Flash | Citi Research

08 Sep 2026 15:57:21 ET | 9 pages

Artificial Intelligence

Genspark and Cresta Intelligence - Day 1 Takeaways from Citi's Global TMT Conference

Heath Terry[AC]

Shelby Spencer

Ashley Kim

Janna Withrow

CITI'S TAKE

We hosted Wen Sang, the Co-Founder and COO of Genspark, and Ping Wu, the CEO of Cresta Intelligence, for a fireside chat on September 8, 2026, at our Global TMT Conference in NYC. The discussion spanned the current state of agentic AI adoption, token economics and cost management, and the future of work. Key takeaways below.

Agentic AI adoption and AGI diffusion still in its infancy. With ~1B knowledge workers still largely untouched, Genspark expects the volume of knowledge work to increase significantly as agentic AI adoption grows, just as more code has been written since AI-code generation was introduced to ~30M global software developers. The company pointed to early enterprise case studies as evidence of the broader knowledge work opportunity, with Grant Thornton initially deploying Genspark to accelerate RFP responses and more recently exploring deeper integrations into other business areas, framing AI as beginning to compress the ~60% of time knowledge workers spend on non-strategic tasks. Cresta framed contact center AI value across four distinct buckets: eliminating calls that should not occur, automating containment and deflection, augmenting human agent productivity via handle time reduction and back-office QA automation, and enabling net-new revenue conversations that were previously uneconomical, consistent with Jevons' paradox as costs decline.

Model capabilities are advancing, but adoption gaps remain substantial. Both speakers emphasized that while model capabilities are advancing rapidly at the frontier, production-grade agentic AI for general knowledge work remains in early stages, and that distinguishing demos from real production capabilities is increasingly important. Genspark noted that frontier labs are pushing the capability ceiling while open-weight models raise the floor, and that leading financial institutions are deliberately running tech stacks six months behind the frontier to preserve workflow predictability and control. Cresta flagged that there is still work to be done at the fundamental model capability layer (e.g. agents cannot yet reason within sub-200 millisecond latency thresholds, self-improving models remain a work in progress) and the adoption layer for customer intelligence (e.g. tribal knowledge not yet captured, messy enterprise environments, scenarios where real-time APIs are not available). The speakers agreed that coding has pulled ahead as the dominant use case given better verifiability, cost measurability, and bottoms-up adoption dynamics, with nearly 100% of Genspark's own code now written by AI, but that agentic AI for knowledge work will take time given the broader surface area ($30-50T globally, ~1B workers), diversity/complexity of tasks, and longer IT procurement cycles.

See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliati



Token economics and model routing are emerging as core infrastructure decisions. Genspark highlighted its proprietary routing and mixture-of-agents architectures across 70+ frontier and open-weight models, deploying senior models in advisory roles and lighter, purpose-built models for high-volume, lower-complexity tasks. The company measures value on cost-per-task-completed rather than cost-per-token-consumed and noted that model-agnosticism is important to prevent single-provider lock-in, while noting that close partnerships with all frontier labs ensures early access and premium rates. Cresta noted that routing for contact center use cases is relatively static compared to coding given the fixed nature of call center/CX tasks, with small fine-tuned typing models handling keystroke acceleration, and larger multi-billion parameter text and voice models applied to more complex reasoning, and overall token volume naturally bounded by call volumes.