Simon Aytes

New York, NY | (313) 806-6429 | simon@aytes.net | /in/simonaytes | saytes.io

PROFESSIONAL EXPERIENCE

SENIOR DATA SCIENTIST January 2026 – Present

Penta Group, LLC — London, UK (Hybrid)

- Designed, built, and shipped a full-stack, API-first internal AI tooling platform using FastAPI and React/TypeScript, replacing a legacy Voila/Jupyter environment and integrating with existing platform infrastructure.
- Drove platform adoption to 1,500+ monthly analysis runs by 180+ users company-wide, with outputs feeding nearly all client deliverables and API access designed for developers and AI agents.
- Built and maintained production LLM-powered workflows for generation, summarization, document analysis, and analyst-facing automation, translating internal stakeholder needs into reliable tools used in live delivery.
- Re-architected always-on EC2/SageMaker notebook workflows into dynamically executed AWS ECS workers using PostgreSQL, CodePipeline, CodeBuild, and CloudWatch, projected to reduce related infrastructure costs by 50%+.

INDEPENDENT LLM AUTOMATION CONSULTANT January 2025 – December 2025

Freelance — Remote

- Scoped, designed, and deployed a production LLM document-intelligence system for a small-business client, owning delivery from discovery with non-technical stakeholders through production rollout.
- Built an invoice-processing system handling approximately 3,000 invoices per month across 100+ vendor formats, using LLM-driven OCR and OpenAI structured outputs with Pydantic schemas to convert documents into structured data.
- Delivered measurable workflow impact, saving approximately 2,000 working hours and $30,000+ in annual operating costs while enabling the client team to scale without additional headcount.

GRADUATE RESEARCH ASSISTANT February 2024 – December 2025

KAIST MLAI Lab — Seoul, South Korea

- First-authored Sketch-of-Thought: Efficient LLM Reasoning via Cognitive-Inspired Sketching, published at EMNLP 2025, introducing an inference-time method that reduced reasoning tokens by approximately 84% on average.
- Researched efficient LLM reasoning and test-time compute, including training-free inference-time methods and GRPO/SFT approaches for reducing reasoning-token usage.

DATA SCIENTIST September 2021 – February 2024

Penta Group, LLC — New York, NY

- Built NLP and social-listening systems for media intelligence, including automated topic extraction and large-scale text classification across social and traditional news data; delivered client-facing analyses for Fortune 500 stakeholders.
- Embedded with internal teams to identify workflow bottlenecks and automation opportunities; built reusable dashboards and reporting tools that reduced manual data-processing time by approximately 90%.

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST) December 2025

M.S., Artificial Intelligence, Advisor: Prof. Sung Ju Hwang

Lehman College, CUNY December 2022

B.S., Computer Science (Minor: Data Science), Summa Cum Laude

TECHNICAL SKILLS

ML & AI	Python, PyTorch, LLMs, NLP, RAG, scikit-learn, Pandas, NumPy, SQL
TOOLS	OpenAI SDK, FastAPI, PostgreSQL, TypeScript, React, Git
INFRASTRUCTURE	AWS (ECS, EC2, S3, SageMaker, CodePipeline, CodeBuild), CI/CD, Docker, Linux