MUHAMMAD ENRIZKY BRILLIAN

SOFTWARE ENGINEER | AI INFRASTRUCTURE, MODEL SERVING, DATA PLATFORMS

billy.suharno@gmail.com | linkedin.com/in/enrizky-brillian | github.com/billy-enrizky | Toronto, ON

SUMMARY

Software engineer in IBM's AI and Automation Practice. Built self-hosted inference and data platforms that cut inference cost, improved analyst efficiency, and reached pre-production. Created OpenBrowser, an open-source browser agent with 53,327 PyPI downloads and two research papers.

TECHNICAL SKILLS

Languages	Python, SQL, TypeScript, JavaScript, R
LLM Serving	Self-hosted vLLM, SageMaker, Bedrock, speculative decoding/MTP, FP8 quantization, tensor parallelism, KV cache, OpenAI-compatible APIs, streaming SSE, structured outputs
Data	BigQuery, Snowflake, Apache Spark/PySpark, Pandas, NumPy, ETL, partitioning, clustering, data modeling
Backend	Flask, FastAPI, Node.js, Express, REST APIs, MCP, Pydantic, Zod, OOP, concurrency, multi-threaded execution
Browser	Raw CDP, Chrome Manifest V3, ONNX Runtime Web, WebGPU, WASM
Cloud	AWS (ECS, Lambda, S3, VPC, IAM, SageMaker, Bedrock, Cognito, WAF), GCP (BigQuery, Cloud Run, GCS, Scheduler), Terraform, Docker, CI/CD
Testing	pytest, unit, integration, contract, regression, and end-to-end tests
Core	Algorithms, data structures, data-intensive systems, reliability, observability

EXPERIENCE

Software Engineer, AI and Automation Practice

May 2025 – Present

IBM, Toronto, ON

- Deployed Gemma 4 31B as a self-hosted SageMaker endpoint with vLLM 0.32.0, FP8 weight compression, TP=4 on four A10G GPUs, and MTP speculative decoding with 4 draft tokens. Tuned an 8,192-token context, 4,096 batched tokens, 8 concurrent sequences, and 0.92 GPU memory utilization. Production mean acceptance length reached 4.56 with a 45,161-token KV cache (5.51x the 8,192-token window). This kept regulated data in the VPC and removed the managed-model dependency.
- Benchmarked a 31B vLLM deployment on eight A100 GPUs with DP=2 and TP=4. Throughput reached 10,410.60 tok/s at N=512 with zero errors at N=1024, and cost fell from $132.56 per million tokens at N=1 to $0.77 at N=512, a 172x reduction. The result set the concurrency target for production capacity planning.
- Built a Modal-hosted speculative-decoding benchmark for Gemma 4 E2B-it with a 77,993,476-parameter drafter. Across A10, A100-80GB, H100, and B200, 3-run A/B tests at 16 requests and concurrency 1 measured up to 2.06x MTP speedup on A10 code prompts and an 8.9–26.8x vLLM/Transformers gap. The benchmark gives teams a hardware-specific rule for when MTP lowers latency and when it regresses.
- Parallelized generation with a 10-worker ThreadPoolExecutor and built dual Flask services with streaming vLLM calls, JSON-schema output, token-budget management, retries, and background jobs. A controlled full-draft comparison cut generation from 7m48s to 4m52s (1.6x), and a six-week PoC demonstrated 50%+ drafting time savings.
- Built a nightly enterprise finance data platform on GCP that publishes a 1,400,883-row BigQuery fact table with staging, schema and row-count validation, duplicate detection, a run lock, and atomic publication. It replaced manual month-end pulls and tie-outs, supported a >50% Phase 1 analyst efficiency gain, and tied January–September control totals within $0.11.
- Built a pre-production AWS platform for regulated work packages with Python Flask on ECS Fargate, a Vue 3 frontend, asynchronous S3-backed jobs, VPC isolation, Cognito, WAF, and Terraform. The platform reached pre-production across 984 files in 24 weeks.

Data Scientist

Jan 2024 – Aug 2025

Sanofi, Toronto, ON

- Built Python and SQL ETL for high-volume vaccine manufacturing records, with quality-control checkpoints across production lines. The pipeline made production data available for quality analytics and reduced manual preparation.
- Built a two-click Snowflake and Streamlit workflow that moved manual reporting from days to seconds. Led a cross-functional team and introduced sprint planning and CI/CD automation, supporting a 40% team productivity gain.

SELECTED PROJECTS

- Automated 6–7 vendor invoice templates and a point-of-sale close. SME and director sign-off confirmed >75% invoice efficiency, 56–75 analyst hours per year, and 54–72 additional hours per year from the POS workflow.
- Delivered a six-agent text-to-SQL platform at 93% accuracy across five production databases in eight weeks, enabling non-SQL staff to query operations data.

OPEN SOURCE AND RESEARCH

OpenBrowser | github.com/billy-enrizky/openbrowser-ai 53,327 PyPI downloads | 279 GitHub stars

- Built and published OpenBrowser, a Python framework that drives Chrome over raw CDP and executes LLM-written Python in a persistent namespace. It ships a CLI, a single-tool MCP server, a persistent daemon, a FastAPI and Next.js web app, and 14 LLM provider adapters. 53,327 PyPI downloads, 279 GitHub stars, and 52 releases show developer adoption.
- Benchmarked OpenBrowser against three browser-automation tools across six tasks and three runs. It used 36,010 tokens and 84.8 seconds versus 77,123–94,130 tokens and 99.0–118.3 seconds, all at 100% task accuracy. The result cuts token cost and response time for agent workflows.
- Built CI with 128 pytest files and 7,007 test functions across Ubuntu and macOS, and automated releases to PyPI, TestPyPI, MCP Registry, and Homebrew. The pipeline shipped 52 versions through 0.1.54 without manual packaging.

Context Atlas (Jev + Laya) | Chrome Web Store Manifest V3 extension

- Built a source-linked page-search extension. Laya runs local ONNX inference with WebGPU and WASM fallback. Jev provides optional cloud ranking over bounded passages. It highlights the exact source sentence, keeps history local, and limits queries to 400 characters and 160 passages. The default on-device path keeps page text local and avoids cloud inference for routine searches.

Multi-Token Prediction, benchmarked | openbrowser.me/mtp Published benchmark study

- Ran 144/144 successful A/B runs across A10, A100-80GB, H100, and B200 and three prompt regimes. vLLM was 8.9–26.8x faster than Transformers, and A10 was the only GPU with a durable MTP win at 1.38x–2.06x. The article gives teams a hardware-specific rule for when speculative decoding lowers latency and when it regresses.

Browser-in-the-Loop: Reinforcement Fine-Tuning LLM Agents for Web Form Filling STAD68 final paper, 2026

- QLoRA-tuned Qwen3-8B on 992 demonstrations, then ran online GRPO against live Chromium forms. Validation reward rose from 0.614 to 0.670 (+9.1%, p=0.007), and test reward rose 5.4%. The result reduces failed form workflows and manual recovery.

Concentrate or Collapse: When Reinforcement Learning Meets Diffusion Language Models for Web Planning

STAD80 paper, 2026

- Evaluated Qwen3-8B, ReFusion-8B, and FS-DFM-1.3B on 1,240 FormFactory tasks with browser-native Pass@1. FS-DFM scored 0/124 greedy despite 76.6–87.9% stochastic parser validity, while ReFusion MDPO reached 11/124 and 8.9%. The evaluation exposes policies that look competent but fail exact browser execution, reducing deployment risk.

ADDITIONAL BUSINESS IMPACT

- Automated seven SAP configuration use cases at 51 seconds average, replacing 15+ manual steps per request and cutting inference cost 75%.
- Built a five-agent cross-cloud A2A platform with Entra ID JWT propagation, AWS Lambda SigV4, Bicep, and Arize Phoenix. 124 unit tests, three orchestration modes, and 16 RBAC roles. The platform cut API calls 73%.

RESEARCH ARTIFACTS

Code: github.com/billy-enrizky/openbrowser-ai. Models: Qwen3-8B FormFactory GRPO and ReFusion-8B MDPO.

SYSTEMS LAB

- Built Baton, an 8,957-line Rust CLI for six agent harnesses, with native session import, per-turn git checkpoints, and parallel provider threads. 218 of 220 soak rounds passed, improving recovery safety and multi-agent throughput.
- Optimized H200 kernels for Muse Spark: 1.53x dense projection, 1.96x MoE W4A16, 1.90x attention decode, and 1.77x decoder. The gains reduce inference time for dense and MoE layers.

TEACHING

Teaching Assistant, University of Toronto 2023 – 2026

Taught calculus, data science, and probability for 1,029+ hours. Led tutorials for 30+ students, review seminars for 400+ students, and graded assessments for 650+ students. Impact: clearer exam preparation and targeted feedback for 650+ students.

EDUCATION

University of Toronto June 2026

Honours Bachelor of Science with High Distinction

Data Science and Machine Learning Specialist, Computer Science Major, Economics Minor. CGPA 3.87/4.00.

Relevant coursework: Algorithms, Data Structures, Software Design, Software Engineering, Computer Organization.

Awards: Dean's List (2023–2025), T-CAIREM Research Award, and Advanced Indonesian Scholarship.

Georgia Institute of Technology 2026 – Present

Master of Science in Computer Science, Artificial Intelligence Specialization