Yunfeng Zhang

Chappaqua, NY · zywind@gmail.com · linkedin.com/in/zywind · Google Scholar

Summary

- Staff ML Engineer with 10+ years building production AI/ML systems at scale, specializing in Responsible AI, MLOps, and LLM-powered products.
- Deep expertise in MLOps, AI fairness/safety, content moderation, and end-to-end ML infrastructure across GCP, AWS, Databricks, and on-prem environments.
- Published 50+ papers with 7,600+ citations; 12 granted patents.

Experience

Superhuman (formerly Grammarly) — NYC

Machine Learning Engineer, Core Product — Aug 2025 – Present

- Improved search quality for Superhuman Go by implementing reranking and optimizing agentic search invocation, delivering measurably better results for users.
- Led ML development for Grammarly's agent platform, building the system that proactively recommends agents to users based on context — from zero to production, including the recommendation model, serving infrastructure, and evaluation framework.
- Diagnosed and fixed critical performance bottlenecks in a production AI agent platform, roughly tripling its throughput and making the evaluation pipeline 8× faster.

Tech Lead, Responsible AI Team — Feb 2024 – Jul 2025

- Led the development and strategy for Superhuman's Responsible AI initiatives, including content moderation services and AI safety/bias evaluation tools.
- Spearheaded an 8× improvement in the precision of the content moderation service, resulting in a 70% reduction in user complaints.
- Built an automated AI risk assessment platform that replaced a manual, multi-day process with single-command execution, expanding coverage to 6 risk dimensions and enabling evaluation of numerous LLM and product features — reducing assessment timelines from weeks to days.
- Co-developed a state-of-the-art multilingual NER model for PII detection and redaction, directly enabling Superhuman's compliance with global privacy and data safety requirements.

EvolutionIQ — NYC

Senior Machine Learning Engineer — Feb 2023 – Feb 2024

- Led ML development for Individual Disability Claim Guidance, driving significant advancements in predictive accuracy and operational efficiency.
- Pioneered the company's first deep dive into model biases and successfully established guidelines and metrics for Responsible AI.
- Successfully introduced two advanced survival modeling frameworks, empowering all product teams with the capability to predict the duration of disability claims accurately.
- Completely revamped the ML training and tuning pipeline, achieving a 2–10× increase in training speed and a marked enhancement in model robustness and reliability.

Twitter Inc. — NYC

Senior Machine Learning Engineer — Jun 2021 – Nov 2022

- Architected and launched Fairness Evaluator, Twitter's first distributed model performance and fairness evaluation system (built on TFX and Apache Beam), adopted company-wide to evaluate and improve recommendation, ranking, and content moderation models.
- Researched and developed new AI fairness and performance metrics, including those published at ACM FAccT 2022.

IBM, T.J. Watson Research Center — Yorktown Heights, NY

Research Staff Member — Jun 2015 – Jun 2021

- Co-created IBM's **AI Fairness 360** and **AI Explainability 360** toolkits — the industry's first open-source ethical AI toolkits — and integrated them into IBM OpenScale. Received **IBM Outstanding Research Accomplishment Award**.
- Prototyped and transferred research into production for Watson OpenScale, IBM AutoAI, and IBM Watson Assistant, covering active learning, drift detection, and conversational AI. Received **IBM Outstanding Accomplishment Award**.

Education

University of Oregon — Eugene, OR

Ph.D. in Computer and Information Science — Jun 2015

Beijing Normal University — Beijing, China

B.S. in Computer and Information Science — Jun 2007

Publications & Patents

Published 50+ papers with 7,600+ citations; 3 granted patents and 10 pending applications. See Google Scholar for the full list. Selected publications:

- De-biasing "bias" measurement
- AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias
- One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques

Technical Skills

- **ML & AI:** LLMs, RAG, PyTorch, TensorFlow, scikit-learn, XGBoost, NLP, AI fairness & safety, content moderation
- **MLOps:** Kubeflow, TFX, Dagster, Airflow, Apache Beam, Apache Spark
- **Cloud & Infrastructure:** GCP, AWS, Databricks, SageMaker, Vertex AI, Kubernetes, Terraform
- **Languages:** Python, SQL, Java, Javascript/Typescript, R