GIANLUIGI VITALE

gianluigi.vitale12@gmail.com | Pistoia, Italia | gianluigivitale.github.io/portfolio
github.com/GianluigiVitale | LinkedIn | Google Scholar | OpenReview | ORCID: 0009-0007-6740-2695

INTERESSI DI RICERCA

Sistemi di machine learning, infrastrutture di serving per LLM, portabilità tra acceleratori, kernel per TPU e comportamento numerico, benchmarking ML, affidabilità dell'inferenza e valutazione della sicurezza degli LLM.

FORMAZIONE

Laurea triennale in Ingegneria Informatica — Universitas Mercatorum, Roma *ago. 2024–mar. 2027*

Media: 29,35/30 · voto di laurea atteso: 110/110 e lode — esami tutti sostenuti; tesi in corso

Tesi: Natural Language Is Flatland for Machine Intelligence: AI Systems Should Decouple Their Reasoning Substrate from Human Language

Relatore: Prof. Rocco Pietrini

Full Stack Web Development Bootcamp — *Boolean SRL (700 ore)* *gen.–giu. 2020*

Diploma di maturità scientifica — Liceo Scientifico S. Cannizzaro, Palermo *2019*

PUBBLICAZIONI

G. Vitale, "Decomposing One Professional-Framing Pipeline: Which Components Shift LLM Safety Boundaries?" **NeurIPS 2026**, main track (poster). Accettato, in corso di pubblicazione.
Articolo | Poster

G. Vitale, "DriftBench: Measuring and Predicting Infrastructure Drift in LLM Serving Systems," **MLSys 2026**, Bellevue, WA (USA), maggio 2026. Presentazione orale.
Articolo | Artefatto (badge ACM: Available, Functional, Reproduced) | Intervento

MANOSCRITTI IN REVISIONE

G. Vitale, R. Pietrini, "Cross-Generation Numerical Agreement of LLM Inference on TPUs," in revisione a **MLSys 2027**.

PREMI E RICONOSCIMENTI

Distinguished Artifact Evaluator Award — *Artifact Evaluation Committee di SOSP 2026* *2026*

Terzo premio — *International Mathematics Competition for University Students (IMC 2026)* *2026*

171° posto ex aequo su 448 (top 40%); partecipanti inviati per lo più dai propri atenei (circa 180, di 50 paesi).
Risultati | Sito IMC

ESPERIENZA DI RICERCA

- **DriftBench e validazione su TPU:** Realizzazione di un framework di misurazione e previsione basato su 236.985 coppie prompt-risposta in 105 configurazioni. Il Portability Risk Index generalizza a hardware mai visto (R^2 = 0,909). Estensione del risultato alle TPU v5e/v6e, dove la prevedibilità si è confermata anche con la nuova aritmetica (R^2 = 0,72–0,96).
- **Ablazione fattoriale del framing nei jailbreak (NeurIPS 2026):** Disegno fattoriale 2×2×2 definito a priori, applicato alla pipeline di professional framing STF: 5.000 prove etichettate manualmente su tre modelli frontier chiusi nello studio principale, 20.000 in totale su quattro fattoriali da 5.000 prove, e verifiche comportamentali su sei modelli open-weight. Il ruolo professionale (persona) è risultato il fattore principale dell'accettazione delle richieste (OR = 6,5), la Linguistic Substitution quello della profondità delle risposte (coefficiente cumulative-logit β = 0,91); la giustificazione morale non ha mostrato alcun effetto positivo rilevabile.
- **Caratterizzazione bit a bit della MXU delle TPU (in revisione a MLSys 2027; con il Prof. Rocco Pietrini):** Analizzando dall'esterno le TPU v4, v5e e v6e, abbiamo ricondotto le differenze tra generazioni nei prodotti matriciali bf16 all'ordine di riduzione tra le passate fissato dal compilatore. Un emulatore su CPU ha riprodotto bit per bit 34,6 milioni di nuovi output su tutte e tre le generazioni (bf16, precisione predefinita, quattro regimi di input standard). Far coincidere i soli prodotti matriciali non basta a far concordare il modello; fissando l'intero programma (whole-program pinning) logit e testo greedy diventano identici su tutte e tre le generazioni, e l'output vLLM identico su v5e/v6e.

- **DeepSeek V4-Flash su TPU:** Porting completato su 32 chip (v4-64), con validazione di tutte le modalità di attenzione e prefill e decode funzionanti. Un supporto equivalente è poi arrivato in vLLM TPU inference grazie a sviluppi paralleli nel resto dell'ecosistema.
- **Kernel DSA per TPU:** Un'implementazione identica bit per bit al riferimento fino a 8K token e in grado di arrivare a 256K. Lavoro in corso: è alla base del porting di GLM-5.3, ora su TPU v4.
- **Genomica long-context con l'ICAR-CNR:** Collaborazione con i genetisti dell'ICAR-CNR per verificare se i modelli del DNA a contesto lungo codifichino il linkage disequilibrium.

COMITATI E ATTIVITÀ DI REVISIONE

Membro del Program Committee — *MLSys 2027* *2026–2027*

Revisione degli articoli sottomessi al programma tecnico di MLSys 2027.

Revisore etico — *NeurIPS 2026* *2026*

Valutazione degli articoli sottomessi sotto il profilo etico: possibili danni per la società, rischi per l'integrità della ricerca e rispetto delle linee guida etiche della conferenza.

Membro dell'Artifact Evaluation Committee — *SOSP 2026* *2026*

Valutazione di artefatti di ricerca sul serving di modelli con uso efficiente della memoria e sull'analisi della sicurezza della memoria nei kernel GPU dei sistemi di inferenza LLM. Verifica di disponibilità, funzionamento e riproducibilità, oltre che della coerenza con i risultati sperimentali dichiarati negli articoli.

Membro dell'Artifact Evaluation Committee — *MLSys 2026* *2026*

Valutazione di quattro articoli accettati su ottimizzazione dei kernel di attenzione, infrastrutture di profiling per il ML, disaggregazione della memoria GPU e inferenza distribuita. Riproduzione dei risultati sperimentali su cluster 8×H100 a noleggio (RunPod) e su infrastruttura AMD MI350X (AMD University Program), con verifica di disponibilità, funzionamento e coerenza con quanto dichiarato in ciascun articolo.

ESPERIENZA PROFESSIONALE

Systems Engineer — *Archivio di Stato di Pistoia, Ministero della Cultura* *apr. 2024–oggi*

Progettazione e messa in produzione di un sistema di retrieval-augmented generation (RAG) su oltre 10.000 documenti amministrativi interni, che risponde alle domande del personale citando le fonti. Realizzazione di un'interfaccia di consultazione pubblica basata sull'AI, che permette a personale e cittadini di interrogare in linguaggio naturale l'intero database dell'inventario.

Preparazione al concorso e studio autonomo *apr. 2021–mar. 2024*

Preparazione al concorso per il Nucleo Speciale Tutela Privacy e Frodi Tecnologiche della Guardia di Finanza: completata la preparazione militare richiesta (forza, agilità, resistenza) e superate tutte le prove tecniche e fisiche. In parallelo, studio autonomo dell'informatica; poi vinto il concorso pubblico per Systems Engineer (apr. 2024).

Web Developer — *Informamuse S.R.L.* *ott. 2020–apr. 2021*

Sviluppo di siti e applicazioni web con HTML, CSS, JavaScript, PHP, Laravel, Vue.js e MySQL. Gestione completa dei progetti, dalla progettazione alla messa in produzione.

COMPETENZE TECNICHE

Linguaggi: Python, JavaScript, PHP, SQL, HTML/CSS, Bash

Framework: PyTorch, JAX, Pallas, Hugging Face Transformers

ML/Sistemi: vLLM, tpu-inference, SGLang, TensorRT-LLM, CUDA, Docker

Infrastruttura: Google Cloud TPU (v4, v5e, v6e), amministrazione di sistemi Linux, reti, provisioning di GPU in cloud (RunPod, AMD University Program, Chameleon Cloud), CI/CD

Strumenti di ricerca: LaTeX, Git, Jupyter, scipy, statsmodels

RISORSE DI CALCOLO

- Google Cloud TPU Research Cloud (TRC) — attivo da maggio 2026. Allocazione attuale di 320 chip: 32 chip v4 on-demand e 288 chip spot tra v4, v5e e v6e, in zone USA e UE.
- Account AMD University Program (attivo da marzo 2026); allocazione su Chameleon Cloud (2.000 SU); esperienza su cluster GPU RunPod (H100, H200, B200, MI300X)

LINGUE

Italiano (madrelingua), **Inglese** (C1 — livello accademico, pubblicazioni scientifiche)