

BRANDON BORBA

Applied AI Engineer | Production AI Agents & Copilots | LLMs · VLMs · Full-Stack Systems
San Jose, CA | 408-455-6332 | brandon@borbai.com | borbai.com | linkedin.com/in/brandonborba

SUMMARY

Applied AI engineer who ships. As the sole AI engineer at a 750+ person PCB manufacturer (ITAR-constrained), designed and deployed a production multi-stage LLM/VLM pipeline now processing ~280 jobs/week at 98.8% accuracy, returning ~117 hours/week (~2.9 FTEs) to the team. Equally at home shipping full-stack production software: ERP-integrated Django apps and large-scale legacy migrations, delivered end-to-end, with results reviewed directly by the CEO and executive team.

SKILLS

Generative & Agentic AI: LLMs, Vision-Language Models (VLMs), agent orchestration, multi-stage AI pipelines, agentic RAG, copilot experiences, prompt engineering, structured extraction, fine-tuning (Qwen 7B), quantized model serving, evaluation harnesses

Frameworks & Libraries: Python, PyTorch, Hugging Face Transformers, vLLM, Ollama, OpenVINO, FastAPI, Django, htmx, Tkinter, PyQt

Deployment & Infrastructure: On-prem inference, Docker, NVIDIA DGX, GPU serving, CUDA-level debugging, AWS, AWS GovCloud, headless services, ERP/system integration, AI-augmented development (Claude Code, Cursor)

Data & Integration: SQL Server, ODB++, multi-source data integration, OCR pipelines, document intelligence, data migration, ITAR-compliant environments, Genesis 2000 (PCB CAM)

EXPERIENCE

AdvancedPCB — Santa Clara, CA

Mar 2025 – Present

AI Automation Engineer — Applied AI Systems (sole AI engineer)

- Built and shipped a production multi-stage AI pipeline processing ~280 jobs/week at 98.8% extraction accuracy, manual-audited across 1,000+ production jobs. Orchestrates a vision-language model across page classification, fab-drawing extraction, stackup parsing, ODB++ integration, and cross-pass reconciliation, then auto-populates downstream CAM/quoting systems; cut per-job entry from ~30 to ~5 minutes, returning ~117 hours/week (~2.9 FTEs).
- Deployed a quantized 122B-parameter MoE (Qwen3.5-122B-A10B, AWQ 4-bit) on-prem via vLLM in Docker on a single NVIDIA DGX Spark that I evaluated, procured with CEO approval, and stood up inside a fully cloud-restricted ITAR network: 91 days continuous uptime with zero restarts, 47,973 requests with zero server errors, 488.5M tokens processed, and a 136-second measured kill-to-healthy recovery drill, unattended.
- Drove model selection through empirical evaluation of ~20 open-weight candidates (7B–122B): started with a two-model pipeline (Qwen 7B for OCR, GPT-OSS 20B for reasoning), fine-tuned Qwen 7B for structured outputs, then consolidated to a single multimodal model once capability matched, simplifying the system without sacrificing accuracy; currently benchmarking Qwen3.8-27B against the same 1,000-job gold set, within days of its Aug 2026 open-weight release, to decide whether it replaces the deployed 122B.
- Designing an evidence-cited planning co-pilot (CEO-assigned, in progress) that attempts every traveler field with citations from ERP records, fab drawings, and prior builds, flagging deltas from the previous build so planners audit rather than re-derive in a quality-critical, human-in-the-loop workflow.
- Automated a multi-site CAM hand-copy step by rewiring the form-filler to pull stackup and scaling values live from the planning system's REST API; ported the UI from PyQt to tkinter to run on a legacy RHEL 6 server, validated 101/101 offline harness assertions plus 6 real jobs end-to-end, and rolled out across 4 manufacturing sites.
- Ported a ~30,000-line, 30-module legacy Perl application (IPC-2221 coupon generator driving the Orbotech/Genesis 2000 CAM system) to ~25,000 lines of Python across 37 modules; proved byte-for-byte identical output with a dual-interpreter protocol-replay oracle and built 704 automated tests on a codebase that had zero, then shipped all 9 management-requested features (automated board markings, XRF measurement via REST, config-driven behavior gates) across 25+ releases in ~6 weeks, root-causing a latent bug in the 20-year-old original along the way. The port replaced the Perl in production.
- Built a greenfield Django + htmx production scheduling app integrated with the ProCIM ERP across two manufacturing sites, replacing manual spreadsheet workflows; delivered across 8 reviewed-and-merged pull requests (~14K LOC) with SQL Server migrations, test coverage, and production fixes. Separately eliminated a ~\$20K/year Smartsheet subscription with three native Django tracking apps and lossless historical migration via a live API sync.
- Presented progress, ROI, and technical decisions to executive and board-level stakeholders, supporting continued investment in applied AI.

Borbai — Independent AI Automation Practice

2024 – Present

Founder / Engineer

- Build and maintain AI-driven workflow automation, lead-generation systems, and CRM integrations for 5 ongoing SMB clients across the Bay Area and Los Angeles. End-to-end ownership: scoping, build, deployment, and iteration.

Summit Interconnect — Santa Clara, CA

Summers 2021, 2022

Developer Intern / Lab Support Engineer

- Programmed PCB manufacturing automations using Genesis 2000 and advanced Excel logic; maintained run documentation and validated parts in a high-spec manufacturing environment.

EDUCATION

University of San Diego — B.B.A., Business Analytics

2024

Valley Christian High School — San Jose, CA

2020

Volunteer: Georgia Travis Family Shelter