

TaSain Thomas

(586) 873-0656 | tasain@onwheelsdetailing.com | github.com/T-Thomas1 | linkedin.com/in/tasain-thomas-539867249

SUMMARY

AI Engineer who ships production-grade LLM systems. Deploys fine-tuned VLMs, RAG pipelines, and autonomous agents on Kubernetes with a core focus on safety, red-teaming, and adversarial robustness. Owns the full ML lifecycle - from LoRA fine-tuning and transformer architectures to zero-trust cloud deployment - while maintaining 99.9% uptime. Translates research into business-ready features in Agile environments. Open to onsite.

EDUCATION

Wayne State University - B.S. Computer Science | May 2026 | GPA: 3.4/4.0

Trustworthy AI Coursework: LoRA/QLoRA, RLHF/DPO/GRPO, Red Teaming, Machine Unlearning (TOFU/WMDP), Agentic AI Safety, Grounded VLMs (LISA/GLaMM), Chain-of-Thought Reasoning, HarmBench/SafeChain

TECHNICAL SKILLS

AI/ML: PyTorch, TensorFlow, Hugging Face, LangChain, LoRA/QLoRA, RAG, Vector Databases, Transformer Architectures, Attention Mechanisms, Scikit-learn, Pandas, NumPy

MLOps & Cloud: Kubernetes, Docker, AWS/Azure, DigitalOcean, Nginx, CI/CD, Proxmox, GPU Passthrough, Model Serving, Model Monitoring, Helm, Terraform

Security: Prompt Injection Defense, Red Teaming, Zero Trust Architecture, Kali Linux, Wireshark, Adversarial Robustness, OWASP Top 10 for LLMs

Languages & Tools: Python, SQL, Java, C++, JavaScript, PostgreSQL, MySQL, Git, Agile/Scrum, TDD

AI ENGINEERING PROJECTS

Secure AI Pipeline for BMI Estimation | github.com/T-Thomas1/BMI-Estimator | Apr 2026

- 89% accuracy (MAE 1.24, MAPE 3.9%) via LoRA fine-tuning 7B VLM; 35M params, 95% cost reduction
- Privacy layers: face blurring, synthetic data (BEDLAM), LLM-as-Judge (GPT-5.4 nano)
- 40% hallucination/refusal reduction; CoT reasoning + zero-trust validation > HarmBench benchmarks

MLOps & AI Agent Engineer | Enterprise Infrastructure | Mar 2026 - Present

- Auto-scaling Kubernetes inference cluster (Hermes Agent + DeepSeek), GPU passthrough, 99.9% uptime
- Red-team testing (Kali Linux) embedded in CI/CD - pre-production prompt injection patching
- Automated monitoring & alerting pipeline - cron-scripts, real-time incident response

Full-Stack Engineer | SaaS Booking Platform | github.com/T-Thomas1/on-wheels_detailing-crm

- Production CRM (Python/JS) + Stripe/Telegram APIs; multi-location, 6-tier pricing
- DigitalOcean, Nginx, systemd, Cloudflare (DNS/SSL/CI/CD) - 24/7 uptime, 100% incident handling

Inventory Management System | Detroit Business | github.com/T-Thomas1/inventory-website

- Custom full-stack inventory web application for a local Detroit business

Magna Automation - Production Line Assignment Engine | Sep 2023 - Sep 2024

- C++/SQLite3 worker-to-station matching system; evaluateFit() scoring algorithm balancing
- line familiarity, station expertise, and employee rank; 40% reduction in assignment time

PROFESSIONAL EXPERIENCE

Industry 4.0 / Automation Intern (Rotational) | Magna Electric Vehicle Structures

Sep 2023 - Sep 2024

- 4 automation projects on-time, API integrations for real-time production data pipelines
- 33% throughput via Vuforia Model Targets; 25% debug reduction via custom C# logic
- 480-hour rotational program across 2 manufacturing roles

Registered Behavior Technician | BlueMind Therapy LLC | Sep 2024 - Present

- Data-driven decision-making: real-time behavioral data analysis, daily intervention adjustments
- 100% audit-ready compliance; mirrors model drift monitoring and A/B testing in production ML

Keywords: RAG, Vector Databases, Transformer Architectures, Attention Mechanisms, Model Serving,

Model Monitoring, Helm, Terraform, Zero Trust Architecture, Adversarial Robustness, A/B Testing