Caio Sousa Santos

Rio de Janeiro, RJ caiosousasantos.cs@gmail.com rendercv.com caio-santos-85647914
Rhombk

Experience

Co-Founder & CTO, Nexus AI – San Francisco, CA — June 2023 – present

- Built foundation model infrastructure serving 2M+ monthly API requests with 99.97% uptime
- Raised $18M Series A led by Sequoia Capital, with participation from a16z and Founders Fund
- Scaled engineering team from 3 to 28 across ML research, platform, and applied AI divisions
- Developed proprietary inference optimization reducing latency by 73% compared to baseline

Research Intern, NVIDIA Research – Santa Clara, CA — May 2022 – Aug 2022

- Designed sparse attention mechanism reducing transformer memory footprint by 4.2x
- Co-authored paper accepted at NeurIPS 2022 (spotlight presentation, top 5% of submissions)

Projects

Penguin — Jan 2023 – present

Open-source library for high-performance LLM inference kernels

- Achieved 2.8x speedup over baseline attention implementations on A100 GPUs
- Adopted by 3 major AI labs, 8,500+ GitHub stars, 200+ contributors

Ziggurat — Jan 2021

Automated neural network pruning toolkit with differentiable masks

- Reduced model size by 90% with less than 1% accuracy degradation on ImageNet
- Featured in PyTorch ecosystem tools, 4,200+ GitHub stars

Skills

Languages: Python, C++, CUDA, Rust,

Infrastructure: Kubernetes, Ray, distributed training, AWS, GCP

Linguas Estrangeiras: Inglês fluente.

Any Section Title

You can use any section title you want.

You can choose any entry type for the section: `TextEntry`, `ExperienceEntry`, `EducationEntry`, `PublicationEntry`, `BulletEntry`, `NumberedEntry`, or `ReversedNumberedEntry`.

Markdown syntax is supported everywhere.

The `design` field in YAML gives you control over almost any aspect of your CV design.

See the documentation for more details.

Education

Princeton University, PhD in Computer Science – Princeton, NJ Sept 2018 – May 2023

- Thesis: Efficient Neural Architecture Search for Resource-Constrained Deployment
- Advisor: Prof. Sanjeev Arora
- NSF Graduate Research Fellowship, Siebel Scholar (Class of 2022)