Punctuated Stagnation

Mahesh Madhav, CPU Architecture Fellow and Media Producer
Ampere Computing, Portland, OR
Architecture 2.0 Workshop (ISCA-53): June 27, 2026

1. The Age of the AI Apprentice

We are living in an age of perceived miracles. Large Language Models (LLMs), trained on the vast expanse of human knowledge, appear superintelligent. Standing on the shoulders of giants, they can synthesize, reason, and create with breathtaking speed. They are the ultimate apprentice, capable of exploring the known universe of ideas, finding novel connections in its corners, and even pushing just beyond its current borders.

This has created a "Great Spike" on the graph of human progress. For decades, the merit of our field—be it knowledge or productivity—has grown exponentially through human effort. Now, AI has injected a near-vertical surge of capability (see Figure 1). This spike is both our greatest opportunity and our most profound threat. My thesis is this: the surge will inevitably plateau as AI exhausts the existing human-generated knowledge it feeds on. The existential question for our field is not if this plateau will be reached, but how long it will take. The answer will determine whether computer architecture continues to innovate or enters a long winter of stagnation.

This may sound like a human-centric perspective, and it is. We are, after all, human. Let me be clear, however: this is not a distress call about individual job security. It is an alarm sounded for the future of technological progress itself. The risk is not that architects become unemployed, but that computer architecture itself ceases to advance.

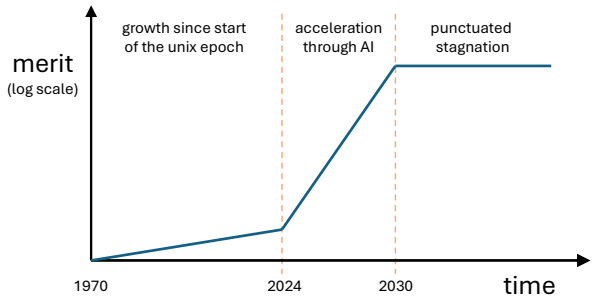

Fig. 1: Apparent progress of knowledge over time, projected

2. The Currency of Suffering

The cartoonist Matthew Inman, "The Oatmeal," observed that the disappointment we feel with AI art comes from the understanding that great human art comes from great suffering [4]. Engineering is no different. Deep knowledge, the kind that pushes a field forward, is not downloaded; it is earned through struggle. It is the currency of suffering that buys intuition.

I have seen this firsthand. At Intel, I remember watching two senior architects work on Crystalwell, the Haswell derivative that featured an embedded DRAM cache on package [6]. This was a high-risk, high-reward project. They poured nights and weekends into designing and verifying the feature in pre-silicon, and later worked odd hours managing chip-in-house and bring-up activities. When they finally got the eDRAM feature working, after years painstaking effort, tears were streaming from their eyes. These two men in their mid-forties were crying rivers of joy, for their feature was functional and their hard work, their belief, and their shared struggle had come to fruition. The intuition they gained in that crucible of effort comes from the subtle interactions, the unexpected failure modes, the gut feel for what to debug next; this cannot be distilled into a dataset.

More recently, at Ampere Computing during the pandemic quarantine, we were in the middle of post-silicon validation for the Ampere Altra SoC. This work requires people to be physically present in the lab. In our Vietnam office, a dozen engineers volunteered to be sequestered onsite to lead physical debug. They ate, slept, and worked inside the building for three months, not once leaving. These employees were critical to Ampere's ability to execute and deliver a functioning product. To help them through the mental strain of isolation, the rest of the company, including senior leaders, rallied in support. We held video chats and played board games with them remotely. They suffered through the quarantine, but in doing so, they gained a great deal of knowledge and formed a fellowship that will last a lifetime. This is how intuition is built. It is the scar tissue of experience. It is the non-verbal, non-quantifiable wisdom that knows when a result feels wrong or when a design path, though logically sound, will lead to ruin.

3. The Two Wings of an Architect

To navigate the complexities of our field, a great architect needs two wings to fly. One wing is data-driven analysis: the rigorous, methodical exploration of the design space through simulation, regression, and measurement. The other wing is intuition: the hard-won, experience-based judgment that allows for creative leaps and the wisdom to know which data actually matters.

A project with only the wing of data becomes an exercise in endless optimization, flying in ever-tighter circles around

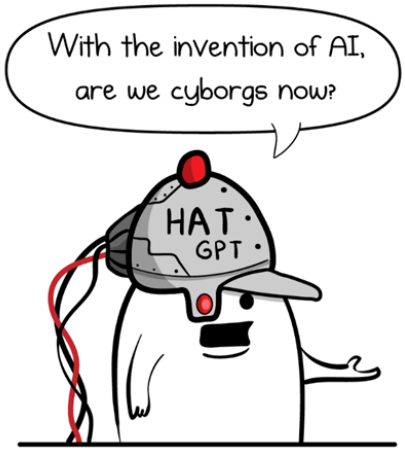

Fig. 2: Drawing by a bonafide Human Artist [4]

a local maximum. A project with only the wing of intuition may take a courageous leap, but without data to ground it, it risks flying blind. True forward progress, the kind that creates new markets and new capabilities, requires both wings working in tandem.

AI is a powerful engine we can strap to the "data" wing. It gives us unprecedented analytical force, capable of running millions of simulations and finding correlations we might never have seen. But an engine without a pilot and a balanced airframe is a recipe for disaster. If the "intuition" wing is weak or atrophies, the project is thrown into an uncontrollable spin, drowning in a vortex of data correlations devoid of causal understanding.

4. The Atrophy of Expertise

The intuition wing, like any muscle, requires constant exercise. If you don't use it, you lose it. The United States continues to build aircraft carriers not just for military strength, but because the only way to retain the institutional knowledge of how to build and maintain them is to keep doing it. When NASA's Apollo program was shuttered, they lost the tacit knowledge and muscle memory of an entire generation [8]. It took fifty years and the dedicated Artemis program to "relearn" how to go to the moon [9].

This is the future we risk with an unguided over-reliance on AI. As Ivan Sutherland has so eloquently shared [14], it takes courage to start a new endeavor. AI lowers the bar for starting, but it doesn't eliminate the need for judgment. If we lose the foundational ability to judge the quality of the work—if we accept AI-generated "slop" without question—our intuition wing begins to atrophy. This is the path to deskilling [2, 3]. We are already seeing a Dunning-Kruger effect ripple through industries, where access to an LLM is mistaken for personal expertise.

5. A Race Against Apathy

We face two possible futures, defined by the duration of this AI-driven productivity spike, or the slope of the second phase of the graph in Figure 1.

Scenario 1: The Short Spike (The Benign Shock). In this future, we reach the AI productivity plateau in a year or two. The limits of AI's ability to innovate on its own become clear quickly. The disruption is absorbed. Students continue to study architecture, knowing that human ingenuity is still the prime mover. Senior architects remain in the field, using AI as a powerful co-pilot as we are"Centaurs and Cyborgs" a taxonomy that industry researchers have proposed recently [10, 11]. Humans do not relinquish the controls and the vital apprenticeship of our industry continues.

Scenario 2: The Long Plateau (The Great Deskilling). This scenario is more dire. A five or six-year wave of "the AI can do that" would be catastrophic. This prolonged period of apparent AI omnipotence would trigger a brain drain [3]. Students, seeing a field seemingly automated, would choose other careers. Senior practitioners would naturally retire. The final attrition would be from mid-career architects, as they would find no reason to stay and babysit an AI. Because when the architects go extinct, the field enters a death spiral. By the time the AI hits its innovation plateau, there will be no architects left with the intuition to push beyond it. The cost of a new chip tapeout—upwards of $10 million for masks alone—is too high a price for an AI to learn from its mistakes without a human guide.

6. The Call to Action

AI cannot create and train on knowledge recursively; the model is prone to lexical explosion and semantic collapse [5, 15]. When trained on its own synthetic input, it forgets what it has learned [13]. Continued use of AI regresses our written content to a terrible mean [12], so that all academic papers start to look and sound the same [7]. The effort to get past unoriginal work is required in order to generate something original [1]. The only source of new, foundational knowledge is the human architect—the pilot who can integrate the thrust from the AI engine with their own earned insight. But if those pilots are gone, the semiconductor industry will be caught in a whirlpool of "punctuated stagnation," trying to move forward but to no avail.

Therefore, our mission is clear: we must reach the AI productivity plateau as fast as possible. We must aggressively use these tools to exhaust the existing knowledge base, quickly revealing their limits and re-asserting the primacy of human creativity. We must shorten this window before our field's capacity for genuine innovation atrophies.

This is a call for computer architects to remain active pilots, not become passive passengers. An architect must know how to fly with both wings. The apprenticeship of new architects—placing them in situations where they must struggle, learn, and build their own intuition—is how we strengthen the intuition wing for the next generation. We must use AI as an instrument, but never forget the feel of the controls that only comes from deep, earned insight. We must race to the top of this curve not to celebrate the power of AI, but to preserve the human practice of compute architecture itself. The future of silicon depends on it.

References

[1] T. Chiang, "Why A.I. Isn't Going to Make Art," *The New Yorker*, 2024. [Online]. Available: https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art

[2] K. Crowston and F. Bolici, "Deskilling and Upskilling with AI Systems," *Information Research an international electronic journal*, vol. 30, no. iConf, p. 1009–1023, Mar. 2025. [Online]. Available: https://publicera.kb.se/ir/article/view/47143

[3] Defiant-Act-7439, "My senior engineers have stopped thinking for themselves," *Reddit*, 2026. [Online]. Available: https://www.reddit.com/r/cscareerquestions/comments/1tkccsi/my_senior_engineers_have_stopped_thinking_for/

[4] M. Inman, "A Cartoonist's Review of AI Art," 2025. [Online]. Available: https://theoatmeal.com/comics/ai_art

[5] W. Kong, S. Lai, J. Piao, and J. Evans, "Multi-LLM Systems Exhibit Robust Semantic Collapse," 2026. [Online]. Available: https://arxiv.org/abs/2605.17193

[6] C. Lam, "Broadwell's eDRAM: VCache before VCache was Cool," 2024. [Online]. Available: https://chipsandcheese.com/p/broadwells-edram-vcache-before-vcache

[7] R. Levesque, "The Shape of Enshittification," *Substack*, 2026. [Online]. Available: https://thedigitalcontrarian.substack.com/p/the-shape-of-enshittification-104

[8] J. M. Logsdon, *After Apollo?: Richard Nixon and the American Space Program*, ser. Palgrave Studies in the History of Science and Technology. New York: Palgrave Macmillan, 2015. [Online]. Available: https://airandspace.si.edu/stories/editorial/why-did-we-stop-going-moon

[9] E. Margolis and G. Clark, "Artemis II's Long Countdown: A space historian explains why it has taken over 50 years to return to the moon," 2026. [Online]. Available: https://phys.org/news/2026-03-artemis-ii-countdown-space-historian.html

[10] S. Randazzo, H. Lifshitz-Assaf, K. Kellogg, F. Dell'Acqua, E. R. Mollick, F. Candelon, and K. R. Lakhani, "Cyborgs, Centaurs and Self-Automators: The Three Modes of Human-GenAI Knowledge Work and Their Implications for Skilling and the Future of Expertise," *Harvard Business School Technology and Operations Mgt. Unit Working Paper No. 26-036*, Dec 2025. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4921696

[11] P. Ranganathan, "Unicorns, Centaurs, and Cyborgs: Co-design Powering the Intelligence Era," in *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ser. ASPLOS '26. New York, NY, USA: Association for Computing Machinery, 2026, p. 2. [Online]. Available: https://doi.org/10.1145/3779212.3795611

[12] J. Russell, R. Rajendhran, C. M. Pham, M. Iyyer, and J. Wieting, "StoryScope: Investigating Idiosyncrasies in AI Fiction," 2026. [Online]. Available: https://arxiv.org/abs/2604.03136

[13] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "The Curse of Recursion: Training on Generated Data Makes Models Forget," 2024. [Online]. Available: https://arxiv.org/abs/2305.17493

[14] I. Sutherland, "Technology and Courage," in *CMU Computer Science, A 25th Anniversary Commemorative*. New York, NY, USA: Association for Computing Machinery, 1991, pp. 425–447. [Online]. Available: https://labs.oracle.com/pls/apex/f?p=94065:10:112314537568218:2628

[15] L. Wang, X. Shi, G. Li, J. Li, X. Zhang, Y. Dong, W. Jiao, and H. Mei, "Theoretical Proof that Auto-regressive Language Models Collapse when Real-world Data is a Finite Set," 2025. [Online]. Available: https://arxiv.org/abs/2412.14872

Acknowledgment

I absolutely did use generative AI to help with this position paper! After collecting my stories in prose backed by real citations, I used Gemini 2.5 to refactor my thoughts to craft the outline and narrative. All the text was reviewed and revised by me.