VISHNU ANILKUMAR

Lead Machine Learning Engineer · Production GenAI · LLM Agents & MCP · MLOps
Bengaluru, India · +91 8943372700 · vishnuanilkumar.engineer@gmail.com
linkedin.com/in/vishnuverse · github.com/vishnuverse · vishnuanilkumar.com

May 02, 2026

Dear Hiring Team,

Over the last two years I've built and shipped the AI Python backend for a production GenAI platform that grew from zero to **$2M+ ARR** — a 46-route orchestrator that fans a single ad brief through 12 downstream stages: prompt intelligence, diffusion generation, a 7-tool image editor, a product-to-video studio with camera choreography, multi-provider LLM ad copy, and brand-QA. Under it sits a unified provider abstraction that routes across Baseten, FAL, Replicate, OpenAI, Anthropic, and Gemini behind a webhook-driven async fan-out, with cost-tier draft/polish routing that cut generation spend **40–60%** at no quality cost.

In parallel I shipped two production **MCP servers** — one for workspace-scoped competitor-ad intelligence (11 tools, Streamable HTTP, integrated with Claude Desktop and Cursor), one for creative QA with multi-modal routing across Claude Vision, GPT-4o, and Gemini. What I'm most proud of is making this serviceable at scale: **9 zero-downtime SQL migrations**, two isolated ARQ worker pools, 1,000-variant bulk fan-out with parent-child task aggregation, and full New Relic-backed observability.

If you're wrestling with the production end of GenAI — cost, reliability, brand-voice grounding at scale — I'd love to talk.

Best regards,

Vishnu Anilkumar

vishnuanilkumar.engineer@gmail.com · +91 8943372700