

Zoran Stojanovic

Senior Applied AI & Automation Engineer | Multi-Agent Orchestration & Deterministic Systems

📍 Belgrade, Serbia • Remote (Full US / EU Overlap) • Ready for B2B Contractor / EOR

Kaggle Top 1.2% Expert (#734 / 61,348)

✉ kiza101288@gmail.com 🌐 kiza-portfolio.pages.dev 🐙 github.com/kizabgd123 📄 kaggle.com/kiza123123
🌐 linkedin.com/in/kizabgd123

⚡ PROFESSIONAL SUMMARY

Senior Applied AI & Automation Engineer with 7+ years of expertise in architecting **autonomous multi-agent workflows, deterministic LLM governance runtimes, and high-reliability machine learning systems**. Creator of **JudgeGuard**, an open-source multi-layered deterministic verification engine turning probabilistic LLMs into zero-trust, production-grade automation systems with automated reality-grounding. Ranked in the **Top 1.2% worldwide on Kaggle (Expert, #734 of 61,348+)**. Proven engineering capability building AI developer tools for in-browser runtimes (**WebContainers / Bolt.new**), enterprise operations automation, universal RAG engines (ChromaDB / Llama-3.3), and rigorously calibrated healthcare/clinical risk predictors (MedGemma, Brier Score calibration).

</> TECHNICAL STACK & CORE COMPETENCIES

AGENTIC SYSTEMS & LLM GOVERNANCE

Multi-Agent DAGs Adversarial Debate

Elo Consensus JudgeGuard Safety Engine

LangChain Function Calling

Zero-Trust Guardrails

RAG & CONTEXT OPTIMIZATION

ChromaDB Pinecone Dense/Sparse Hybrid

Context Window Pruning Surgical Token Diffs

ResearchProAI Ingestion

APPLIED ML & PREDICTIVE MODELING

Probability Calibration (Brier) LightGBM

CatBoost PyTorch HuggingFace LoRA/PEFT

Vision Ensembles

SOFTWARE ENGINEERING & ENVIRONMENTS

Python (Expert) Go TypeScript/JS FastAPI

WebContainers Docker Cloudflare Workers

Git CI/CD

 FEATURED PRODUCTION & OPEN-SOURCE ENGINEERING**JudgeGuard Core & Autonomous Multi-Agent Governance Runtime** [Lead Architect](#)

Python • Multi-Agent • Security

- Designed and open-sourced an enterprise multi-tier governance runtime enforcing deterministic verification on autonomous AI agents before and after code execution.
- Engineered **Layer 00** (regex blocks on destructive actions), **Layer 000** (Ollama Gemma 4 gatekeeper), **Layer 3** (semantic drift detection under <0.40 threshold), and **Layer 4** (physical reality verification inspecting disk state, SHA256 hashes, and non-empty byte outputs).
- Eliminated mock hallucination loops and false positive completion claims, enforcing mandatory git checkpointing with SHA recording on every successful task.

Bolt.new & WebContainer Optimization Engine (git_bolt_jules_gate.py) [46k+ Tokens Saved / Run](#)

Python • WebContainers • DevTools

- Engineered an automated context-reduction and surgical diff generator specifically designed for browser-based IDEs (**WebContainers / Bolt.new**).
- Developed token-cleaning protocols isolating unneeded cache directories (``node_modules``, ``.cache``, ``dist``), saving **46,000+ prompt tokens per session** and preventing context overflow in browser runtimes.
- Integrated live LLM telemetry tracking latency, prompt/completion tokens, and cost estimates per execution.

Autonomous Developer Agent v2.0 (Google Gemma 4 Competition) [43% Faster Execution](#)

Gemma 4 31B • Kaggle • Coder Agent

- Engineered a single-agent autonomous coder running Gemma 4 31B, calibrated for competitive problem-solving across held-out, unseen software repositories.
- Formulated a strict **call-paced doctrine** (Calls 1–8: semantic read, Calls 9–12: atomic edit, Calls 13–20: test verification, Call 20+: diff checkpoint), reducing execution latency by **43%** while staying within a 12-hour air-gapped evaluation ceiling.

Clinical Risk Prediction Pipeline & Healthcare AI (MedGemma) [Top 1.2% Ranked](#)

MedGemma • Scikit-Learn • RSNA Apex 0.9576

- Built an end-to-end clinical risk prediction pipeline, evaluating 15 model iterations to optimize the trade-off between **Brier Score (probability calibration)** and F1 score, preventing overconfident false positives in medical decision support.
- Created *MyHealth Decoder* (MedGemma Patient Advocate) for the Google Med-Gemma Impact Challenge; achieved **0.9576 Apex score** in the RSNA Knee Abnormality Detection medical imaging competition.

Diskort — Asynchronous Go Automation Suite [123+ GitHub Stars](#) [Go](#) • High-Concurrency • Automation

- Authored a high-throughput, concurrent automation and bot orchestration suite in Go, achieving 123+ GitHub stars with zero memory leaks under prolonged execution.



PROFESSIONAL EXPERIENCE

Senior Applied AI & Automation Engineer • [Autonomous Systems & Consulting](#)

2023 – PRESENT • REMOTE

- Architected production multi-agent workflows and autonomous copilots for international clients, automating technical intake, PO validation, and CRM/ERP integrations.
- Designed and deployed production RAG solutions utilizing ChromaDB, hybrid BM25/vector search, and dynamic rerankers, achieving **>94% retrieval accuracy** on unstructured corporate data.
- Built zero-trust LLM middleware and pre/post-execution verification systems that reduced unhandled LLM exceptions and hallucinations by **99% in production**.
- Optimized agent prompt context consumption through surgical diffing and AST-based code analysis, lowering API token expenditure by **35–50%**.

Lead AI Automation & Machine Learning Engineer • [Analytics & Core Systems](#)

2020 – 2023 • REMOTE

- Spearheaded development of automated data pipelines and machine learning models for high-stakes tabular datasets and business intelligence systems.
- Implemented automated ETL and feature engineering pipelines in Python, SQL, and Docker, reducing manual operational processing time by **80%**.
- Enforced probability calibration and validation standards across predictive models, ensuring strict compliance with operational risk parameters.
- Integrated Datadog Synthetics, automated unit/integration testing, and CI/CD pipelines across distributed services.

Data Scientist & Software Engineer • [Systems & Model Deployment](#)

2017 – 2020

- Engineered predictive algorithms, classification models, and data exploration tools using Python (pandas, scikit-learn, NumPy).
- Developed RESTful APIs with FastAPI and Flask to serve real-time model inferences to client-facing web applications.
- Automated database ETL workflows and designed optimized SQL schemas for high-throughput transactional databases.



HONORS, AWARDS & COMMUNITY IMPACT

- **Kaggle Notebooks Expert:** Ranked **#734 out of 61,348+ worldwide (Top 1.2%)**. Awarded multiple medals across competitive ML & deep learning.
- **GitHub Developer Program Member:** Active open-source maintainer with 60+ public and private repositories focusing on autonomous systems and developer tooling.
- **Open Source Contributor:** Creator of *Diskort* (123+ GitHub Stars) and *JudgeGuard* reality-grounding framework.
- **Participant:** Google Med-Gemma Impact Challenge (Clinical Decision Support & Medical LLMs).