

Willow Hughes

📍 San Francisco, CA ✉ john.doe@email.com 🌐 rendercv.com 🔗 rendercv 🐙 rendercv

Welcome to RenderCV

RenderCV reads a CV written in a YAML file, and generates a PDF with professional typography.
See the [documentation](#) for more details.

Education

- | | | |
|------------|--|-----------------------|
| PhD | Princeton University , Computer Science | Princeton, NJ |
| | <ul style="list-style-type: none">Thesis: Efficient Neural Architecture Search for Resource-Constrained DeploymentAdvisor: Prof. Sanjeev AroraNSF Graduate Research Fellowship, Siebel Scholar (Class of 2022) | Sept 2018 – May 2023 |
| BS | Boğaziçi University , Computer Engineering | Istanbul, Türkiye |
| | <ul style="list-style-type: none">GPA: 3.97/4.00, ValedictorianFulbright Scholarship recipient for graduate studies | Sept 2014 – June 2018 |

Experience

- | | |
|--|---|
| Nexus AI , Co-Founder & CTO | San Francisco, CA |
| <ul style="list-style-type: none">Built foundation model infrastructure serving 2M+ monthly API requests with 99.97% uptimeRaised \$18M Series A led by Sequoia Capital, with participation from a16z and Founders FundScaled engineering team from 3 to 28 across ML research, platform, and applied AI divisionsDeveloped proprietary inference optimization reducing latency by 73% compared to baseline | June 2023 – present 2 years 8 months |
| NVIDIA Research , Research Intern | Santa Clara, CA |
| <ul style="list-style-type: none">Designed sparse attention mechanism reducing transformer memory footprint by 4.2xCo-authored paper accepted at NeurIPS 2022 (spotlight presentation, top 5% of submissions) | May 2022 – Aug 2022 4 months |
| Google DeepMind , Research Intern | London, UK |
| <ul style="list-style-type: none">Developed reinforcement learning algorithms for multi-agent coordinationPublished research at top-tier venues with significant academic impact<ul style="list-style-type: none">ICML 2022 main conference paper, cited 340+ times within two yearsNeurIPS 2022 workshop paper on emergent communication protocolsInvited journal extension in JMLR (2023) | May 2021 – Aug 2021 4 months |
| Apple ML Research , Research Intern | Cupertino, CA |
| <ul style="list-style-type: none">Created on-device neural network compression pipeline deployed across 50M+ devicesFiled 2 patents on efficient model quantization techniques for edge inference | May 2020 – Aug 2020 4 months |
| Microsoft Research , Research Intern | Redmond, WA |
| <ul style="list-style-type: none">Implemented novel self-supervised learning framework for low-resource language modelingResearch integrated into Azure Cognitive Services, reducing training data requirements by 60% | May 2019 – Aug 2019 4 months |

Projects

| | |
|---|--------------------|
| FlashInfer Open-source library for high-performance LLM inference kernels <ul style="list-style-type: none">Achieved 2.8x speedup over baseline attention implementations on A100 GPUsAdopted by 3 major AI labs, 8,500+ GitHub stars, 200+ contributors | Jan 2023 – present |
| NeuralPrune Automated neural network pruning toolkit with differentiable masks <ul style="list-style-type: none">Reduced model size by 90% with less than 1% accuracy degradation on ImageNetFeatured in PyTorch ecosystem tools, 4,200+ GitHub stars | Jan 2021 |

Publications

| | |
|---|-----------|
| Sparse Mixture-of-Experts at Scale: Efficient Routing for Trillion-Parameter Models <i>John Doe, Sarah Williams, David Park</i> 10.1234/neurips.2023.1234 (NeurIPS 2023) | July 2023 |
| Neural Architecture Search via Differentiable Pruning <i>James Liu, John Doe</i> 10.1234/neurips.2022.5678 (NeurIPS 2022, Spotlight) | Dec 2022 |
| Multi-Agent Reinforcement Learning with Emergent Communication <i>Maria Garcia, John Doe, Tom Anderson</i> 10.1234/icml.2022.9012 (ICML 2022) | July 2022 |
| On-Device Model Compression via Learned Quantization <i>John Doe, Kevin Wu</i> 10.1234/iclr.2021.3456 (ICLR 2021, Best Paper Award) | May 2021 |

Selected Honors

- MIT Technology Review 35 Under 35 Innovators (2024)
- Forbes 30 Under 30 in Enterprise Technology (2024)
- ACM Doctoral Dissertation Award Honorable Mention (2023)
- Google PhD Fellowship in Machine Learning (2020 – 2023)
- Fulbright Scholarship for Graduate Studies (2018)

Skills

Languages: Python, C++, CUDA, Rust, Julia
ML Frameworks: PyTorch, JAX, TensorFlow, Triton, ONNX
Infrastructure: Kubernetes, Ray, distributed training, AWS, GCP
Research Areas: Neural architecture search, model compression, efficient inference, multi-agent RL

Patents

- Adaptive Quantization for Neural Network Inference on Edge Devices (US Patent 11,234,567)
- Dynamic Sparsity Patterns for Efficient Transformer Attention (US Patent 11,345,678)
- Hardware-Aware Neural Architecture Search Method (US Patent 11,456,789)

Invited Talks

- Scaling Laws for Efficient Inference — Stanford HAI Symposium (2024)
- Building AI Infrastructure for the Next Decade — TechCrunch Disrupt (2024)
- From Research to Production: Lessons in ML Systems — NeurIPS Workshop (2023)
- Efficient Deep Learning: A Practitioner’s Perspective — Google Tech Talk (2022)

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.