Dylan Patel

From hardware forums to an organization that measures, models, and intervenes in the semiconductor and AI industries

How a self-taught semiconductor analyst built and financed a research institution

Dylan Patel founded SemiAnalysis in 2020, a business that explains semiconductors and the physical and economic systems behind artificial intelligence. He is its chief executive and a research leader, based publicly in the San Francisco Bay Area. What began with one analyst’s writing and consulting has grown into an organization of engineers, industrial researchers, financial analysts, and software specialists. Its work follows a chip through manufacturing, installation, and use, asking at each stage what a technical capability costs and whether it can become a functioning business.[1, 2]

The scale of that change is visible in what Patel can now finance. A newsletter can interpret an announcement; SemiAnalysis can spend months testing a vendor’s software, commission a compiler engineer to find failures, or investigate a transistor through microscopy. Patel still writes, travels, and argues about technology. Increasingly, however, his occupation also involves deciding which questions deserve an expensive investigation and assembling people capable of answering them. His authority and his dependence on others have grown together.

[External image / 外部圖片] Portrait of Dylan Patel
Dylan Patel in SemiAnalysis’s official team portrait. SemiAnalysis; photographer not stated · Rights retained by the source and respective rights holders; no open license stated. · Image source

Repairing an Xbox and learning to look beyond the product

Patel grew up in rural Georgia with immigrant parents. They ran a motel, where the family lived with his grandparents, and later a gas station; he worked in their businesses. He remembers approaching technology with an economic interest rooted in that household.[3, 4] A product could appeal to him as a user while posing a different question about its maker: how much performance could a company obtain from the hardware it sold, and at what cost?

An Xbox failure sent him to internet forums for repair advice. He tried fixes, then kept returning to hardware discussions and eventually moderating them.[5] Moderation added a responsibility to an interest: helping sustain an exchange as well as learning from it. The technical world was available through this continuing activity, rather than through a

professional credential.

His interests moved from gaming into phones, their components, and eventually datacenters. Internship earnings became investments in semiconductor companies, whose financial statements and suppliers he followed.[6] The movement was not an announced career plan. New products redirected his attention, and the businesses behind them offered a second way to understand technologies he already cared about.

He reports graduating from the University of Georgia with majors including data analytics, risk management, and legal studies.[7, 8, 9] College data-science work preceded a job modeling financial risk from hurricanes, earthquakes, and wildfires.[10] This supplied a quantitative apprenticeship alongside the one pursued online. The later analyst's conjunction of technology and money did not come from semiconductor engineering training; it brought different kinds of experience to the industry's questions.

Early 2020 brought disillusionment with his job and his grandmother's death after dementia and an accident. During the pandemic he stayed with his brother in Nashville. After an antagonist exposed his anonymous identity, he briefly withdrew, then published as himself.[11] SemiAnalysis emerged from this unsettled period. A long-established online practice acquired a public name while the arrangement supporting his previous work was coming apart.

He also remembers a period in Minnesota after college and about a year and a half keeping bees. Those episodes resist neat placement in a career ladder. In February 2026, he rejected an inevitable childhood-to-career story. He recalled separate interests and successive steps that only eventually met in the business.[12]

Making explanation into paid work

In 2020, he traveled through national parks with a truck and tent, working from motels. Investors and companies seeking hardware explanations became consulting clients through his posts.[13] The mobile life combined earning and self-education. He was useful to people who understood their own field but needed to know what was happening in an adjoining one; his occupation could begin in the gaps between their expertise.

Patel recalls raising consulting prices as demand grew.[14] People who appreciated a post could pay for an explanation tailored to a decision. Publication and consulting reinforced one another, but the work was still attached to the person and to particular assignments. This was a market for his judgment before it was a staffed research institution.

Doug O'Laughlin helped change Patel's reluctance to charge for the writing itself. The two were fellow semiconductor writers, and Patel initially thought he could improve on O'Laughlin's publication. O'Laughlin urged him to leave anonymous WordPress posting behind, use his name, and charge on Substack. Patel resisted before eventually taking the advice. Their relationship matters because the intervention was specific: it

encouraged an already active researcher to make a continuing paid product out of work he was giving away.[15]

O'Laughlin brought a separate formation. At Bowie Capital in Texas, he learned investment research under Cory Whitaker. His publication, Fabricated Knowledge, joined semiconductor investment analysis to primers and industrial history, asking what happened to businesses as the established course of Moore's law became harder to sustain. He was an adjacent practitioner with experience in judging firms and explaining their technical circumstances. When he later joined SemiAnalysis, that experience entered its leadership; he is now identified as president, responsible for semiconductor strategy, competitive intelligence, and financial research.[16, 17]

In late 2021, a client did not pay for his photoresist report. Patel published it as paid content and recalls forty subscriptions overnight; he continued mixing free and paid research.[18] Photoresist is applied to wafers before lithography. A specialized inquiry commissioned for one buyer had found another route to readers, and subscriptions offered a continuing exchange around the work.

The change did more than add revenue to an unchanged hobby. A report could now be sold to many readers, while continuing consulting preserved contact with particular clients' problems. SemiAnalysis eventually separated the paid newsletter from institutional models and consulting, with different arrangements for corporate and group access. The newsletter's roughly fortnightly research updates bought an individual reader continuing explanation. An institution could instead purchase maintained spreadsheets or help applying an analysis to its own circumstances. These became distinct obligations: keeping a subscriber informed, keeping a model usable, and answering a client's question.[19, 20]

He dates first hiring to mid-2022, recruiting across the technology stack, while beginning broad technical-conference attendance that year.[21] Expansion was both organizational and personal: colleagues increased what the firm could investigate, while repeated travel increased what its founder could understand and whom he could ask. The same industrial system could now be approached through several working specialisms, rather than only one person's reading.

At lithography and photomask conferences, he recalls initially understanding little, reading afterward, and returning with a better grasp. Conversations helped connect published research to present commercial production. By his account, the schedule grew to roughly forty conferences annually.[22] A possible process described in a paper still had to be related to usable tools, materials, suppliers, and costs. Repeated participation gave his industrial explanations something that reading alone could not supply.

Entering the AI argument

By 2023, SemiAnalysis was reaching readers well beyond the original hardware discussion. In May, Patel and Afzal Ahmad published an internal Google employee's argument about

the competitive power of open models. Its provocative claim, reflected in the title Google: We Have No Moat, and Neither Does OpenAI, traveled into the larger debate about who would control AI. The editors explained that the text had appeared in a public Discord, that they had checked its authenticity and removed internal links, and that they disagreed with parts of it. They were publishing an employee's argument, not announcing Google's position.[23]

Two months later, Patel and Gerald Wong attempted a different kind of intervention: reconstructing GPT-4's architecture, infrastructure, training, and inference economics. Here the object was not an identifiable employee document but a reported synthesis assembled from sources and technical reasoning. The distinction shaped the work's authority. Their detailed account concerned a closed model and remained an external reconstruction, even where its estimates were precise.[24]

There was a consequential information gap to work in. OpenAI's own technical report withheld detailed architecture, model size, hardware, training computation, and other particulars, citing competition and safety. For people trying to assess the resources required by frontier AI, the developer's public document left many industrial questions unanswered. SemiAnalysis could offer a way to act before those facts were disclosed. It could not turn its reconstruction into an OpenAI disclosure merely by giving the reader a coherent account.[25]

These two publications show Patel's changing position. He was no longer only translating specialist material for readers farther away from it. He was helping determine which documents and estimates entered the public argument, and under what framing. The audience's interest was partly produced by secrecy: the more consequential a hidden system became, the more valuable an informed account of it could be. That circumstance rewarded both genuine cross-industry knowledge and confident presentation. The distinction between the two would become harder to leave implicit as SemiAnalysis's audience widened.

Following a chip into a datacenter

The next stage of work made the system around the chip increasingly central. In March 2024, Patel, Dan Nishball, and Jeremie Elia examined AI's datacenter-energy demands. They tried to reconcile accelerator shipments with a North American facility database and estimates for other regions, including work with Structure Research. Regional development, including Johor Bahru, complicated aggregate stories about how much computing power the industry could bring online. A shipped accelerator needed a building, cooling, electricity, and a workable installation. Those were separate constraints rather than automatic consequences of demand for AI.[26]

Nishball's responsibilities help explain this expansion. SemiAnalysis identifies his fields as networking, optics, AI infrastructure, and cloud economics, including leadership of its

AI Cloud TCO work. Connecting those subjects changes the question an analyst can answer. A processor's advertised performance does not tell a buyer what a cluster will deliver after network costs, utilization, and operating conditions are included. Patel's breadth was now joined to a colleague's work on the links between physical systems and financial outcomes.[27]

The June 2024 account of a 100,000-H100 cluster, written by Patel and Nishball, pursued that problem at scale. It considered capital, networking, power, failures, checkpointing, and utilization. A failed component or interruption could affect useful output across a system; accumulating more accelerators did not simply multiply the amount of productive training. Patel's AI Engineer presentation that year carried related distinctions to a software-building audience, explaining the stages of generating a response, their different demands on memory or computation, batching, and the cost of a large cluster. Hardware analysis had become part of understanding what software could economically do.[28, 29]

The company's September 2024 treatment of the memory wall returned this expanding inquiry to physical components. Moving data quickly and cheaply enough to feed computation introduced difficulties that faster logic alone could not solve. High-bandwidth memory, packaging costs, memory scaling, and alternative approaches had to be considered together. AI's apparently software-led development remained attached to a manufacturing world with its own limits and costs.[30]

That manufacturing world also demanded regional and linguistic expertise. Wega Wei-Chia Chu's stated responsibilities include outsourced assembly and test, packaging, fabrication processes, and Asian supply chains, with Mandarin and English capabilities. This is one of the ways the firm acquires knowledge that cannot be gathered solely from English-language corporate announcements. Patel's industry network is consequential, but a collaborator's command of a regional production world is a competence of its own.[31]

The wafer-fab model now sold by SemiAnalysis turns manufacturing research into a maintained account of equipment demand. It links capacity and process-node ramps to the steps required to make chips, then estimates tool requirements through assumptions about throughput, uptime, and rework. Lithography, deposition, etch, tracks, and metrology do not grow in identical proportions. Gate-all-around transistors, backside power delivery, or high-numerical-aperture lithography can change how much equipment is needed for a given output. A customer buys a spreadsheet and its continuing explanation, not merely a prediction delivered once.[32]

The datacenter model works along the other end of the chain. Its described workflow combines property records, permits, utility information, public-record requests, satellite observation, and computer vision. It separates owners, operators, and ultimate customers, and follows projects from land and permissions through construction, cooling and generators, energization, and live computing capacity. Accelerator shipments provide another check. The distinctions matter to an analyst trying to locate a bottleneck: permission for a building is different from a powered site, and both differ from usable

servers. The privately sold database remains less inspectable than this published account of its method.[33]

In October 2024, Patel and Nishball’s neocloud playbook addressed people attempting to turn those resources into a business. It followed procurement, financing, deployment, ongoing operations, and returns, distinguishing types of providers and brokers. Buyers had to obtain equipment and customers under conditions that might change before installation was finished. The research now spoke to operators as well as investors: someone building an AI cloud needed a business model, not only a view about which chip was fastest.[34]

The AI Cloud TCO product formalizes much of that practical inquiry. Its costs include servers, networking, power, colocation, financing, depreciation, repairs, utilization, and rental terms. Measured workload performance can then be connected to the price of computation or tokens. Buyers can change contract terms, residual values, and other assumptions. This is an area where engineering knowledge does not replace business judgment; it makes that judgment more specific. A machine that is attractive under one financing and occupancy arrangement may be unattractive under another.[35]

These models are a different form of labor from the essay that first made Patel recognizable. Information must be entered, checked, linked, and updated, and customers must be able to understand changes. The firm’s advertised wafer-fab roles combine capacity and utilization research, pricing and capital intensity, financial statements, Excel and Python, client briefings, and conference travel. The job descriptions make the recurring work visible without establishing when each position was filled. The early act of explaining an industry has become an obligation to maintain a usable account of it.[36]

A comparison that required five months

At the end of 2024, SemiAnalysis published work in which its own production process became especially tangible. Patel, Nishball, and Reyk had spent five months comparing AMD’s MI300X with NVIDIA H100 and H200 systems for training. They established configurations, ran workloads, investigated discrepancies, sent code and intermediate results to both vendors, reported problems, and tested again. Publication was delayed to incorporate fixes, and the article included reproducibility instructions and a review cutoff.[37]

The criticism concerned how difficult it was to obtain useful training performance from AMD’s available software stack. That was more demanding than comparing advertised hardware specifications. The test depended on a working configuration, and a failed or poorly supported software path could prevent the hardware’s theoretical capability from becoming something a customer could use. Time spent debugging was therefore part of the subject, rather than merely inconvenience on the way to a number.[38]

The inquiry also changed the relationship with the manufacturer. In April 2025, the team reported that Lisa Su had contacted them after the article and asked detailed questions

about bug reports. They described subsequent changes in AMD’s developer engagement and proposed further organizational and technical improvements. This was their account of the response, rather than proof that the article alone caused each change. It nevertheless shows a researcher in a new position: the work could supply an agenda for engineers improving the product under examination.[39]

A public exchange with AMD’s Anush Elangovan kept disagreement about the test visible. SemiAnalysis objected to a later comparison’s attention implementations and fairness across configurations; Elangovan described the containers, alternative attempts, and continuing work on reported issues. Container images, kernels, versions, and failures were not secondary details. Changing them could change the result. Here Patel’s enterprise operated among technically capable parties who could resist its framing and answer its criticism, rather than an audience asked only to trust a verdict.[40]

This work required resources beyond ordinary publication: machines, engineering assistance, repeated runs, and specialists’ time. InferenceMAX, launched in October 2025, made continued inference testing an institutional undertaking. It compared hardware and software across configurations and considered throughput alongside responsiveness. AMD, NVIDIA, and cloud providers supplied material support, including hardware access and engineering help. Those relationships made more work possible while also giving the suppliers a stake in how and when their systems were tested.[41]

The project, subsequently renamed InferenceX, opened an important part of the process. Its Apache 2.0 repository organizes automated testing of inference systems and supporting components, records the suite’s evolution, and provides contribution procedures. The public results site connects charts to GitHub Actions runs, commands, logs, artifacts, configurations, and historical results. A reader can follow a plotted result toward the procedure that produced it. The hardware and cost of rerunning an experiment remain substantial, but the claim is no longer available only as a finished article.[42, 43]

ClusterMAX brings related testing to the experience of cloud customers. Its overview describes real workloads and assessments of multinode networking, storage, reliability, documentation, security arrangements, service commitments, and user feedback. A provider is treated as something a customer must operate within, rather than a nominal inventory of GPUs. Ratings still incorporate choices about tests and weights, but the object of analysis has changed: useful computing depends on an organization keeping a service usable as well as on a machine’s speed.[44]

The expansion helps explain why the earlier paid-writing decision was consequential. Sustaining repeated comparisons is expensive even when a public article attracts a large audience. The firm needs clients and differentiated products, yet part of its most inspectable work is available to people who do not buy a subscription. Patel’s task includes financing research whose public value and direct saleability do not coincide perfectly.

An institution with several kinds of expertise

By February 2026, Patel described a sixty-person business. He was recruiting globally and spoke of an expanding interest in AI usage and its economics during 2025–26. The movement farther up the stack introduced a staffing problem of its own: the person who had been working on usage analysis had been hired by Anthropic. Patel could present public work as an opportunity for researchers while also confronting the loss of a capable colleague to the industry the firm followed.[45]

The firm also expanded through established publications. In September 2025, Chips and Wafers announced that it was entering SemiAnalysis, with continuity for its readers and the prospect of expanded work. That brought an established publication and its practitioners into a larger organization. It does not establish an acquisition price or ownership terms, but it gives a concrete instance of expansion through another working practice, rather than through anonymous headcount alone.[46]

Investor-oriented work has its own specialists. Simi Sherman brings stated buy-side experience and curates ChipBook for investors and executives. Chaim Eisenberg's background includes semiconductor research at an Israeli hedge fund and participation in the same product. Alongside O'Laughlin's leadership, their roles show the difference between learning how a process works and organizing what a financial decision-maker needs to know about it. Neither skill automatically subsumes the other.[47, 48, 49]

The Tokenomics model extends the chain of research from installed computing and workload-dependent token production into applications, demand, revenue, profitability, and subsequent demand for hardware. It covers model providers, APIs, software, and infrastructure. This is a consequential widening of what the company sells: the question is not only how much computation can be supplied, but which uses will pay for it. A future application's demand is less directly observable than a shipment or completed building, so this part of the business relies particularly heavily on linked assumptions.[50]

Patel's practical position has changed with this expansion. Starting inquiries across fields now requires organizing differentiated expertise and retaining it, while maintaining products for audiences with different needs. The organization can investigate more than its founder could execute personally. That capacity makes his editorial and financial decisions more important without making him the author of every result.

Paying for inquiry and building a laboratory

Justin Lebar's May 2026 account makes Patel's role as a funder unusually concrete. A former compiler engineer at Google, Waymo, and OpenAI, Lebar joined as a contractor to investigate miscompilations using AI-assisted fuzzing. Randomly generated programs were compared to find transformations that changed their behavior. Apparent failures then had to be reduced and checked; plausible-sounding agent reports could be wrong. He linked

reproducible cases and described reporting findings to vendors.[51]

Lebar credited Patel's willingness to pay for the token expenditure that the work required. The contribution was distinct from Lebar's engineering: Patel supplied backing for an inquiry whose answer would emerge from running, failing, and checking, rather than from commissioning an essay with a conclusion already in view. The account does not disclose an audited spending total. It does show how a research entrepreneur can make technical work possible without personally performing its central operations.[52]

Physical investigation had also become part of the enterprise. In September 2026, the Panther Lake teardown examined Intel's 18A implementation through SemiAnalysis's STEEL laboratory work, using cross-sectional microscopy and elemental analysis to investigate transistor structures, wiring, and backside power delivery. The article distinguished the density of a representative logic cell, measured or modeled, from an assertion about an entire product. Patel appeared in a substantial collaborative byline. The images and measurements depended on the laboratory and its technical authors, not simply on a founder's access to industry contacts.[53]

Laboratory and benchmarking roles, including advertised work in Hillsboro, Oregon, make this an occupational expansion as well as an editorial one. The business now calls for people who can handle tools and samples, produce repeatable measurements, maintain software tests, and explain the results. A newsletter begun through learning and argument has acquired facilities and commissioned practices capable of generating some of the facts it analyzes.[54]

These investigations put useful resistance inside the institution's own work. An attention kernel, a compiler transformation, or a physical transistor can behave differently from what an explanation predicts. What has changed is the ability to pay for the people and apparatus needed to pursue that resistance. Much of the firm's industrial intelligence still depends on sources and inference; these projects add other ways to put a claim under pressure.

Speaking about infrastructure and power

As SemiAnalysis's technical work expanded, Patel also became a public interpreter of AI's industrial and political circumstances. His 2025 AI Engineer talk discussed China's efforts to build large AI systems under hardware constraints and the significance of Gulf investment and accelerator access. He called sanctions "completely useless" in a discussion of Huawei's supply chains. This was a forceful policy judgment, rather than a conclusion settled by a chip specification. His cross-industry knowledge could illuminate the material consequences of policy while leaving political aims, values, and competing evidence open to argument.[55]

The DeepSeek controversy showed how the firm's estimates could enter that argument through other institutions. In April 2025 congressional testimony, Gregory Allen used

SemiAnalysis estimates to challenge the idea that a reported training expense represented the cost of a whole frontier AI enterprise. DeepSeek's own V3 report described 2.788 million H800 GPU-hours for the full training covered by that paper. The two accounts asked different questions: resources used for a defined training run, and the resources of an organization that also employed staff, maintained infrastructure, and conducted other experiments. The narrower figure need not be false for it to leave the broader question unanswered. The testimony carried SemiAnalysis's analysis into Washington; it did not independently turn those estimates into a disclosure from DeepSeek.[56, 57]

Patel's August 2026 conversation with Dwarkesh Patel confronted another boundary, between present constraints and future concentration. Dylan argued that continuing growth could give leading AI laboratories command of much more of the world's usable computation. Dwarkesh pressed the obvious counterargument: enormous prospective revenue should invite enormous expansion of manufacturing. Dylan emphasized the delays across specialized suppliers and the mismatch between cash generation and capital requirements. Demand, pricing, and deployment conditions remained assumptions in the scenario, not facts about an already observed 2028.[58]

The argument draws on the particular industry Patel learned to follow. A financial incentive does not instantly create a fabrication process, a tool supplier's output, or an energized datacenter. Equally, the identification of a bottleneck does not establish how long it will last or which organization will benefit. His public style can make the connections vivid; its continuing test is whether the linked explanation survives changes in the conditions that make it plausible.

Infrastructure research also encounters people with claims other than those of buyers and operators. In September 2026, Maya Barkin, Reyk, Elia, and Patel examined datacenter moratoriums, separating restrictions on new permissions from interruptions to already authorized construction. Their database combined public records, satellite work, AI-assisted research, and human validation, mapping policies to jurisdictions, parcels, and delays. This gave a more granular account than a headline count of restrictions. It answered a capacity question, however, without deciding whether local objections about electricity, land, or other uses were justified. The industry's ability to deliver a project and a community's reasons to oppose it remain different parts of the public problem.[59]

Research, capital, and the responsibilities of scale

The business's financial subject matter cannot be separated from its own changing financial position. In September 2026, Nishball, Oliver Kennon, and Terence Ong examined NVIDIA's infrastructure backstops: commitments that could help cloud operators secure funding, together with possible exposure if utilization or demand disappointed. Patel was not in the byline. The article is a useful instance of research proceeding through colleagues' authored judgments, while pursuing the same concern with how a technical expansion is financed.[60]

SemiAnalysis’ s compliance policies acknowledge the sensitivities of a business serving investors and interacting closely with industry. They describe an external compliance program, procedures for material nonpublic information and personal trading, and Series 16 supervisory-analyst review of institutional research. That review expressly excludes the newsletter. The published policies also distinguish the firm’ s work from investment advice and securities recommendations. They are a statement of arrangements, not an audit of their implementation, and their differentiated scope matters to readers of the different products.[61]

A July 29, 2026 Form D gives a concrete record of a further move. SemiAnalysis Capital Fund I, LP filed for a venture-capital offering with a $400 million target, identifying Patel through the general partner’ s management. At that filing it reported zero securities sold and zero investors. The target was therefore an intended offering, not $400 million already raised. The record marks a proposed investment enterprise associated with the research name; it does not disclose a completed portfolio or Patel’ s personal earnings.[62]

These developments change the stakes of the relationships that helped Patel learn. Supplier conversations can yield valuable information; vendor support can make a costly benchmark possible; investor clients can sustain specialist research. An associated investment enterprise introduces another prospective interest in the same industrial world. The problem is not resolved by presuming wrongdoing or by treating a published compliance policy as complete reassurance. Disclosure, separation of roles, and workable legal agreements become practical responsibilities of the institution Patel has built.[63]

What Patel’ s occupation has become

The demand for AI infrastructure gave a particular conjunction of abilities an unusually consequential market. Patel could connect technical questions to economic ones, while the organization around him could pursue more of the industrial system than a single analyst. That enlargement is not simply a matter of having more information. It changes the kinds of questions the firm can afford to ask, the work it must maintain, and the audiences to which it becomes answerable.

The more distinctive later development is that he did not remain solely a commentator benefiting from that demand. SemiAnalysis added maintained models, iterative tests, commissioned software inquiry, and physical investigation. Patel’ s work increasingly lies in sustaining the conditions for those different kinds of research: choosing problems, paying for investigations, bringing specialists into an organization, and selling products capable of supporting further work. The firm’ s collective authorship is part of that achievement, not a qualification appended to it.

There is an unresolved tension within the result. Industrial relationships and money allow the research to become more concrete, while bringing new interests into its production. The benchmark that must be tuned with a vendor, the model that must be useful to a

customer, and the proposed fund connected to the research brand place Patel in different positions. His future authority will depend on how those positions are managed as much as on the acuity of a prediction. The institution has acquired more ways to find out what is happening. It has also acquired more reasons to make clear who is doing the finding, who is paying, and what remains an inference.

Research cutoff: October 4, 2026, with a focused rereading of the cited formation interviews on October 5. Source-specific qualifications and chronological limits are retained in the notes.

Sources

1. SemiAnalysis, Dylan Patel, institutional biography, undated.↩
2. Dylan Patel, Public LinkedIn profile, location metadata, accessed October 4, 2026.↩
3. Sequoia Capital, Dylan Patel on hardware–software co-design, Training Data, episode 92, undated transcript. Retrospective account; it gives two years of quantitative employment. Recollected childhood ages are not assigned calendar dates.↩
4. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩
5. Sequoia Capital, Dylan Patel on hardware–software co-design, Training Data, episode 92, undated transcript. Retrospective account; it gives two years of quantitative employment. Recollected childhood ages are not assigned calendar dates.↩
6. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩
7. Abram Brown, The Information, April 17, 2026. The degree fields are presented as Patel’s reported educational background, not independently established qualifications.↩
8. University of Georgia Alumni Association, 2017 Virtual Plaque, Senior Signature roster. It lists Dylan Sanjay Patel, matching the middle name used in the 2024 interview, but is a donor roster rather than a transcript and does not establish the reported majors.↩
9. University of Georgia Alumni Association, Senior Signature, program description, undated. The program is for graduating students; this corroborates the university context rather than every degree detail.↩
10. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩
11. Sequoia Capital, Dylan Patel on hardware–software co-design, Training Data, episode

92, undated transcript. Retrospective account; it gives two years of quantitative employment. Recollected childhood ages are not assigned calendar dates.↩

12. Latent Space, In-Context Cooking with Dylan Patel, February 28, 2026, opening and closing transcript discussions. Retrospective episodes; staffing and recruitment are Patel’s dated self-report.↩

13. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩

14. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩

15. Latent Space, In-Context Cooking with Dylan Patel, February 28, 2026, opening and closing transcript discussions. Retrospective episodes; staffing and recruitment are Patel’s dated self-report.↩

16. Doug O’Laughlin, About Fabricated Knowledge, first-person professional account, undated.↩

17. SemiAnalysis, Doug O’Laughlin, institutional biography, undated.↩

18. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩

19. SemiAnalysis, About, institutional description, undated.↩

20. SemiAnalysis, FAQ, subscription and model-access terms, undated.↩

21. Dwarkesh Patel, Asianometry and Dylan Patel, October 2, 2024, transcript from 01:50:20. Retrospective account; it gives three post-college employment years. Former employer and exact calendar remain unconfirmed.↩

22. Sequoia Capital, Dylan Patel on hardware–software co-design, Training Data, episode 92, undated transcript. Retrospective account; it gives two years of quantitative employment. Recollected childhood ages are not assigned calendar dates.↩

23. Dylan Patel and Afzal Ahmad, Google: We Have No Moat, and Neither Does OpenAI, May 4, 2023; source introduction and attribution.↩

24. Dylan Patel and Gerald Wong, GPT-4 Architecture, Infrastructure, Training, and Inference, July 10, 2023; accessible introductory material.↩

25. OpenAI, GPT-4 Technical Report, 2023, section 2, “Scope and Limitations.”↩

26. Dylan Patel, Dan Nishball, and Jeremie Elia, AI Datacenter Energy Dilemma, March 13, 2024.↩

27. SemiAnalysis, Dan Nishball, institutional biography, undated.↩

28. Dylan Patel and Dan Nishball, 100,000 H100 Clusters, June 17, 2024.↩

29. Dylan Patel, Compute and System Design for Next-Generation Frontier Models, AI Engineer, 2024; talk transcript.↩

30. SemiAnalysis, The Memory Wall, September 3, 2024.↩

31. SemiAnalysis, Wega Wei-Chia Chu, institutional biography, undated.↩

32. SemiAnalysis, Wafer Fab Model, methodology and product documentation, undated.↩

33. SemiAnalysis, Datacenter Industry Model, methodology and product documentation, undated.↩

34. Dylan Patel and Dan Nishball, AI Neocloud Playbook and Anatomy, October 3, 2024.↩

35. SemiAnalysis, AI Cloud TCO Model, methodology and product documentation, undated.↩

36. SemiAnalysis, Careers, job descriptions, undated; whole page treated as one source. Advertisements establish requested tasks, not filled positions. Specific specialists' recruitment dates and the complete calendar of model introductions remain unestablished; current biographies and products are not treated as a dated hiring or launch sequence.↩

37. Dylan Patel, Dan Nishball, and Reyk, MI300X versus H100 versus H200: Training, December 22, 2024.↩

38. Dylan Patel, Dan Nishball, and Reyk, MI300X versus H100 versus H200: Training, December 22, 2024.↩

39. SemiAnalysis, AMD 2.0: A New Sense of Urgency, April 23, 2025; authors' account of vendor interaction.↩

40. SemiAnalysis and Anush Elangovan, Public AMD benchmarking exchange, LinkedIn post and reply, undated.↩

41. SemiAnalysis, InferenceMAX launch, October 9, 2025; methodology and support disclosures.↩

42. SemiAnalysisAI, InferenceX, public repository and README, current October 4, 2026.↩

43. SemiAnalysis, InferenceX results and documentation, current October 4, 2026. Documentation was inspected; the complete benchmark suite was not rerun.↩

44. SemiAnalysis, ClusterMAX Overview, methodology description, current October 4, 2026.↩

45. Latent Space, In-Context Cooking with Dylan Patel, February 28, 2026, opening and closing transcript discussions. Retrospective episodes; staffing and recruitment are Patel's dated self-report.↩

46. Chips and Wafers, Joining SemiAnalysis, September 1, 2025.↩

47. SemiAnalysis, Simi Sherman, institutional biography, undated.↩

48. SemiAnalysis, Chaim Eisenberg, institutional biography, undated.↩

49. SemiAnalysis, Doug O’ Laughlin, institutional biography, undated.↩

50. SemiAnalysis, Tokenomics Model, methodology and product documentation, undated.↩

51. Justin Lebar, Finding Miscompiles for Fun, Not Profit, May 28, 2026.↩

52. Justin Lebar, Finding Miscompiles for Fun, Not Profit, May 28, 2026.↩

53. SemiAnalysis, Intel Panther Lake Teardown, September 26, 2026; STEEL laboratory collaboration.↩

54. SemiAnalysis, Careers, job descriptions, undated; whole page treated as one source. Advertisements establish requested tasks, not filled positions. Specific specialists’ recruitment dates and the complete calendar of model introductions remain unestablished; current biographies and products are not treated as a dated hiring or launch sequence.↩

55. Dylan Patel, Geopolitics and AI Infrastructure, AI Engineer, 2025; talk transcript.↩

56. Gregory C. Allen, DeepSeek: A Deep Dive, congressional testimony, April 8, 2025.↩

57. DeepSeek-AI, DeepSeek-V3 Technical Report, December 2024, revised February 2025, abstract.↩

58. Dwarkesh Patel, Dylan Patel interview, August 25, 2026; opening discussion of supply, financing, and compute concentration.↩

59. SemiAnalysis, Datacenter moratorium analysis, September 15, 2026.↩

60. Dan Nishball, Oliver Kennon, and Terence Ong, NVIDIA’s Backstop Universe, September 11, 2026.↩

61. SemiAnalysis, Compliance Policies, published policy description, undated.↩

62. US Securities and Exchange Commission, SemiAnalysis Capital Fund I, LP: Form D, filed July 29, 2026, items 13–14 and signature.↩

63. San Francisco Superior Court, SemiAnalysis LLC v. Wei Zhou: July order, CGC-26-635328, July 20, 2026; DocketAlarm facsimile/OCR. The order compelled arbitration and stayed litigation. This procedural disposition establishes neither party’s accusations, and no developmental consequence beyond the company’s legal relationships is inferred.↩