

Joel Pfeiffer (publishes as Joseph J. Pfeiffer III)

Principal Machine Learning Scientist — Large-Scale ML Systems: Training, Inference & Applied LLMs

Seattle, WA · jpfeiffe@gmail.com · linkedin.com/in/joel-pfeiffer-57801737 · github.com/jpfeiffe · jpfeiffe.github.io

Principal ML Scientist, 11 years shipping production ML at Microsoft Advertising: foundation-model personalization, GPU-efficient LLM serving and training, multi-agent harnesses, auto-bidding models (response curves, forecasting, decisions), privacy-preserving modeling. Ph.D. in relational ML; publications at WWW, WSDM, ICDM, AAAI; web-standards proposals behind Microsoft Edge's Ad Selection API.

EXPERIENCE

Microsoft Seattle / Bellevue, WA

Jun 2015 – Present

Principal ML Scientist (Mar 2022 – present) · **Senior ML Scientist** (Sep 2017 – Mar 2022) · **Applied Researcher** (Jun 2015 – Aug 2017)

Set technical direction for cross-team initiatives; principal design reviewer for large-scale training and serving systems; senior ML hiring and mentoring.

User Foundation Models & Personalization · 2023 – 2025

- Built the user-intent pipeline for Microsoft Audience Network personalization: GPT-4 generates intent queries from each user's behavior sequence in offline batch; the pipeline retrieves matching search ads and indexes them for online display serving; first launch drove a **double-digit revenue lift**, deployed globally, with a distilled v2 for serving cost.
- Built the offline experimentation and flighting framework for rapid model/prompt iteration; modernized offline modeling infrastructure and ran production flight analysis.

Auto-Bidding — Response-Curve, Calibration, Forecasting & Decision Models · 2025 – present

- Own the campaign dynamics models in Microsoft's auto-bidding system (millions of campaigns at P99 ~200 ms): monotonic campaign response-curve models and dual-to-KPI calibration — allocating spend across a **multi-billion-dollar ads marketplace**.
- Built the online calibration layer (hourly adaptive updates; a dozen prototyped variants, landed in production and piloted live) and integrated Chronos-2 time-series forecasting for response-curve extrapolation and traffic forecasts.
- *In progress*: typed-decision models for the bidding loop — choosing search/native allocation and dual adjustments per campaign.

GPU Efficiency — Serving & Training · 2026

- Unblocked Microsoft Commerce model serving: took the small-LM scorer to **~4× baseline throughput on one H100** — CPU-side feeding fixes, Triton dynamic batching, CUDA graphs, 4-bit inference; shipped with weights-versioned CI/CD and canaries.
- Turned that benchmark-change-remeasure loop into a multi-agent harness (Git-orchestrated, hook-enforced lanes, evidence ledger) that found and landed those optimizations; a sibling ranking team then applied it to their legacy models.
- *In progress*: agent-assisted TorchRec port of the production ranking trainer (multi-billion-key sparse embeddings; 8× GPU-reduction target at AUC parity).

Generative Ads (Brandshowroom) · 2025

- Built a multimodal generative ad-assembly system (FAISS retrieval, end-to-end MinGRU sequence model, RLAIIF loop); architecture adopted by later GenAI ads work.

Privacy-Preserving Modeling (Parakeet v2, Masked LARK, Ad Selection API) · 2020 – 2022

- Architected the early prototypes and designs for Microsoft's privacy-preserving ads stack — noisy client-side ranking, masked/secure-aggregation learning — and contributed to Microsoft Edge's TEE-based Ad Selection API.

Signals Modeling for Syndication Partners · 2017 – 2022

- Led Signals modeling for Microsoft Advertising's syndication partners for five years; consolidated the per-partner models into one — **up to +10% AUC** for individual partners, with large compute and serving-cost reductions.

Earlier · 2009 – 2017

- Click prediction, model compression, personalization, and bandits at Microsoft (2015 – 2017); relational ML with Jennifer Neville at Purdue (2009 – 2015); internships at Microsoft Research, LLNL, and LivingSocial.

EDUCATION

Ph.D., Computer Science — Purdue University · Advisor: Jennifer Neville

2009 – 2015

M.S. Statistics, Purdue (2014) · M.S. CS, CU Boulder (2009) · B.S. CS & Math, New Mexico State (2006)

SELECTED PUBLICATIONS, STANDARDS & SERVICE

as Joseph J. Pfeiffer III · 16 papers (Google Scholar) · WICG: MaskedLARK, PARAKEET V2 · RMLlib · Reviewer: KDD, ICML, NeurIPS, AAAI

- Masked LARK: Masked Learning, Aggregation and Reporting workflow. arXiv, 2021.
- Causal Transfer Random Forest: Combining Logged Data and Randomized Experiments for Robust Prediction. WSDM, 2021.
- Overcoming Relational Learning Biases to Accurately Predict Preferences in Large Scale Networks. WWW, 2015.
- Incorporating Assortativity and Degree Dependence into Scalable Network Models. AAAI, 2015.
- Composite Likelihood Data Augmentation for Within-Network Statistical Relational Learning. ICDM, 2014.
- Attributed Graph Models: Modeling Network Structure with Correlated Attributes. WWW, 2014.