# SuperAI Debrief

**Wednesday, June 10, 2026 - Singapore**

## Executive Summary

The morning sessions established the macro thesis — the model layer is commoditizing, and value is migrating to deployment and context — while the afternoon showed what is being built one layer up. The infrastructure stack of the agent era is rapidly taking shape: agent-native search (Exa: "search is infrastructure for AI"), agent governance (Google's "manage agents like employees" — identity, registry, performance management), and durable enterprise assets that survive model swaps (Tencent Cloud's "durable assets" frame).

The commoditization debate also matured over the afternoon. Commoditization is real but partial — lower-tier, routine tasks commoditize, while high-value verticals (professional-grade 4K/8K generation, scope-problem domains) do not. This dovetails precisely with Kevin's morning scale-vs-scope framework. The Stripe session then added the business-model-layer conclusion: as AI collapses the four historic constraints on company formation (building, global expansion, pricing, discovery/commerce), payments, billing, and trust infrastructure become more important, not less — "AI makes software companies easier to start; Stripe wants to tax the moment they become real businesses."

The closing AGI session serves as this report's conclusion. When everyone has AI, advantages cancel out in adversarial domains (the "polytheistic" scenario); in a world where LLMs produce "the average," the scarce premium shifts to authenticity and individuality; and as fabrication becomes trivially cheap, a verification and auditing economy grows. The four focus areas derived from the day: (1) personal, private, programmable AI; (2) physical AI; (3) human communities as a complement to digital AI; (4) verification, proctoring, and auditing.

## Session Summaries (Chronological)

### 1. Capital Investment Trends in Tech — 09:38

Benedict Evans (Analyst) — "AI Eats the World"

Hyperscaler capex has surged from roughly $400B in 2025 to roughly $700B in 2026 (oil and gas combined is about $1T), with roughly one third flowing to Nvidia — heading toward $100B per quarter if nothing slows. Meta and Alphabet are both pushing toward or past 50% of revenue going to capex; Alphabet raised $85B in equity and Meta is considering the same. Bottlenecks: GPU/memory supply, power (backed up multiple years), gas turbines, and political backlash against data centers in the US.

The primary demand driver is agent coding. Anthropic's revenue run rate is accelerating sharply, and Uber spent its entire annual AI budget in Q1 alone.

The core thesis: LLMs are infrastructure, not platform monopolies. Capital-intensive, no network effects, no fundamental differentiation between labs — all building with the same data, chips, and infrastructure. The analogy: mobile data grew 1,500-2,000x over 15 years while telco stocks went nowhere — the commodity trap. Sam Altman's "intelligence as a utility like electricity" framing faces exactly this challenge.

The deployment reality check was sobering. ChatGPT/Gemini usage is "mile wide, inch deep"; in enterprise, far more people say "I'll use it next week" than use it daily. Per the Atlanta Fed's December 2025 CFO survey, easy-to-deploy categories (productivity, insights, customer service) show roughly 2-3% impact, while hard-to-deploy categories (cost savings, new revenue) barely register.

The framework offered: Absorb, then Innovate, then Disrupt — first do old things better with new tools, then do things only possible with the new technology, finally redefine the question entirely. Like the barcode — deployed to save checkout time, it enabled inventory accuracy and ultimately created Walmart. When something becomes cheap or free, the interesting question is not cost savings but what was previously impossible that now becomes

possible.

### 2. Frontier AI Deployment Insights — 10:03

Geoff Soon (VP Revenue APAC, Mistral) · Cherie Shi (Global Business Manager, MiniMax) · Hemant Mohapatra (Partner, Lightspeed India) · Zixuan Li (Head, Z.ai)

The definition of "frontier" itself is changing. Geoff Soon (Mistral) redefined it as what you can effectively deploy in government and enterprise with durable, controllable capabilities; Cherie Shi (MiniMax) as how many real users use your model daily. Benchmarks are no longer the metric. The gap between open and closed weights narrows every six months, and beyond a point each 1% improvement requires 10x more capital ($10B to $100B) — so the big balance-sheet players win the scale and price war.

Kevin's investment framework was the highlight. The AI model supercycle runs in two phases: extractive (who produces the most intelligent tokens for the least cost per watt) shifting to distributive (who solves real customer problems with that intelligence) — and the world is at the transition point now. Being the best "token extractor" is losing relevance. Layered on top is the scale-versus-scope distinction: scale problems like coding — solved once, replicated cheaply at volume — will be commoditized by baseline models, while scope problems — cancer research, materials science, mathematics, where the problem definition keeps changing as you go deeper — are where durable value lives. Areas to watch now: physical world models, bio, materials science, mathematics, game simulation. Also useful: the Ford/GM analogy (early AI = make the best model, any color as long as it's black; the next phase = go-to-market, brand, segmentation) and the Swiss

Army knife versus specialist tool distinction (general models for POC and hypothesis; once value is proven, take enterprise context into a specialized model).

On the competitive landscape: MiniMax differentiates via multimodal unification (merging LLM, Hailuo video, speech, and music into one model improves both generation and understanding); M3, released about a week ago, offers strong agentic and visual/video understanding with a 1M-token context window at roughly one third the price of Claude Sonnet 4.6, targeting the most cost-efficient frontier model globally. Over 70% of revenue is from global markets, with strong open-source demand from Southeast Asia for local data center deployment. Mistral's edge is partnerships with industrial heavyweights holding proprietary data — ASML in high-tech manufacturing, Airbus announced at the Paris summit — plus the acquisition of Emmy (surrogate modeling that makes digital twin simulation 10-100x faster). Its challenges: the EU's 27 fragmented regulatory markets and significantly less funding liquidity versus the US. Positioning: Europe as a credible third alternative to US and China AI.

An honest diagnosis of open-source dynamics: open weights are not free or cheap, merely fungible (like oil between gas stations); since open weights cannot be monetized standalone, you must own the full stack — infrastructure, compute, energy, deployment — to make the token economics work. Companies that open-source today will face pressure to close once they have something worth protecting. On the most underestimated success factor on the AGI path: Geoff Soon named the distribution of GPU/compute access (geopolitical shocks could heavily constrain capability spread); Cherie Shi named cost reduction; Kevin named founder imagination — too many founders are "feeding oil back to horses hoping they run faster," when what is needed is the vision to reimagine workflows entirely, not just automate existing ones.

### 3. Secret War: Old Web vs New — 10:45

Ariel Shulman (CPO, Bright Data)

Ariel (Bright Data) framed the core problem: the web was built in 1989 for humans, but AI agents are now the primary consumers — and the web does not know how to handle this yet. Google's 29-year data moat spans text (Gmail/Drive), images (Photos), video (YouTube), commerce (Chrome), local (Maps), behavior (Android), and expert knowledge (Scholar). The DOJ antitrust ruling forces Google to share some data, but it keeps its secret sauce (PageRank, spam-detection algorithms). The implication for builders is clear: you have to crawl and build your own data infrastructure.

The most important insight concerned how the web now fights back against bots. Instead of blocking detected bots, websites now serve them disinformation (the "honeypot"): a bot sees "no tickets available" where humans see inventory; a bot sees a $399 room rate where humans see $199. Because agents do not sanity-check like humans — an agent sees $399, updates 5,000 records, and sends 5,000 emails with wrong pricing — the biggest hidden risk is not blocked requests but plausible, false data arriving with no indication anything is wrong.

Data freshness decays at different rates by category: e-commerce pricing and availability in about one day, news in 1-2 days, finance in 5-7 days, general blogs much longer. Bright Data itself: 15

years in web data infrastructure,

scraping 50B+ pages per day (3x Google's advertised search volume), parsing and archiving 3M+ pages per day (more than the Wayback Machine), 150M+ IP addresses, 2.5 petabytes scraped daily — delivered as trusted HTML, JSON, Markdown, or CSV via API. It has won federal lawsuits in California against both Meta and Elon Musk, establishing the precedent that publicly accessible data (no login, no paywall) is fair game.

### 4. AI Trust in Consumer Markets — 11:23

Panel discussion — consumer AI & trust (private session; speakers withheld)

One global bank's top AI use cases: drafting and information extraction, client meeting preparation, and senior leadership visibility — CEOs sharing their AI tool usage weekly drives adoption. Some companies run three-tier AI literacy tests, with the top tier requiring employees to build agent workflows. New teams are scrambling to rebuild corporate foundations "AI-first," unlearning old processes.

On consumer trust: willingness is highest for routine, low-stakes tasks and drops sharply for higher-stakes or less transparent interactions. One panelist's moat discussion stood out: moats in AI are not static — they must be continuously rebuilt. The primary lever is user acquisition speed — build critical mass before full customization kicks in, after which user attention and capital follow. With OpenAI and Anthropic pointing everything in one direction, smaller players must decide: defend a niche or sell. Building a user interface is buying time.

From the founder panel: One founder (manufacturing) noted that AI democratizes ideation, making execution the differentiator; another panelist (international expansion) emphasized paying customers over frontier R&D with no near-term revenue, and that overseas expansion still requires building teams with shared culture and delivery accountability — AI does not replace this. A fundraising insight: how you prepare the data room now matters — be mindful of AI picking up unintended signals in due diligence. On going global from Singapore: the small home market forces international thinking from day one, AI companies can target globally without being anchored to one country (at the cost of organizational complexity), and Chinese founders moving into US and Asian markets often underestimate the need for local go-to-market teams.

### 5. Cost of Intelligence — 12:15

Evan Conrad (CEO, San Francisco Compute) · Carmen Li (CEO, Silicon Data) · Val Bercovici (Chief AI Officer, WEKA) · Anand Iyer (Founder, Canonical)

GPU cost is fundamentally a risk premium, not a unit cost: guaranteed on-demand runs up to $10/hr while preemptible spot runs around $0.50/hr. Token costs have risen roughly 100x in the past year; cache tokens are now about 80% of what agents consume, with wild variance in cache-token pricing across providers. Demand elasticity is extreme — every 100x drop in unit inference cost yields roughly a 10,000x increase in consumption.

The subsidized intelligence era is ending. $20 subscriptions have become $100-200 plans, and "unlimited" is now capped (the Anthropic example). CIOs and CFOs are actively capping token

spend and pushing model routing and budget harnesses. Businesses running 100% on-demand GPUs face margin compression tied directly to utilization.

On infrastructure: "memory" has two distinct meanings that should not be conflated — the software/AI view (context window, KV cache, prompt state) versus the hardware/supply-chain view (HBM on the GPU package, DDR5 DRAM, NVMe storage). Compute and memory remain tightly coupled, which drives rate limits and pricing. Reinforcement learning is now central to pre-training, not just fine-tuning, making batch inference a major memory

workload. Google's split of its TPU line into TPU8 (training) and TPU8i (inference) reflects the fundamentally different compute/memory balance between the two workloads.

The biggest thesis: a compute futures market. Compute will be the largest human resource — currently about $600B/yr in spot markets, with a derivatives market that could reach $6-8T. A CME partnership enables futures contracts on compute; buyers can exit or sell contracts within ten days rather than being locked in (the analogy: NASDAQ as a transaction layer plus a financial data layer normalizing pricing across providers). Energy is a necessary input but not the primary GPU cost driver ("air to breathe"); the real energy constraint is credit and financing — hyperscalers are near capex limits, frontier labs are not investment-grade, and everyone else cannot finance at the scale needed to build new power plants.

Outlook for 2027: data centers moving globally is the dominant near-term trend. Sovereign AI is a non-issue in the US and critical everywhere else. Countries with power assets but no fiber access create geopolitical trade opportunities (undersea cable access deals). "Gross Token Product" was proposed as an economic metric alongside GDP, with token export likely to emerge as a major national export category. New chip architectures beyond the GPU are coming (inference processors like Cerebras, sparse/dense network accelerators), and non-proliferation treaties for AI at the scale of frontier labs and sovereign-AI nations were flagged as an emerging concern.

### 6. AI Productivity Initiatives — 14:39

Oscar Hjelde (ML/AI Engineer, NBIM — Norges Bank Investment Management)

A fund investing in 60+ countries, holding on average 1.5% of listed equities and covering 7,000 companies with only 700 people (a quarter of GIC's headcount) shared how its CEO set a target three years ago of becoming 20% more productive through AI — and how it was achieved.

Five success factors. First, full management buy-in with visible weekly CEO commitment. Second, bombarding the organization with events year-round so that it was "impossible not to learn about AI if you worked at the fund" — mandatory upskilling, tech days across four international offices, and sharing both what works and what fails. Third, an internal AI ambassadors network meeting weekly — people who, alongside their day jobs, drive transformation; for example, a dedicated lawyer trained on AI who knows how best to leverage it in the legal function. Fourth, an internal AI team wearing three hats: internal consultants (advice and upskilling), forward-deployed engineers (sitting close to the business on management-prioritized use cases), and platform architects (a common platform on which anyone can spin up agents and MCPs).

The journey ran in three "orchard" phases: democratization (open the gates — Claude Code, Cursor, and similar tools for everyone; pick the low-hanging fruit), identification (prioritize the higher, more valuable fruit with the C-suite — "chief projects"), and dedicated delivery with business ownership.

Two flagship use cases. Bad Apples: screening qualitative risks (supply chain, reputational) across 7,000 portfolio companies with an agentic system ingesting vendor data, news, financial reports, government filings, and — critically — local media, where language barriers previously blocked early signals. A deliberately low bar for flagging (accepting false positives), with human risk analysts reviewing — human-in-the-loop by design. CMAS (Company Meeting Assistant and Simulation): portfolio managers hold roughly 3,200 company meetings a year at about three hours of prep each; CMAS pulls from the investment thesis, prior meeting notes, filings, and news to draft an agenda — not to replace preparation but to offer a new angle, leveraging the fund's strategic advantage of corporate access at scale.

Four key lessons: (1) developer teams shrank from "two-pizza" to "half-pizza" (two developers plus one PM), with faster feedback loops and shorter iterations — and more pressure on the organization to produce specs. (2) "Build for the model that exists six months from now" (borrowed from Anthropic's Claude Code team) — extrapolate the exponential curve to the model that will exist when the project ships. (3) "Harness is cheap, context is expensive" — harnesses can be outsourced to the big labs or built quickly, but the context unique to your business (knowledge, MCPs, skills, documented internal processes) is the expensive, slow-to-build real asset. (4) Lincoln's "give me six hours to chop down a tree and I will spend the first four sharpening the axe" — the compounding returns from reinvesting in your own tools are the source of competitive advantage.

### 7. Stripe — AI Commerce and the Collapse of Four Constraints

Abhi Tiwari (Head of Global Product, Stripe) — "Inside the Rising AI Economy"

Core thesis: AI is collapsing four historic constraints on company formation and growth — building, global expansion, pricing/monetization, and product discovery/commerce — creating a world where tiny teams go from idea to global revenue far faster than before.

The cost of building is collapsing. The early "human intelligence" cost of starting a software business — engineers, designers, PMs, marketers, finance, support — is being replaced by AI coding agents. Agent traffic to Stripe's developer docs is rising fast and may exceed human traffic by year-end; iOS app releases have re-accelerated since agentic coding took off; and Stripe Atlas startups are reaching first payment faster — a February 2026 cohort went from incorporation to first payment in about six weeks. The bottleneck shifts from "can you build it?" to taste, judgment, product choice, and execution quality.

The cost of going global is collapsing. Local teams for payments, compliance, support, and operations are being replaced by APIs. AI companies internationalize far earlier than SaaS did: old SaaS reached roughly 25 countries in year one and 50 by year three; AI companies reach roughly 42 countries in year one and 120 by year three. Around half of top AI company revenue now comes

from outside the home country, versus roughly one third a few years ago. Smaller but skilled markets — Singapore, Iceland, Estonia, Luxembourg, Switzerland — become more addressable. Global from day one is becoming the default, and localization (currency, payment methods, checkout optimization) is now a direct revenue lever.

AI breaks old SaaS pricing. Software pricing evolved from perpetual licenses to subscriptions to per-seat SaaS, but AI has a different cost structure: every inference carries real, variable compute cost, so flat subscriptions risk power users burning disproportionate compute. At the same time, AI creates measurable outcomes — resolved support tickets, legal documents generated, workflows completed, code commits. Pricing therefore moves from flat seat-based subscriptions toward usage-based pricing, credit systems, hybrid subscription-plus-credits, and eventually outcome/workflow-based pricing. Examples: Intercom charges by support resolution; legal AI tools charge per document; Replit-like developer platforms use subscriptions plus AI credits. An AI company still purely per-seat may be under-monetizing or mispricing compute risk — this is one of the biggest unsolved AI business-model questions, and hybrid pricing is probably the near-term winning model. (The mirror image of the morning's Cost of Intelligence session: CFOs capping token spend and routing models.)

Discovery and commerce are changing. AI agents shorten the path from "I want something" to "I bought it." Roughly half of AI users already use AI for shopping, and AI-assisted shopping converts at a cited 12.3% versus 3.1% without — roughly 4x. Agentic commerce proceeds in levels: Level 1, agents helping complete checkout (already live); Level 2, AI shopping assistants finding and comparing products (also live); higher levels, agents initiating purchases or managing purchase flows autonomously. Discovery may move from keyword search and marketplaces to agent-mediated recommendation and checkout. But payments and trust are the chokepoint: for agents to buy on behalf of users, the hard problems are user control, product/catalog accuracy, integration across many agent protocols, fraud prevention, and ownership of the merchant-customer relationship.

Strategic read. First, AI lowers the floor, not the ceiling — more people can build companies, which also means more noise; the edge moves to taste, distribution, workflow ownership, trust, and pricing design. Second, "solo founder to global revenue" becomes plausible. Third, this is not just an AI hype talk — it is Stripe positioning itself as AI economic infrastructure: if agents build, sell, price, and buy things, Stripe wants to sit underneath identity, payments, billing, fraud, checkout, localization, and agent permissions. The second-order investment implication: the beneficiaries are not only model labs or app-layer startups — infrastructure winners may be the platforms handling AI-native billing, usage-based pricing, global payments, checkout localization, agent permissions, and fraud/risk for autonomous transactions. In one line: AI makes software companies easier to start; Stripe wants to tax the moment they become real businesses.

### 8. AI Applications in the Real Economy — 15:44

Ailing Teng (GM Developer Business, StepFun) · Kan Yang (Head of Solution, BytePlus) · Cameron Zhou (Head of Operations Compute, Tencent Cloud) · Grace Shao (Founder/Analyst, AI Proem)

Enterprise adoption of generative AI is running about 3x faster than executives expected (Tencent Cloud internal research), spreading bottom-up and employee-driven — mirroring the iOS/BYOD wave. As a result, enterprise investment focus has shifted to governance: making agent AI secure and business-environment-friendly. Open source is the primary driver of cutting-edge agent tooling, but enterprise feature sets lag behind — and that gap is itself an opportunity: enterprises need help deploying open-source agents safely.

The panel consensus: 2025-2026 is "the agent year." Agent loops demand search, retrieval, and action — fundamentally more complex than the chat/LLM era. Ailing Teng's (StepFun) AI-native organization thesis stood out: product and engineering merge into single roles — "I'm not hiring product managers anymore; I have product engineering" — with material implications for headcount structures and hiring profiles. New interaction paradigms beyond the screen are emerging, including ambient and voice interactions while mobile. StepFun treats its open-source community as a moat — "you can't close your eyes and ears and just train models" — with community feedback looping into model quality, and local/on-device performance requiring extensive pre-launch community work. StepFun recently filed for a Hong Kong listing.

The session also delivered the most refined take on commoditization: it is real but partial. Lower-tier, routine tasks commoditize; high-value verticals do not. BytePlus's example: as basic video generation commoditized, it moved to real-image/real-voice reference inputs and 4K/8K generation — and users pay significantly more for professional-grade output. Another BytePlus insight: the shortage is not ideas but demo-to-production conversion, and AI collapses that cost.

Cameron's (Tencent Cloud) enterprise framework resonates precisely with NBIM's "context is expensive": don't bet on specific models or agents — build durable assets. Assets means AI-friendly data formats, internal benchmarks, governance policies, and knowledge bases — things that persist across model swaps and compound in value over time. "Benchmarks on GitHub are one thing — benchmarks on your own business are more valuable." Tencent Cloud is launching Core Pro on this principle: flexible policy configurations that live outside the agent (preventing hallucination from overriding policy), compatible with whichever agents go viral — swap models, retain accumulated assets. (Resource: the AI Proem newsletter by Grace Shao — deep dives on APAC/China AI, on Substack.)

### 9. AI Search Trends — 16:24

Will Bryk (CEO, Exa)

Humans search roughly 15B times a day in 2026, and AI systems are approaching the same volume. Exa predicts AI will surpass human search volume by the end of 2026, with a long-term projection of AI searching 1,000x more than humans at roughly 10x YoY growth. The logic is simple: GPT-5 holds roughly 10TB of information while the web is millions of terabytes — search is not optional for AI; it is infrastructure. And agents need fundamentally different search than humans: humans type a few keywords and lazily scan a few links; agents issue complex queries, need thousands of documents, structured output, and sub-200ms latency.

Exa closed a $250M Series C around May 2026 at a $2.2B valuation, serving 5,000+ companies and 400,000+ developers — customers include Cursor and Cognition (coding agents), HubSpot (go-to-market agents), banks, hedge funds, and PE firms. Core capabilities: deep search (minutes to hours — e.g., "every YC-funded AI startup, with batch and status"), fast search (sub-200ms, built for real-time voice agents), token compression (returning roughly the 100 most important tokens per document instead of 1,000, cutting downstream compute costs), structured JSON output, and per-customer configurability (domain whitelists/blacklists, content-type filtering — effectively 5,000+ custom search engines). The company describes a novel retrieval method developed through years of ML research — undisclosed, but framed as a genuine breakthrough — and positions itself as "the Anthropic of search," a deep research organization rather than an API wrapper (recent senior hires from Meta retrieval infrastructure, Yandex indexing, and Google Zurich research). Will Bryk's framing: "There are more space programs in the world than true search engines."

### 10. Agentic Era Insights — 16:42

Moe Abdula (VP Customer Engineering, Google Cloud)

Per Gartner, enterprise agent adoption has gone from roughly 5% in 2025 to roughly 40% now. The Fortune 500 projection: 100 agents per employee — at an average of 60,000 employees, effectively 3x the workforce. Most agents today come out-of-the-box (Salesforce, ServiceNow), but custom-built agents are growing fast, and as industry expertise gets codified into sellable agents, the prediction is that next year's "Super AI" conference

becomes "Super Agent."

Moe's core thesis: manage agents like employees, applying human onboarding principles. (1) Identity — give agents an ID, enabling audit trails, access control, and accountability. (2) Registry — a source of truth for what each agent can and cannot do. (3) Gallery — a single catalog of all agents, whether built, acquired, or out-of-box. (4) Skills — train agents on specific tasks (e.g., a conference scheduling agent that reads calendars and batches speakers). (5) Workflow training — business users, not IT, define agent behavior via no-code/low-code tools. Google internally has 150,000 agents, roughly 100,000 of them built by business users rather than engineers. (6) Collaboration with boundaries — agents work in parallel but scoped to defined data and agent sets. (7) Performance monitoring — observe outcomes, not just task completion, and retire underperforming agents.

At scale a zero-trust model is critical: identity plus gateways plus shared protocols plus registries plus durability. Human-in-the-loop remains essential because agents act on real systems — credit cards, bookings, campaigns. Highlights included Grab's real-time voice translation demo (launched the night before; facilitating 10M+ language-barrier calls per month, with the travel segment up 10x in three years — a live English-Vietnamese driver call), and the Gemini Enterprise platform covering the full stack from build to scale to govern to optimize: agent runtime, session memory, registry, policy controls, and Model Armor (detecting and removing prompt-injection threats).

### 11. AGI Insights and Predictions — 17:13

Balaji Srinivasan (The Network State) & Benedict Evans (Analyst) — "After the Breakthrough"

The most thought-provoking session to close the day. The AGI debate framing: polytheistic (many competing models) versus monotheistic (one model wins) — with the polytheistic view more empirically likely. In adversarial, time-varying domains (markets, law, politics), everyone having AI just raises the floor — advantages cancel out. Lawyers using AI to redline faster simply speed up both sides equally.

Three patterns of job displacement: (1) the job disappears entirely — the elevator operator, where the task is fully automated; (2) the job expands — accounting grew as computational power enabled more complexity; (3) the job is replaced by a comparable but different job over 10-20 years — the music industry via streaming. And the Jevons paradox applied to AI — three outcomes when something gets cheaper: (1) more utilization (coding autocomplete); (2) complementary goods become premium — cheap digital AI makes physical AI more valuable (robots, hardware you cannot simply download); (3) "cheap" becomes derogatory — AI spam and low-quality content at scale.

The "average" problem is central: LLMs produce what most people would say — valuable for commodity tasks (NDAs, spreadsheets), but originality scores low in that model. Therefore authenticity and individuality become the scarce premium.

Emerging opportunity areas: (1) a verification and proctoring economy — AI makes fabrication trivially easy, so a large chunk of economic activity shifts toward auditing and verification; (2) physical AI — unlike digital work ("is an essay ever done?"), task completion is measurable in XYZ coordinates, and Chinese-style robotics is increasingly visible; (3) offline human communities — the arts-and-crafts-movement analogy: a reaction to mass-produced digital content drives demand for in-person, handmade, authentic experiences (Never School / NS.com cited as an example of the trend); (4) industry-specific disruption is fractal and non-obvious — in advertising AI automates paperwork, not creative (counterintuitive); in consulting the excitement is AI-powered customer surveys (10,000 respondents in 24 hours), not PowerPoint generation. Only people inside an industry know where the real billion-dollar opportunities are. On education, the "webware" thesis: AI is only as good as the human prompting it — offline exams, pencil-and-paper tests, and longhand prompts ensure genuine understanding before AI use; the goal is to be upstream of AI, not downstream.

The four focus areas distilled in this session: (1) personal, private, programmable AI; (2) physical AI — robots in the real world; (3) human communities as a complement to digital AI; (4) verification, proctoring, and auditing.

## Cross-Cutting Themes

**1. Model commoditization — agreed, but partial.** The morning declared commoditization (the telco trap, extractive-to-distributive, fungible tokens); the afternoon (BytePlus's move upmarket to

4K/8K, the scope-problem frame) showed it is confined to lower-tier routine work. Premium persists in high-value verticals and scope-problem domains.

**2. The agent infrastructure stack is taking shape.** Agent-native search (Exa), agent governance and identity (Google's seven principles, zero trust, Model Armor), trusted web data (Bright Data), policy that lives outside the agent (Tencent Core Pro), and agent payments, permissions, and billing (Stripe) — the picks-and-shovels layer of

the agent era is forming fast. Gartner's 5%-to-40% adoption jump and the "100 agents per employee" projection are the demand-side evidence.

**3. Context and durable assets are the new moat.** NBIM ("harness is cheap, context is expensive"), Kevin ("take enterprise context into the specialized model"), Tencent ("durable assets — data formats, benchmarks, governance, knowledge bases that survive model swaps"), and Mistral (proprietary-data partnerships with ASML and Airbus) — four sessions converged on the same proposition.

**4. The rise of the trust and verification economy.** Bright Data's honeypots (plausible false data served to agents), Google's Model Armor (prompt-injection defense), and the AGI session's verification/proctoring economy thesis — the cheaper AI makes fabrication and error, the more expensive and valuable verification becomes.

**5. The end of subsidies, financialization, and pricing redesign.** Token-price normalization, CFO spend caps, and a compute futures market ($600B spot toward a $6-8T derivatives market) — plus, at the app layer, per-seat SaaS mismatched with AI's variable compute costs, moving toward usage, credits, hybrid, and outcome-based models (Stripe). From infrastructure to applications, the AI economy is shifting from growth-at-all-costs to capital efficiency, pricing precision, and financial infrastructure.

**6. Scarcity shifts from the average to the authentic.** In a world where LLMs supply "the average" in unlimited quantity, the premium moves to authenticity, individuality, in-person experience, and physical presence — the things that cannot be downloaded. Jevons's complementary-goods logic explains both physical AI and human communities at once.

**7. AI lowers the floor, not the ceiling.** As the four constraints collapse (Stripe: Atlas cohorts reaching first payment in six weeks; AI companies in 42 countries in year one), the floor for company formation drops — but so does the noise floor rise. The edge moves to taste, judgment, distribution, workflow ownership, trust, and pricing design — echoing the AGI session's "only insiders know where the real opportunities are."