemporary *empirical* critic; same neurosymbolic thesis since 1998. Leans on Apple's 2025 "*Illusion of Thinking*" (RL reasoners still fail out-of-distribution, e.g. Tower of Hanoi). **K flags the Apple paper is itself contested** (output-length limits) and that Marcus is "a partisan in a live dispute, not a neutral referee."
- **Chomsky** — explanation vs description; LLMs are "lumbering statistical engines… stuck in a prehuman phase." No reversal.
- **Bender/Gebru** — "stochastic parrots"; meaning needs communicative intent; reaffirmed after GPT-4.
- **The reversal cutting the other way — Hinton.** A *builder* who moved toward "they understand / may already be conscious," reasoning from copyability/bandwidth (digital minds merge weight updates, ~trillion bits, vs human language at ~10 bits/sec).

K's synthesis line: the field is a **three-way split** — builders move toward understanding/consciousness, philosophical skeptics hold firm, and Marcus sits in the middle arguing the *current architecture* won't get there regardless.

5. The builder's frame — Karpathy, software, verifiability

From the video summaries `[VID]`:

- **Software 1.0 → 2.0 → 3.0**: explicit rules → learned weights → *prompting*. LLMs are a new kind of programmable computer.
- **Verifiability is the axis of automation.** "Traditional computers automate what you can specify in code; LLMs automate what you can *verify*." Where outputs can be auto-verified, throw RL at it and it progresses fastest.

- **Vibe coding vs agentic engineering**: vibe coding *raises the floor* (programming for everyone); agentic engineering *preserves the quality bar* of professional software while going faster.
- "**Summoning ghosts, not building animals**"; agents are like *interns* — you still own aesthetics, judgment, taste, oversight. Extreme "**jaggedness**": refactors a 100k-line codebase yet tells you to *walk* to a car wash 50m away.

6. Compute economics — the AI supercycle

From the Stanford "Economics of the AI Supercycle" class summary `[VID, Apoorv Agrawal, Stanford Spring '26]`:

- **The "five-layer cake"** (Jensen's framing): energy → chips → power interconnects → memory → data centers.
- **Inverted-triangle value chain.** Unlike the cloud/software pyramid, AI value today is concentrated at the *infrastructure* base, not the application top. Open question whether/when it inverts.
- **Not zero marginal cost.** AI inference runs on expensive compute — a fundamental break from software economics.
- **Nvidia ~75% gross margin** on data-center revenue; structural dominance. ~**$300bn of revenue** to fight over, ~half from hyperscalers.
- **Consumer monetization**: ChatGPT ~1bn users monetized at ~$10; "how do we get from 1bn to 4bn — not sure knowledge work is the answer." Ads likely beat subscriptions long-term.
- **The most sought-after unknown**: Nvidia's *inference vs training* fleet share — the "timing mismatch" question (building 5-year semis capacity against revenue that hasn't arrived).
- **Mercor** `[VID]`: models are hitting scaling-law constraints; progress now needs *curated datasets from top-tier domain experts*, power-law distributed. Models still need humans to train/verify — "a very long road" to that ending.

6b. The 2026 compute-economics update (delta 2026-07-02)

A cluster of 2026 interviews sharply enriches — and partly re-prices — the picture above. All `[VID]`, verbal claims; dated inline.

Demand / revenue run-rate.

- Anthropic went from **~$9B ARR at the start of 2026 to >$30B by end of Q** `[VID, Krishna Rao, 2026-05-13]`; Dylan Patel puts it at "*$35–40B, probably $40–45B by the time this airs*" — while flagging that **compute has not grown to the same degree** as revenue `[VID, Dylan Patel, 2026-04-27]`. Gavin Baker's version: Anthropic **added ~$11B of annualized**

run-rate in a single month, "matching ten years of cloud business building" `[VID, Gavin Baker, 2026-05-20]`.

- Nvidia guided to **~$91B quarterly revenue** vs the ~$11B guide of May 2023 that "initiated the boom"; DC revenue splits **~50/50 hyperscaler vs. enterprise tail** (Rasgon uses this to rebut concentration fears); Nvidia projects **~$20B of CPU revenue this year — ≈ Intel + AMD's entire DC chip business combined**; hyperscaler capex cited at **Google ~$180B, Amazon ~$220B** `[VID, Stacy Rasgon, 2026-06-08]`. (⚠️ This eclipses the node's earlier "~$300bn to fight over" figure from the Stanford class — kept as a time-series, not overwritten: ~April framing = whole-industry pool; ~June framing = Nvidia alone implying ~$364B annualized. See Conflict Log #5.)

Cost / token trajectory.

- Tokens are expected to fall to **~1/10 of current cost long-term, triggering an explosion in enterprise consumption** `[VID, Nikesh Arora, 2026-06-22]`. Krishna Rao rejects the "faster car burns more fuel" analogy: frontier models raise **capability and token-efficiency simultaneously**, efficiency compounding between releases `[VID]`.
- A vivid supply-management datapoint: **Claude Opus generates ~70% fewer tokens for the same prompt** vs earlier — Gavin Baker reads this as a *deliberate deprecation to manage a compute shortage* `[VID, Gavin Baker, 2026-05-20]`.

Hardware / supply chain / depreciation.

- Bottlenecks have moved *upstream* into HBM, lasers, copper foil, logic — "aggressively pushing gross margins upstream," complex supply chains → longer lead times `[VID, Dylan Patel]`.
- A single **Blackwell rack ≈ 3,000 lbs / ~100 kW**; a liquid-cooled Nvidia rack ≈ **$4–5M with >1M components** `[VID, Gavin Baker; Jensen Huang]`.
- **GPU useful life re-rated to ~10–15 years** (decoupling prefill/inference), and **rising rental prices for older Ampere GPUs** are used to *debunk the short-seller "rapid obsolescence" thesis* `[VID, Gavin Baker; Stacy Rasgon]`. Dot-com analogy inverted: 99% of dark fiber sat unused in 2000; today GPU capacity is fully utilized `[VID, Gavin Baker]`.
- **Nvidia–TSMC operate on a handshake, no contract**; TSMC is the leading-edge gatekeeper; Broadcom = premier custom-ASIC supplier `[VID, Gavin Baker]`.

Capital commitments.

- Anthropic has signed **>$100B of "layer-cake" compute commitments** — ~5-gigawatt deals with Google, Broadcom, and Amazon (TPUs + Trainium), landing infrastructure continuously into 2027 — and runs **three chip platforms fungibly (Trainium / TPU / Nvidia)** to route workloads by availability/economics. Procurement is governed by a "cone of uncertainty"

(long-lead supply vs. volatile demand) with a *non-negotiable floor allocation for model development* `[VID, Krishna Rao, 2026-05-13]`.
- Nvidia placed **~$500B of firm purchase orders** partly to force suppliers to build advanced fabs inside the US; AI-native software startups are *finally reporting positive gross margins* `[VID, Jensen Huang, 2026-05-05]`.

Gatekeeping thesis (Dylan Patel). As tokens aggregate value among fewer, larger spenders, *"who has Mythos? top banks"* — capability gets **gatekept to elite customers who can spend billions on dedicated clusters**, and labs "release to a fewer and fewer set of customers" to prevent distillation `[VID, Dylan Patel, 2026-04-27]`. This is the opposite bet to the "intelligence is commoditizing" thesis in §9 — logged in Conflict Log #7.

7. AGI timelines & AI-for-science

- **Hassabis** `[VID]`: ~"three-quarters of the way to AGI," timeline **2030** (consistent with DeepMind's original 20-years-from-2010 call). AlphaFold = proof AI can solve "root-node" problems and unlock downstream science (Isomorphic → drug discovery from years to weeks). Speculative: *information* may be as fundamental as energy/matter. Build the tool first; consciousness/agency questions second.
- **Context graphs** `[VID, Ashu Garg / Foundation Capital]`: the "missing memory layer" for agents — institutional memory via "decision traces." Agents need to move from single-player chat to multiplayer with state + memory. Claimed as "the next trillion-dollar AI opportunity." (Note the resonance with *this* wiki exercise — compiled persistent context as the layer above raw capture. The 2026 delta makes this a major theme — see §9.)

8. The 2026 capex / bubble debate & software repricing (delta 2026-07-02)

A genuinely new axis the vault did not hold at the last compile: *is the AI build-out a bubble, and where does the software value go?* Three positions coexist, unreconciled.

Bull — "just getting started." Alex Sacerdote: the foundation-model layer has consolidated into a **three-horse oligopoly (OpenAI, Anthropic, Gemini)**; **enterprise AI is <1% penetrated** on a near-vertical adoption curve; coding tools triggered a 2025 enterprise shift (developers spending ~$100/day on tokens); the hardware stack is **de-commoditizing** (liquid-cooling at physical limits, thick copper + bendable fiber). His fund shorted legacy application software as IT budgets divert to tokens `[VID, Alex Sacerdote, 2026-06-12]`. Jensen (1,000× compute) and Krishna Rao (returns to the frontier "remarkably high") sit on this side.

Bear — "this is a bubble." Seth Klarman: the market shows *"bubbly characteristics"* — extreme multiples paid on an unpredictable distant future; **valuations for leading foundation-model firms are "fundamentally unproven" because they remain dependent on eating massive cash**

inputs. He is playing it via *optionality* (buying land with power access for future data-center sites) and distressed credit, not by chasing the leaders `[VID, Seth Klarman, 2026-06-12]`. Marc Rowan supplies the macro strain: **~$800B of capex from just four public tech firms this year**; **10 stocks ≈ 50% of the S&P 500** (concentration risk for retirement systems); and — pointedly — "*disastrous PE returns are expected from recent legacy-software vintages*" because managers paid premium prices that failed to anticipate AI disruption `[VID, Marc Rowan, 2026-05-27]`.

Value-migration / repricing (the "so what" both sides half-share). AI applications have **net-destroyed billions of dollars of software market value while infrastructure providers capture the profits** `[VID, Gavin Baker]`. System-of-record incumbents (Salesforce et al.) face valuation pressure as software shifts to "opinionated networks" `[VID, Nikesh Arora]`. Harshil Mathur's blunt version: **"AI is compressing every moat" — execution speed becomes the only moat**, because AI collapses the technical build `[VID, Harshil Mathur, 2026-05-06]`.

> **K's own datapoint on the register of these claims:** Klarman notes a consumer footwear brand (Allbirds) drew "*bubble hype after adding AI terminology to its name*" `[VID, Klarman]` — a reminder that the label is being priced, not just the capability.

9. The agentic value chain & "context as the moat" (delta 2026-07-02)

The richest new source is Vibhav Viswanathan's five-post "Alpha & Agents" arc `[SUB, May–Jun 2026]` (CEO, Pascal AI Labs; deployed across institutions ~$1T AUM). It supplies a structured value-chain map and a cost model, and it rhymes exactly with the "context graphs / memory layer" thread in §7 and with *this wiki's own thesis*.

The core equation: `Performance = Intelligence × Context`. Intelligence is commoditizing — everyone gets the same API. The **durable edge is proprietary context**: house views, surviving memos, PM sizing notes, the institutional record of what worked across regimes. "Firms that understand this are not building AI tools — they are building institutional memory that happens to run on AI." Nikesh Arora's independent echo: "*memory becoming the moat,*" with foundation labs "baking user memory and context directly into their networks" (captive models) `[VID, Nikesh Arora, 2026-06-22]`.

The six-layer agentic value chain (bottom→top) `[SUB]`:

1. **Compute/chips** — Nvidia dominates training; inference commoditizing fast (Google TPU v7, Microsoft Maia 2 collapsing rental yields 60%+).
2. **Foundation models** — practical **duopoly: Anthropic + OpenAI ≈ 89% of startup revenue**; both on a "capital treadmill." Claude Code ~50% of developer workflows.

3. **Infrastructure/tooling** — *most structurally defensible* (orchestration scaffolding; compliance lock-in via EU AI Act, DORA, ISO 42001).
4. **Application layer** — where most VC flows *and won't return*; squeezed from above (models absorb features) and below (incumbents add agentic features at ~zero cost). **Cursor cautionary tale: -30% gross margin, paying Anthropic ~$650M/yr against ~$500M revenue.**
5. **System-of-record incumbents** — binary fate: become the native agent platform or degrade into a passive DB read via API ("double-payment trap" live now).
6. **Distribution/integration** — SIs (Accenture, TCS, Infosys) as last-mile; Microsoft shipped 200K+ agentic licenses through partners.
   Durable rents accrue in **proprietary data with exclusive rights, deep system-of-record integration, and complex regulatory positioning**; value evaporates in thin wrappers, in the models themselves (open-weights approach parity), and in commodity compute.

Token economics — "token yield." Agentic loops re-transmit the whole context each step (LLMs are stateless), so cost balloons ("tokenmaxxing"). The metric that matters is **token yield = business-aligned output ÷ tokens consumed** — set by *architecture, not model* (two firms, same model, same filings can differ 10×). Four levers: context precision (surgical retrieval), model routing (cheap models for mundane steps), memory/caching (prompt + semantic caching breaks the linear cost curve), long-workflow compression `[SUB]`. Pascal's own stack (Sovereign Data Fabric → Cognitive Engine: *Lattice* knowledge graph, *Prism* document intelligence, *Shell* permissioning → agents → Excel/PPT workspace, VPC-deployed) is the reference implementation, and the three "2026 infrastructure problems" are **signal acquisition, signal fragmentation/synthesis, and organisational memory** — the last being the memos/post-mortems/rejected-theses that firms discard at trade execution and lose when an analyst leaves.

10. The investor's frame — the inPractise working group (delta 2026-07-02)

`[WG-K, 2026-06-20]` — a six-person working-group call (William Oliver, Ramesh Narayanaswamy, Hugo Montagne, Josh Tarasoff, **Karthik Tiruvarur**, Michael Gallagher). Because K is a participant, this is semi-primary K thinking, and it maps the *investment-practice* consequences of the whole §1–§5 debate.

- **Ramesh Narayanaswamy** (spine of the call): reality is *reflexive* (Soros) and can't be explicitly coded; **reason is the map, intuition is contact with the territory**. LLMs need not *understand* to be useful — simple verifiable metrics (ROCE, gross-margin stability, low leverage) get ~80% of the way, as factor investing did. But the Chinese Room holds ("they impress, looking"); his falsifiable challenge: **has any LLM produced genuinely new *explanatory* knowledge? He thinks not yet.** The "fragility gap": LLMs learn *how you think* from your memos well, but detect company *fragility* poorly — fragility isn't in the training set.

- **Josh Tarasoff**: public markets are *by design* the arena meant to be commoditized by AI (same information by law) — the residual edge is *irreducibly human* judgment and wild valuation swings. **Tacit knowledge = direct experiential contact** (meeting the CEO counts; an ex-employee interview is a symbolic proxy and doesn't). Lean toward tacit-heavy companies.
- **K's own contributions** `[WG-K]`: the **freed-time / Red Queen question** — if AI compresses 3–4 days into 3–4 hours and *everyone* gets the same superpower, the bar rises → a "Cambrian explosion" of approaches; what do you do with the unlocked time? He supplies the **literature spine** (Sutton's Bitter Lesson; Pearl's ladder as "what separates LLMs from analysts"; a call to review Hinton/Pearl) and the **vertical-vs-horizontal / harness-vs-proprietary-weights** worry: *if context + harness is "all there is," a fund collapses into a sellable skills file* — but **path dependence** (*Deep Simplicity*, *Age of Average*, Hindu philosophy) means identical inputs don't yield identical outputs.
- **Hugo Montagne** (practitioner): models **converge on obvious problems, diverge on taste-defined ones** — and that divergence (temperature/sampling) is what makes them useful; **the harness is load-bearing** — without strong constraints, entropy degrades output. Agnostic on "understanding," focused on elicitation.
- **William Oliver**: three ways of knowing (McGilchrist: reason, intuition, imagination); LLM amplification of gestalt intuition can only ever be marginal; spend AI-freed time "at the edge" — trust, relationships, subtle reality.
- **Michael Gallagher**: **surprise as the core construct** — dopamine encodes *reward-prediction error*, not reward, so surprise is wired into both neuroanatomy and RL; the **RLHF → RLVR shift** as the recent "Cambrian" trigger; the transformer's "shocking simplicity"; cites Max Bennett, *A Brief History of Intelligence*.

The five open synthesis questions the group set: (1) where does AI-freed time earn the highest return; (2) what is irreducibly human in investment judgment, and is it shrinking; (3) can we taxonomize companies by explicit↔tacit knowledge; (4) does edge accrue to harness-engineering or proprietary-weights; (5) has any LLM produced genuinely new explanatory knowledge — can "creativity as frame-collision + surprise" give a falsifiable test.

11. Conflict / tension log — do not collapse these

1. **Bitter Lesson vs neuro-symbolic.** Diametrically opposed. Sutton: human knowledge / structure is a *waste*, compute wins. Marcus/Pearl/ `[PRIMER-K]`: you *cannot* get reliable reasoning without symbolic structure. Both are held with conviction in the vault. ⚠️ Not reconcilable — they are competing bets on the same question.
2. **"Do LLMs reason?"** Builders (Karpathy: useful ghosts; Hinton: they understand) vs critics (Searle/Chomsky/Bender/Marcus: mimicry, syntax≠semantics; Ramesh's "impress, looking"). The empirical referee (Apple "Illusion of Thinking") is itself disputed — K flagged this.

3. **Is emergence real or a measurement artifact?** The Emergent-Abilities note contains its own rebuttal (metric-dependence). Don't cite emergence as settled.
4. **Scaling continues vs plateaued.** Now sharper with 2026 data. *Continues:* Krishna Rao ("scaling laws alive and well, no plateau"), Jensen (1,000× compute demand), Dylan Patel (Mythos = biggest step in 2 years). *Plateaued / needs new science:* Sutskever ("age of research"), Mercor ("hitting scaling-law constraints; needs curated expert data"). Even the "test-time compute" bridge is contested as either *extending* the curve or merely *shifting the cost*. `[VID/TXT/ART]`
5. **Where AI value accrues — and how big the pool is.** (a) *Locus:* infra/Nvidia captures it now (inverted triangle; Gavin Baker "software value destroyed, infra captures") vs the "context/memory is the moat" thesis (Vibhav, Nikesh) vs eventual application-layer inversion (which Vibhav argues *won't* pay off for most apps). (b) *Magnitude, as a time-series not a contradiction:* Stanford class ~"$300bn industry pool to fight over" (~Apr) vs Rasgon's Nvidia-alone ~$91B/quarter guide (~Jun). Recorded as evolution.
6. **AGI by 2030 (Hassabis, optimistic builder) vs "current architecture never gets there" (Marcus/skeptics).**
7. **Is intelligence commoditizing, or gatekept? (NEW)** *Commoditizing:* Vibhav ("intelligence is the commodity; context is the differentiator"; open-weights approach parity) and Nikesh (tokens → 1/10 cost). *Gatekept/concentrating:* Dylan Patel (frontier capability released "to a fewer and fewer set of customers"; "who has Mythos? top banks"; billion-dollar dedicated clusters). These are opposite bets on the same variable.
8. **Is the AI build-out a bubble? (NEW)** Klarman ("this is a bubble," foundation-model valuations "fundamentally unproven") and Rowan (~$800B/4-firm capex, legacy-software vintages face "disastrous returns," S&P concentration) vs Sacerdote ("just getting started," enterprise <1% penetrated) and Jensen/Krishna Rao (returns to frontier "remarkably high"). Unresolved; the vault holds both with conviction.
9. **Nvidia obsolescence (NEW).** Short-seller "GPUs depreciate fast" thesis vs Rasgon/Baker (useful life 10–15 yrs; *rising* Ampere rental prices; full utilization vs dot-com dark fiber). The vault records only the rebuttal side in detail — the bear case is referenced, not sourced (see §12).

12. Disclosure-gap catalogue — what the vault does NOT (yet) contain

- `[STUB]` `LLMs.md` **is empty** — a referenced page with no content.
- `[STUB]` `AI Notes.md` **DeepSeek entry (Jan 28 2025) is corrupted** — the Roam→Obsidian export left only garbled PDF-highlight encoding; the DeepSeek content is effectively unreadable and should be re-captured.
- **No primary scaling-laws papers.** Kaplan et al. 2020 and Hoffmann et al. 2022 (Chinchilla) are *not* in the vault — only secondary treatments (Bitter Religion, Emergent Abilities blog).

The compute-optimal "Chinchilla" result is referenced nowhere directly. (Partial progress: Pearl's *own* talk is now in the vault for the causal-ladder strand — §3.)

- **The AI-capex *bear* case is under-sourced relative to the bull case.** Rasgon/Baker rebut the short thesis in detail, but the vault holds no primary short-seller write-up. Klarman/Rowan give the macro-bubble worry but not a bottoms-up teardown.
- **Reorg note (2026-06-25).** The two summary reports this node was originally built from now live at `Sources/Transcripts/Video_Summaries_Report.md` and `Sources/Transcripts/March and April 2026 Videos.md`; the per-video content has been split into individual `Notes/*.md` files (see `Notes/_YouTube_Index.md`). K's two primers moved to `Sources/Readwise/`. Paths updated in the source digest.
- **Indexed but NOT deep-ingested this pass** (listed per Hard Rule 5 so nothing is silently dropped) — on-topic videos/notes that would enrich a future update:
  - *Builder/AGI:* Sundar Pichai — *History and Future of AI at Google* (2026-04-23); Demis Hassabis — *The Future of Intelligence / Three Quarters of the Way to AGI* (2026-05-01, new detail beyond §7); Greg Brockman — *Human Attention Is the New Bottleneck* (2026-05-01); Jack Dorsey — *Every Company Can Be a Mini-AGI* (2026-04-04); Karpathy — *Skill Issue: Code Agents* (2026-03-22).
  - *Compute/markets:* Gavin Baker — *Inside the Mind of a Tech Investor* (2026-05-15); Bloomberg Tech OpenAI clips (2026-04-27/28); MSFT/Alphabet/Amazon Q1-2026 earnings calls (AI-capex commentary).
  - *Enterprise agents:* Sridhar Ramaswami — *Snowflake: Reliable AI* (2026-04-03) & *How AI Agents Will Transform the Workplace* (2026-06-19); Ex-Third Point Data Scientist — *AI Agents Are Encoding PM Workflows*; Wilson Chan — *Agentic AI in Finance*; *AI vs Human Intuition — Techtopia Ep 24*.
  - *Skeptical-canon adjacent:* Patricia Churchland — *Touching a Nerve / The Self as Brain* (2026-04-30); Patrick Winston — *How to Speak* (2026-04-29).
  - Still un-ingested from the 2026-06-17 list: full transcripts of *Gary Marcus on LLM Scaling*, *Shane Legg — arrival of AGI*, *Fei-Fei Li — spatial intelligence*, *Daniel Guetta — why LLMs hallucinate*; and the `.docx` readings *AI RELATED READING*, *AI and Finance Related Reading*, *ai_inflection_point*, *Software Related Reading*.

13. Open questions (carried from the vault)

- Will pre-training scaling laws hold, and does test-time compute extend the curve or just relocate the cost? `[ART/TXT/VID]`
- Can neural nets *learn* their way to Pearl's Rung 2 (Bengio), or is hand-coded symbolic structure unavoidable (Pearl)? `[PRIMER-K/VID]`
- When/whether the AI value chain inverts toward applications; training-vs-inference share of compute. `[VID]`

- Path from ~1bn to ~4bn monetized users — "not sure knowledge work is the answer." `[VID]`
- Is the apparent reasoning of frontier models enough to trust in high-stakes domains without an external symbolic layer? `[PRIMER-K]` `[CANON-K]`
- **Is intelligence commoditizing (context = moat) or concentrating (capability gatekept to elite spenders)?** `[SUB/VID]`
- **Is the ~$800B/yr AI capex a bubble, and will legacy-software vintages be repriced?** `[VID]`
- **What does the investor do with AI-freed time (Red Queen), and what is irreducibly human in judgment?** `[WG-K]`
- **Has any LLM produced genuinely new *explanatory* knowledge — and can "frame-collision + surprise" make that falsifiable?** `[WG-K]`

Linked vault nodes

AI · AI Notes · LLMs · Artificial Intelligence · Neural Networks · Foundational models · FrontierModels · Generative AI · Hyperscalers · OpenAI · chatGPT · What Is ChatGPT Doing … and Why Does It Work · 2026-06-16 Neuro-Symbolic AI Investor Primer · 2026-06-16 Machines Cannot Equal Humans_SkepticalCanon · scale · Demis Hassabis · Gary Marcus · Judea Pearl · Yoshua Bengio · Vibhav Viswanathan · Pascal AI · inPractise · Josh Tarasoff · William Oliver · Max Bennett · AI and investing

Source-node digest (with dates / type)

- *The Bitter Religion: AI's Holy War Over Scaling Laws* — `[ART]` Michael Dempsey, The Generalist, **Nov 2024**. Sutton's Bitter Lesson, Sutskever ceiling, test-time compute, the open bet.
- *Emergent Abilities of LLMs* — `[ART]` assemblyai, **May 2023**. Critical-scale thresholds, metric-dependence, data ceilings.
- *Neuro-Symbolic AI — Investor Primer* — `[PRIMER-K]` K, **2026-06-16** (`Sources/Readwise/`). System 1/2, Pearl's ladder, Bengio, "frontier LLMs aren't symbolic," Z3, investment framing.
- *The Case That Machines Cannot Equal Humans (Skeptical Canon)* — `[CANON-K]` K, **2026-06-16** (`Sources/Readwise/`). Dreyfus, Searle, Penrose, Marcus, Chomsky, Bender/Gebru, Hinton reversal.
- *Video_Summaries_Report.md* & *March and April 2026 Videos.md* — `[VID]` **2026** (now `Sources/Transcripts/`). Karpathy (Software 3.0), Stanford AI-economics, Hassabis, Mercor, context graphs.
- *Ilya Sutskever — age of scaling → age of research* — `[TXT]` recent. Pre-training as "a scaling recipe"; shift to research.
- *AI Notes.md* (DeepSeek, Jan 2025) — `[STUB]` corrupted export.

- **— Delta 2026-07-02 additions —**
- *Dylan Patel — The Supply and Demand of AI Tokens* — `[VID]` **2026-04-27**. Tokenomics, HBM/laser/copper bottlenecks, Anthropic ARR≫compute, capability gatekeeping, depreciation.
- *Judea Pearl — The Science of Cause and Effect* — `[VID]` **2026-04-13**. Do-calculus, three rungs, counterfactuals, mediation/fairness. (Primary source for §3.)
- *Krishna Rao (Anthropic CFO) — Inside the $100B Compute Commitment* — `[VID]` **2026-05-13**. Scaling "alive," $9B→$30B ARR, 5GW/$100B deals, chip fungibility, cone of uncertainty.
- *Jensen Huang — Leading in the Age of AI* — `[VID]` **2026-05-05**. 1,000× agentic compute, $4–5M racks, $500B POs, startups reaching +gross margin.
- *Gavin Baker — Watts, Wafers, and the Future of AI Infra* — `[VID]` **2026-05-20**. +$11B ARR/month, 70%-fewer-tokens deprecation, 3,000-lb/100kW racks, 10–15yr GPU life, Nvidia–TSMC handshake, software value destroyed.
- *Stacy Rasgon — The AI Semiconductor Boom and What Could End It* — `[VID]` **2026-06-08**. Nvidia $91B/q, 50/50 DC split, $20B CPU rev, hyperscaler capex, Ampere rentals rebut obsolescence.
- *Nikesh Arora — Future of Token Costs / Memory Moat* — `[VID]` **2026-06-22**. Tokens →1/10, captive/memory models, enterprise depth vs consumer breadth.
- *Seth Klarman — This Is a Bubble* — `[VID]` **2026-06-12**. Foundation-model valuations "unproven," land/power optionality, "AI in the name" hype.
- *Alex Sacerdote — Why the AI Boom Is Just Getting Started* — `[VID]` **2026-06-12**. 3-horse model oligopoly, enterprise <1% penetrated, hardware de-commoditizing, shorting legacy software.
- *Marc Rowan — Private Markets, Software Repricing & Capital Allocation* — `[VID]` **2026-05-27**. ~$800B/4-firm capex, S&P concentration, "disastrous" legacy-software PE vintages.
- *Harshil Mathur — AI Is Compressing Every Moat* — `[VID]` **2026-05-06**. Execution speed as the only remaining moat.
- *inPractise AI Working Group Synthesis* — `[WG-K]` **2026-06-20**. Investor frame: reason/intuition, tacit knowledge, freed-time/Red Queen, harness-vs-weights, surprise/RLVR. K participated.
- *Vibhav Viswanathan — "Alpha & Agents" (5 posts)* — `[SUB]` **May–Jun 2026**. Performance = Intelligence × Context, 6-layer agentic value chain, token yield, memory moat, Pascal stack.

Changelog

- **2026-07-02** — Delta update (Claude, Opus 4.8). Integrated 13 new on-topic sources from the post-reorg `Notes/` folder + 2 synthesis docs. Added §6b (2026 compute-economics update), §8 (capex/bubble debate & software repricing), §9 (agentic value chain & "context as moat"), §10 (inPractise investor frame). Added Pearl's own talk as a primary source in §3 and a "scaling is alive" 2026 datapoint in §1. New Conflict-Log entries #7 (intelligence

commoditizing vs gatekept), #8 (bubble vs boom), #9 (Nvidia obsolescence). New provenance tags `[WG-K]`, `[SUB]`. Updated frontmatter (`latest_source_date` 2026-06-16 → 2026-06-25), source-node paths (reorg: reports → `Sources/Transcripts/`, primers → `Sources/Readwise/`), and the gap catalogue's indexed-not-ingested list. No prior content discarded.

- **2026-06-17** — Initial compile (Claude, Opus 4.7). Pilot node #2 (domain kind). Sections 1–10 from concept pages, the two K primers, the two video summary reports, the Sutskever transcript.

Pilot node #2. Same template as the Lenskart node — proving the structure generalizes from a single private company to a heterogeneous learning domain. Now on its first delta, demonstrating the compounding step.