Bruno Pistone

Milan, Italy brn.pistone@gmail.com +39 347 7147432 linkedin.com/in/bpistone medium.com/@brn.pistone

Summary

Proven leader in foundation model customization, distributed training, and agentic AI platform development. Track record of optimizing performance, reducing infrastructure costs, and driving enterprise-wide AI transformation. Recognized for leading cross-functional teams to deliver production-grade, high-impact AI solutions that drive automation, creativity, and measurable business outcomes. Thought leader and refined communicator, excel at bridging research and engineering to deliver production-ready, high-performance AI platforms, mentoring global engineering teams, and influencing product roadmaps.

- Deep Learning Architectures
- Machine Learning Engineering
- Cross-Functional Collaboration
- Product Roadmap Development
- Generative AI Techniques
- AI Ethics & Compliance
- System Design & Scalability
- Enterprise AI Transformation
- Programming & Tools
- Applied AI Research
- LLM System Design
- Data & Model Governance

Accomplishments

- Named among the Top 500 Italians Shaping the Future of Artificial Intelligence (2024) for advancing applied AI innovation and thought leadership.
- Transformed AI strategies for over 50 enterprise customers and startups by architecting and implementing scalable solutions for Generative AI, LLM fine-tuning, and Agentic RAG systems on AWS.
- Increased the adoption of AWS AI services by leading technical deep-dive sessions with Fortune 500 C-suite and engineering teams, resulting in the successful deployment of scaled transformer architectures.
- Delivered over 50+ custom Generative AI and LLM solutions for global strategic enterprise clients aimed at translating complex business requirements into production-ready architectures on AWS.
- Authored over 20+ high-impact technical publications on distributed systems, machine learning, and large-scale deep learning training to drive adoption of scalable AI architectures across global developer communities.
- Directed and expanded the AWS Advanced Computing Technical Field Community to 50+ members.

Career Experience

Amazon Web Services

Sr. WW GenAI Specialist SA – Amazon SageMaker AI Training, 2024 – Present

Architect and implement scalable Generative AI solutions on AWS with focus on driving enterprise AI transformation. Define Go-to-Market (GTM) strategy for Model customization, LLM fine-tuning on Amazon SageMaker AI Training services. Strengthen the global technical community by founding and leading a field group of 50+ AI specialists, fostering best practices in LLM customization and fine-tuning. Contribute to the product roadmap of Amazon SageMaker AI by collaborating with Product Managers, engineering teams on 10+ major feature launches and integrating customer feedback to enhance training and infrastructure capabilities.

- Established AWS as a thought leader in Generative AI by authoring and publishing 5+ technical blogs that became reference architectures for RAG, Agentic AI, and foundation model training.
- Pioneered distributed training architectures for large language models (LLMs) ranging from foundational to over 70B parameters which enabled customers to train complex models that were previously infeasible.
- Drove a 30% reduction in model training time for multiple enterprise clients by architecting optimized GPU clusters that sustained over 85% GPU utilization, directly lowering infrastructure costs.
- Engineered and implemented custom Fully Sharded Data Parallel (FSDP), Distributed Data Parallel (DDP), and other different training strategies that reduced memory consumption by 40%, enabling the training of larger foundation models on existing infrastructure.
- Engineered and open-sourced scalable training frameworks, adopted by a community of over 2 million developers, establishing new industry benchmarks for distributed training efficiency.

Sr. Generative AI/ML Specialist SA – Global Strategic Accounts, 2023 – 2024

Defined and drove the Go-to-Market (GTM) strategy for Generative AI and ML, establishing AWS as a thought leader in the Automotive, Manufacturing, and Financial Services industries. Authored and published five technical blogs and best practices on Generative AI, RAG, and Agentic AI to amplify AWS's technical voice and accelerate customer adoption. Mentored and upskilled a team of three solutions architects and engineers, building a center of excellence for Generative AI and advanced ML topics.

- Designed and architected high-performance RAG systems capable of processing over 1 million documents while maintaining sub-200ms query latency for enterprise-scale applications.
- Engineered and implemented optimized fine-tuning strategies with popular open-source frameworks, to optimize memory and GPU utilizations for model training an inference over 70%.

- Implemented scalable Generative AI solutions for Agentic workflows by integrating 5+ data sources and 20+ internal APIs to interact with internal systems.
- Leveraged a combination of Amazon SageMaker, Amazon Bedrock, and open-source models to build and deploy end-to-end RAG and Agentic AI solutions into production environments.

AI/ML Specialist Solutions Architect – EMEA, 2022 – 2023

Developed and deployed distributed training systems across multi-GPU clusters, which reduced model training timelines for large-scale neural networks from weeks to days and unlocked new avenues for advanced AI research. Implemented production-grade NLP solutions for enterprise clients, automating critical manual workflows and delivering measurable operational cost savings.

- Drove enterprise-wide AI adoption across the EMEA region through tailored solutions for strategic clients, resulting in generating multi-million-dollar cloud revenue.
- Launched a multi-tenant ML platform that centralized model governance and feature stores, standardizing the MLOps lifecycle and enhancing the operational efficiency of an organization of 100+ data scientists.
- Advised on the core roadmap and go-to-market strategy for AWS AI/ML services by translating enterprise customer requirements into actionable product features, ensuring service evolution aligned with market demands.

Machine Learning Engineer, 2020 – 2022

Delivered production-ready ML platforms for enterprise customers across Italy, establishing a foundation for automated, end-to-end model lifecycle management from data ingestion to monitoring. Spearheaded the creation of standardized MLOps templates and best practices that were adopted as a regional standard across the AWS Professional Services EMEA organization.

- Implemented comprehensive MLOps frameworks with integrated CI/CD pipelines and automated testing, reducing deployment risks and accelerating time-to-market for machine learning models.
- Developed and deployed custom NLP models for sentiment analysis and entity extraction, directly addressing specific client use cases and business intelligence requirements.
- Accelerated customer cloud adoption by mentoring technical teams on AWS AI/ML services and cloud-native architecture patterns, building long-term client self-sufficiency.

Machine Learning Reply, Milan - Italy

Senior Data, ML & Cloud Engineer, Machine Learning Engineer, 2019 – 2020

Developed and operationalized production ML models on GCP for predictive analytics and recommendation systems, directly enhancing client customer engagement and operational forecasting. Managed computer vision R&D initiatives and developed custom object detection models that automated retail inventory management and reduced manual inspection costs. Elevated team capabilities by mentoring junior engineers on cloud-native ML development and GCP service optimization.

- Drove end-to-end ML solution delivery for enterprise clients, translating business requirements into technical architectures and deployed models.
- Established MLOps practices using Kubernetes Engine to create a robust foundation for automated model deployment, scaling, and monitoring.
- Mentored 5+ junior engineers on cloud-native ML development practices and GCP service optimization

Eudata Srl, Milan - Italy

Sr. Technical Engineer, 2018 – 2019

Directed a cross-functional team of five engineers, implementing agile methodologies that improved project delivery predictability and speed. Led the technical design and full-stack development of the 'Convy AI' product, delivering advanced conversational AI capabilities through NLP and machine learning. Introduced real-time conversation analytics and sentiment analysis features that processed thousands of daily customer interactions.

- Spearheaded the implementation of voice and chat interfaces for the Eudata Omnichannel solution, enhancing customer engagement channels.
- Improved team output and code quality by establishing rigorous code review processes and technical documentation standards.

Celi Language Technology, Turin - Italy

Full stack Machine Learning engineer, 2016 – 2018

Advanced R&D initiatives for multilingual NLP, developing entity recognition and semantic analysis systems that operated across over 15 languages, including low-resource ones. Collaborated with computational linguists to build multi-language vocabularies and ontologies, expanding the market reach and applicability of NLP solutions.

- Architected and implemented the core search algorithms and linguistic processing capabilities for the Sophia Search engine generator, enabling sophisticated, language-aware information retrieval.
- Trained complex NLP models and integrated with Social Analytics products to provide clients with real-time sentiment analysis and trend detection from social media data.

Telnext / Deloitte Digital, Milan - Italy

Junior analyst and developer, 2015 – 2016

Supported digital transformation for Automotive industry clients by developing and delivering custom applications and CRM integrations.

- Contributed to the design and implementation of Salesforce solutions, streamlining sales and service workflows for automotive customer lifecycle management.

Education

MSc in Computer Science & Engineering | Politecnico di Milano

105/110, 10/2013 - 12/2015

Bachelor's Degree in Computer Science & Engineering | Politecnico di Milano

89/110, 09/2010 - 09/2013

Technical Proficiencies

- Deep Learning Frameworks: PyTorch, Hugging Face Transformers, DeepSpeed, Megatron-LM, FSDP, ZeRO, LoRA/QLoRA, PEFT
- Distributed Systems: torch.distributed, NCCL, Ray, MPI, SageMaker HyperPod, Kubernetes (EKS)
- Optimization: Mixed Precision (BF16, FP16, FP8), Gradient/Activation Checkpointing, learning rate schedulers
- Accelerators & Infra: NVIDIA GPUs, AWS Trainium/Inferentia, multi-node GPU clusters
- Frameworks & Orchestration: LangChain, LlamaIndex, CrewAI, semantic caching, workflow orchestration
- Search & Storage: FAISS, pgvector, Amazon OpenSearch
- Evaluation & Safety: RAGAS, LLM evaluation pipelines, prompt injection testing, monitoring & guardrails
- Other: MLOps, CI/CD for ML, model monitoring, production inference at scale
- Languages: Italian, English

Projects & Blogs

Training large language models on Amazon SageMaker: Best practices
aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/

Train large language models using Hugging Face and AWS Trainium
repost.aws/articles/ARHyFz-RpBR1OiGekJzkg2aw

Accelerate foundation model training and inference with Amazon SageMaker HyperPod and Amazon SageMaker Studio
https://aws.amazon.com/blogs/machine-learning/accelerate-foundation-model-training-and-inference-with-amazon-sagemaker-hyperpod-and-amazon-sagemaker-studio/

Customize DeepSeek-R1 distilled models using Amazon SageMaker HyperPod recipes – Part 1
https://aws.amazon.com/blogs/machine-learning/customize-deepseek-r1-distilled-models-using-amazon-sagemaker-hyperpod-recipes-part-1/

Customize Amazon Nova in Amazon SageMaker AI using Direct Preference Optimization
https://aws.amazon.com/blogs/machine-learning/customize-amazon-nova-in-amazon-sagemaker-ai-using-direct-preference-optimization/

Software-defined Vehicles, GenAI, IoT – The Path to AI-Defined Vehicles
https://aws.amazon.com/it/blogs/industries/software-defined-vehicles-genai-iot-the-path-to-ai-defined-vehicles/

Fine-tune Falcon 7B and other LLMs on Amazon SageMaker with @remote decorator
https://aws.amazon.com/blogs/machine-learning/fine-tune-falcon-7b-and-other-llms-on-amazon-sagemaker-with-remote-decorator/

SageMaker Python SDK: Interactive Distributed Training in Notebooks with PyTorch using SageMaker @remote function
github.com/aws/sagemaker-python-sdk

Ray Cluster setup on Amazon SageMaker training jobs
https://github.com/aws-samples/sample-ray-on-amazon-sagemaker-training-jobs

Build an internal SaaS service with cost and usage tracking for foundation models on Amazon Bedrock
aws.amazon.com/blogs/machine-learning/build-an-internal-saas-service-with-cost-and-usage-tracking-for-foundation-models-on-amazon-bedrock/

Interactive fine-tuning of Foundation Models with Amazon SageMaker Training using @remote decorator
github.com/aws-samples/amazon-sagemaker-llm-fine-tuning-remote-decorator

Multi-tenant Generative AI gateway with cost and usage tracking on AWS
github.com/aws-solutions-library-samples/guidance-for-a-multi-tenant-generative-ai-gateway-with-cost-and-usage-tracking-on-aws