Takara Launches Miru to Solve Enterprise Token Crisis as AI Budgets Soar

Takara's new infrastructure layer solves the 'FinOps paradox', enabling enterprises to slash AI coding spend by 40% while supercharging developer productivity.

LONDON, UNITED KINGDOM, August 11, 2026 /EINPresswire.com/ -- Takara, an AI technology company, today announced the launch of Miru, an AI infrastructure layer purpose-built to optimise the economics of enterprise AI code generation. Enterprise organisations are experiencing a significant FinOps paradox as the use of continuous AI coding assistants continues to become widespread throughout software development teams. Although the average cost per token (the amount charged for processing a given input) continues to decline, the total amount of AI enterprise invoices is rapidly increasing as a result of a large number of context retrieval and loading requests.

> Miru shows a coding agent exactly where to look, so it stops paying to read code it doesn't need. Same agent, same questions, meaningfully cheaper and faster”
>
> *Marty Legg, CEO at Takara*

Most traditional retrieval pipelines will consume anywhere from 80% to 90% of the token budget for an AI coding agent just to search through the repositories and load in context before the agent ever writes one line of code. By transferring the context discovery and indexing of the repositories from expensive GPU-based inference using LLMs (Large Language Models) to an ultra-fast, low-cost CPU infrastructure, Miru cuts token waste by up to 50% and reduce total spend by approximately 40% on AI coding while maintaining the quality of the code generated.

The launch has directly responded to the emerging industry-wide push for standardised tokenomics. This effort is being led by the Linux Foundation and is focused on establishing a framework for governance and the cost of token consumption. Even though tools like Claude Code, Cursor, and GitHub Copilot have significantly increased developer productivity, the fact that these tools have to continuously scan repositories has revealed the performance and cost limitations of this approach and the resulting explosion in token consumption and budget overruns.

According to Takara's CEO Marty Legg, “Software development is the first workload at scale where AI's inefficiency becomes visible on the balance sheet." Legg continued, "Miru shows a coding agent exactly where to look, so it stops paying to read code it doesn't need. Same agent, same questions, meaningfully cheaper and faster - and it works with whatever tools a team already uses and saves up to 60% of the tokens."

Miru is built to enable companies to adopt it easily within their existing tools and platforms without disruption. Miru processes code 50 times faster and indexes 200 times faster compared to other options on the market today, allowing code repositories to stay up-to-date, real-time, and in context. Miru not only provides speed for development teams but also enables strict compliance with enterprise governance policies by enabling sovereign deployment options for customers. Companies can deploy the licensed versions of Miru within their local data centres, or cloud environments ensuring that the source code never leaves their secure, controlled infrastructure.

Jordan Legg, Chief AI Officer at Takara, explained the approach, Miru is a custom designed harness to optimise the developer experience and token consumption. Miru is uniquely underpinned with an ultra-fast embedding model called ds1-code, which runs on ordinary CPU instances, and enables the context and retrieval to be done ultra-fast and for very low cost whilst being incredibly accurate. This means Miru unloads the token heavy context and retrieval into low-cost infrastructure whilst leaving the expensive LLM to do what it is best - reasoning and code generation. – Use the right tool for the right job."

As enterprises move toward autonomous agentic workflows across finance, operations, and engineering, efficient context management becomes paramount. Miru provides the fundamental building blocks to ensure organisations can scale multiple AI agents in the same workspace while achieving a clear financial and operational ROI.

Miru is available immediately via Takara's public API and through private enterprise licensing. For more information, [visit takara.ai](takara.ai).

Carlos Miguens.
Takara
email us here
