

Patrick Ward

Software Engineer · AI Specialist

Farmington, CT · pward17@gmail.com · github.com/pward17 · patrickward.info

SUMMARY

Software engineer and AI specialist who ships production LLM systems end to end: retrieval, agents, tools, billing, Kubernetes and observability. Bootstrapped the Cloud37 platform for building and selling AI agents, running more than 45 models from OpenAI, Anthropic, Google and Perplexity at over \$10,000 a month in model spend. Launched BlinkBook (blinkbook.ai), which earned over \$1M in its first two weeks. A background in marketing technology and field service keeps the focus on business results, not just the model.

AI COMPETENCIES

- **Context engineering:** per-model token budgets, rolling summaries, tiered attachment degradation, lazy-loaded agent skills, and a separate budget for each retrieved document so answers stay accurate and cheap.
- **Vector search and RAG:** embedding pipelines fed by change data capture, structure-aware chunking, hybrid Qdrant and Elasticsearch search with rank fusion, LLM query expansion, self-query filters, reranking, and tenant-scoped retrieval that degrades gracefully.
- **Tools and data access:** function-calling tools, 25 third-party connectors, and a sandbox that lets models run SQL against live business data behind parser allowlists, scoped roles and read-only transactions.
- **Agent architecture:** agent graphs where agents call sub-agents, background sub-agent tasks, LLM routers, deep-research dispatch and a copilot that builds agents.
- **Prompt engineering:** versioned prompts and agents with snapshot and restore, prompt-cache-friendly prompt layout, and refinement driven by test runs.
- **Feedback and evaluation:** user votes sent to traces, scripted multi-turn scenario tests scored by LLM judges, and knowledge-gap reports back to agent owners.
- **Tracing and observability:** LangSmith, Sentry, Prometheus and Grafana on every model call, with per-call cost.
- **Self-improving systems:** loops that turn feedback, test results and knowledge gaps into prompt and knowledge improvements.
- **Prompt-injection defense:** guardrails for untrusted surfaces, retrieved and tool content treated as data, single-pass payload substitution, SSRF blocklists and response redaction.
- **Cost and scale:** over \$10,000 a month in model spend, token accounting across input, output, cached and reasoning tokens, prompt caching on Anthropic and OpenAI, circuit breakers, and reconciliation against provider bills.
- **Business application:** finding where AI pays off in a business and shipping it, including BlinkBook, which earned over \$1M in its first two weeks.
- **Training:** working with many client teams to map their business processes, add AI at the steps where it saves the most time, and teach the team to use it in their daily work.

EXPERIENCE

Cloud37 · Software Engineer and AI Specialist (founded as Xngen, then Xagency) 2024 to present

Cloud37 is a multi-tenant SaaS platform where businesses build, test and sell custom AI agents, with a knowledge library, tools, sub-agents, triggers and paid white-label storefronts. Clients either used it self-serve or had us run it as a done-for-you service, with our team setting up their AI on the platform.

- Launched BlinkBook (blinkbook.ai) for Knowledge Source: the company was founded to build it, I built the demo that started the project, and it earned over \$1M in its first two weeks.
- Bootstrapped the platform codebase in 2024 and stayed a core engineer through launch and scale: about 1,100 commits and 270,000 lines across a Python/FastAPI backend and a Next.js/React frontend.
- Built "Tables", which lets agents query and write business data in SQL safely. Model-written SQL is parsed against an allowlist, rewritten with tenant scope, and run as a per-company read-only Postgres role with timeouts and row caps. 323 tests, including security tests against real Postgres.

- Built the connector framework that gives agents tools in Gmail, Google Calendar, Drive and Sheets, Slack, Notion, GitHub, Outlook, Stripe, QuickBooks and more (16 of 25 toolkits), with per-call usage metering.
- Built token accounting and cost controls across more than 45 models: reconciliation against OpenAI, Anthropic and Google usage APIs, and hierarchical token-usage reporting for admins.
- Built deep research, slash commands and new-model onboarding for the agent chat.
- Built Expert Studio, the white-label client where end users pay for and chat with agents: streaming proxy, HttpOnly-cookie JWT auth with refresh tokens, free signup, and customer domains.
- Kept the RAG pipeline reliable: answers from every available search engine, a Qdrant time limit inside a 45-second budget, a Firecrawl v2 migration, and vision parsing that streams PDF pages one at a time to keep memory low.
- Moved the platform from Railway to DigitalOcean Kubernetes (develop, staging, prod), moved Celery from RabbitMQ to Redis, and added Prometheus, Grafana, Fluent Bit, Sentry and LangSmith tracing.
- Contributed across the team's AI platform: layered conversation memory with hybrid Qdrant and Elasticsearch retrieval, prompt caching, LLM-judged agent scenario tests, sub-agent orchestration and prompt-injection guardrails.

GrowRev Digital · Tech Engineer and Media Buyer, Farmington, CT

2021 to 2024

- Ran Google Ads (Search, YouTube, Display, Performance Max), Facebook and Rumble accounts for clients.
- Used AI to generate ad creative and copy.
- Designed and implemented tracking: custom Facebook and Google events, Google Tag Manager with server-side Conversions API, data-layer pushes, and JavaScript event triggers.
- Built reporting dashboards in Looker Studio.
- Automated client workflows with Zapier, GoHighLevel, Hyros, WordPress, Microsoft and Google Cloud.

WorldWide Tech Services · Service Technician, Farmington, CT

2016 to 2021

- Installed, serviced and managed network equipment for TD Bank.
- Certified on-site technician for IBM, Lenovo, Dell, HP, Sony, Best Buy, Subway and Quest Diagnostics.

TECHNICAL SKILLS

- **Languages:** Python, TypeScript, JavaScript, SQL, Java, PHP, C, HTML/CSS
- **AI and LLM:** OpenAI, Anthropic Claude, Google Gemini (Vertex), Perplexity, LangChain, LangSmith, OpenAI embeddings, Whisper, Firecrawl, LlamaParse, Neo4j
- **Search and data:** PostgreSQL with pgvector, Qdrant, Elasticsearch, MongoDB, Redis
- **Backend:** FastAPI, SQLAlchemy, Alembic, Celery
- **Frontend:** Next.js, React, Tailwind, React Query, Zustand
- **Infrastructure:** Docker, Kubernetes, Kustomize, GitHub Actions, DigitalOcean, Caddy, Prometheus, Grafana, Fluent Bit, Sentry
- **Integrations:** Stripe and Stripe Connect, Clerk, Composio, Google Tag Manager, Meta and Google Conversions APIs
- **Networking:** TCP/IP, DNS, DHCP, VLANs, Ethernet switching

EDUCATION AND CERTIFICATIONS

- **Tunxis Community College**, Farmington, CT: Computer Programming, 2006 to 2008
- Certified on-site service technician: IBM, Lenovo, Dell, HP

VOLUNTEER

- Field Marshal, Bernie Ward Farmington Soccer Labor Day Shootout, 2002 to present