The Fast Product Tagging Engine for the Agentic AI-First Commerce Era

E-commerce is shifting toward agentic AI, and Avian.io demonstrated how to support this evolution by achieving 351 tokens/sec on DeepSeek R1 using Vultr Cloud GPU, accelerated by NVIDIA HGX B200, enabling real-time product tagging at scale.

vultr.com

VULTR avian

Use Case

Fastest Agentic AI Tagging for E-Commerce: 351 TPS with Avian.io and Vultr

AI in commerce, preparing for the agentic shift

Agentic AI is becoming the new front-end of commerce. By 2030, one-sixth of online shoppers – around 500 million people – will interact with brands through their own AI agents[1]. These agents will bypass traditional websites, instead pulling product data directly from APIs and structured sources. IDC notes that this shift from "knowledge to action" will transform GenAI tools like chatbots into autonomous agents executing complex workflows. Commerce is moving from clicks to conversations, and retailers must adapt now.[2]

Vultr and Avian.io deliver the performance and scalability for AI-first commerce by enabling real-time product tagging, categorization, and description generation. This transforms product catalogs into structured, machine-readable data ready to interact with AI agents. According to Gartner, reducing friction, improving agility, and enabling instant response to machine-led interactions are key priorities. With high-speed infrastructure and global GPU resources, this setup supports commerce platforms in preparing catalogs for discovery through agentic AI interfaces.[1]

Challenges: Enabling agentic AI for modern retail

Retailers are entering an AI-native era, but most systems need upgrades to fully support agentic AI. Key areas to address include:

- Product data needs to be machine-structured: Most catalogs are still built for human browsing, but to support AI agents, product data must be consistently tagged and structured.
- Inference must be faster and more efficient: Modern models need low-latency, high-throughput inference, which many platforms struggle to deliver without high infrastructure costs.
- Data systems need to be unified: Product, pricing, inventory, and content often sit in silos. AI needs unified, accessible data to generate real-time responses.
- Infrastructure should be GPU-ready: Legacy cloud stacks aren't built for AI. Retailers need platforms optimized for accelerated computing without orchestration complexity.
- APIs must support agent interaction: AI agents use APIs, not websites. Retailers need systems that handle fast, automated product queries at scale.

How Vultr and Avian.io help retailers compete

Avian.io automates real-time product tagging, categorization, and description generation, turning raw catalogs into structured data for AI agents. Running on Vultr's high-performance GPU infrastructure, the solution supports fast, scalable inference without orchestration complexity or idle compute. Retailers gain a practical way to serve AI-ready catalogs globally, stay visible to agentic interfaces, and compete as commerce shifts from search to AI conversations.

Industry
Retail and eCommerce

Use Case
Agentic AI Tagging

Key Technologies
DeepSeek R1, TensorRT-LLM, NVIDIA HGX B200 (8x Blackwell GPUs), Vultr Cloud GPU, FP4 Quantization, MTP, FusedMoE, Custom CUDA Kernels

Customer Spotlight
Avian

"This benchmark is a milestone for the AI-first commerce ecosystem. Running DeepSeek R1 with TensorRT-LLM on Vultr's HGX B200 platform allowed us to push inference speed to 351 tokens/sec. Vultr gave us the raw GPU power and global scale, without Kubernetes or orchestration complexity. We're proving that it's now feasible to serve structured product data directly to AI agents with enterprise-grade performance."

Jessup Jong,
COO at Avian.io

[1]Gartner, The Future of Digital Commerce in 2030
[2]IDC, Understanding the Shift to Agents

Serverless LLM performance with Avian.io on Vultr Cloud GPU

Avian.io provides serverless AI endpoints optimized for high-speed LLM inference, using custom CUDA kernels, efficient batching, and orchestration techniques. These endpoints run on Vultr Cloud GPU infrastructure, combining low-latency performance with global availability and predictable pricing, making it easy to scale production-ready inference without managing complex infrastructure.

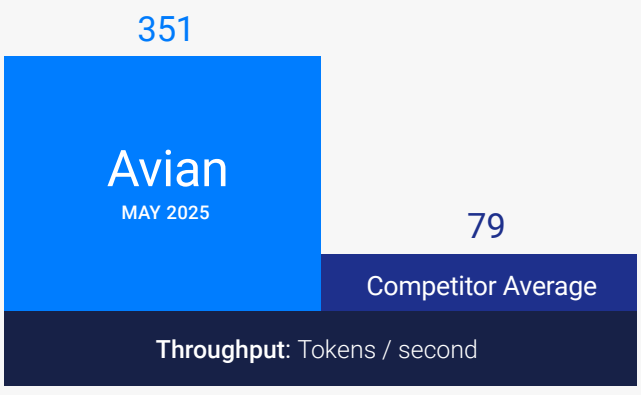
Why it matters for e-commerce

As AI agents like ChatGPT and Gemini become a key way customers discover products, structured data replaces the traditional storefront. Visibility depends on how quickly and accurately information is delivered. Avian.io's 351 TPS benchmark shows that product tags, categories, and descriptions can now be served at agent speed, keeping retailers discoverable, competitive, and ready for AI-first commerce.

Benchmarking real-time catalog structuring

This use case demonstrates how Avian.io's product tagging engine streamlines catalog structuring by generating tags, categories, and descriptions from raw SKU data and turning unstructured catalogs into machine-readable formats. Built on DeepSeek R1 and optimized with advanced inference techniques, the model runs on Vultr Cloud GPU infrastructure, accelerated by NVIDIA HGX B200, delivering low-latency performance for commerce applications workloads.

Benchmark



Tested on Vultr Cloud GPU, accelerated by NVIDIA HGX B200. Results based on average tokens/sec from MT Bench tasks. Accuracy verified. Competitor scores use the higher of the mean or the median from ArtificialAnalysis.

Challenge

Deploying real-time product tagging for agentic commerce requires high token throughput to maintain accuracy, especially for structured data generation. Low-latency inference is essential to meet the demands of AI-driven workflows. Traditional orchestration adds complexity and delays. Many cloud platforms are not optimized for accelerated inference at scale. Keeping catalogs updated without interrupting service is also a key challenge for commerce.

Solution

The setup uses NVIDIA HGX B200s on Vultr Cloud GPU with high memory bandwidth and fifth-gen tensor cores for efficient, high-speed inference. Avian.io applies FP4 quantization, speculative decoding, and custom fused kernels with TensorRT-LLM to maximize throughput. Direct deployment on Vultr eliminates the need for Kubernetes, reducing complexity. The system supports continuous tagging, allowing catalog updates without retraining or downtime.

Results

Avian.io set a world record by reaching 351 tokens per second on DeepSeek R1 for math and reasoning tasks. The benchmark was achieved on Vultr's Cloud GPU platform using NVIDIA HGX B200 with 8x Blackwell GPUs. Avian maintained accuracy while maximizing throughput. It was built on a PyTorch-based TensorRT-LLM stack and enhanced with FP4 quantization and custom inference optimizations like speculative decoding and FusedMoE. The result demonstrates that open-source models can deliver enterprise-grade performance with low-latency, cost-efficient inference, making real-time AI more accessible for practical use in commerce and beyond.

Learn more about
Vultr Cloud GPU

Contact our sales team, or visit vultr.com to learn more.

Additional resources

Avian blog: *Record-breaking AI inference* →
Whitepaper: *Agentic AI white paper* →