Flash | Citi Research

09 Sep 2026 17:12:14 ET | 9 pages

Artificial Intelligence

AI's Open Models: Poolside - Day 2 Takeaways from Citi's Global TMT Conference

CITI'S TAKE

We hosted Jason Warner, the Co-CEO and Co-Founder of Poolside, for a fireside chat on September 9, 2026, at our Global TMT Conference in NYC. The discussion covered Poolside's Model Factory architecture, its recently announced Nvidia partnership, the open-weight model landscape, and the long-term trajectory of AI in knowledge work. Key takeaways below.

Heath Terry[AC]

Shelby Spencer

Ashley Kim

Janna Withrow

Knowledge work evolution and token economics. Warner described knowledge work, a $30T global market representing 25% of global GDP, as a software-on-compute problem, with models increasingly expected to discover and call APIs and tool chains ephemerally, rendering the agent/application distinction less relevant over time. In terms of economics, Warner believes there is significant waste on token spend right now, as the industry is uniformly treating every task as a frontier problem, even though 90% of knowledge work economic value does not require frontier-models. Warner articulated that many organizations have chosen to build on top of closed frontier models as it is viewed as the safer choice, a trend which could shift were open-weight to achieve frontier intelligence. Warner believes pressure from public company shareholders may drive a more rational distribution of intelligence across frontier, frontier-minus-one, and SLM tiers, pointing to energy, compute, and intelligence infrastructure players as critical market participants over the next decade.

Model Factory's differentiated approach to frontier model development. Rather than following conventional pre- and post-training approaches, Poolside built Model Factory as a meta-platform designed to discover efficient model parameters autonomously, drawing on the core insight of Richard Sutton's Bitter Lesson that applying more compute to simpler search problems yields the most durable advances. The approach has produced early results with Laguna S 2.1, trained on ~4,000 H200s in five weeks end-to-end, performing better than comparable open-weight models despite substantially lower parameter counts, with Warner arguing this challenges the assumption that frontier capability requires billions invested in training runs.

Open-weight competitive differentiation. Warner argued that DeepSeek and Qwen's resource constrained operating environments contributed to their model-building competency, forcing a degree of architectural rigor that certain western labs, with more access to compute, did not have to prioritize to the same degree. That gap, in his view, will widen before it narrows. On compute more broadly, Warner

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



pushed back on the training compute shortage narrative, characterizing the constraint as a physics and economics problem, with the largest models trained today not economically viable to serve at scale. Warner described several large hyperscalers' pivot toward inference as reflecting a potential absence of economic justification for continued training capex, rather than supply shortage.