Sumit Pokharel

Tokyo, Japan · pokharel.sumit@proton.me · idosumit.com · linkedin.com/in/sumit-pokharel

Summary

Software Engineer at Rakuten, leading frontend projects for Japan's largest e-commerce checkout system (50M+ users). Independent ML researcher focused on efficient training and model architectures.

ML/Research Projects

moe-emergence

`https://git.idosumit.com/moe-emergence`

Investigating emergent expert specialization in Mixture of Experts architectures.

- Training a compact MoE variant of GPT-2 on code, math, and natural language to visualize which experts become domain specialists.
- Custom router with NoisyTop-k selection and straight-through estimator.
- Load balancing loss (Switch Transformer) and z-loss for router stability.
- Surgical modification of pretrained GPT-2 (replacing MLP layers 8–11 with MoE layers).

seq2seq

`https://git.idosumit.com/seq2seq`

Implementation of Sequence to Sequence Learning with Neural Networks by Sutskever et al. (2014).

- Encoder-decoder LSTMs for German-to-English translation.
- Trained on 2M WMT19 sentence pairs with teacher forcing and gradient clipping.

transformer implementations

`https://git.idosumit.com/transformers`

First-principles implementations of transformer architectures:

- Original Transformer (2017): encoder-decoder for EN-JP translation.
- GPT-2: custom LayerNorm, causal masking, weight tying.
- LLaMA2: RoPE, RMSNorm, grouped-query attention, KV cache, SwiGLU.
- Mistral-7B: sliding window attention, rolling buffer cache.

chibigrad

`https://git.idosumit.com/chibigrad`

Miniature autograd engine in pure NumPy. Automatic differentiation and computational graphs.

Work Experience

Rakuten Group, Inc. — Tokyo

Software Engineer — Apr 2024 – Present

- Led frontend projects for three major initiatives to production: Okihai (contactless delivery), Credit Card Installments, Credit Card Scan (in development).
- Built internal CLI tool using Claude API for auto-generating component boilerplate, improving team productivity.

- Created component dependency graph powered by Claude API that indexes 200+ components for intelligent navigation and analysis.

Best Path Research
Machine Learning Intern Jul 2023 – Aug 2023

- Built Python pipeline for digital text → handwritten image conversion.
- Developed receipt distortion correction system using SAM for segmentation, DocTr-style control point regression for dewarping, and 10K+ synthetically warped training images.

Education

Ritsumeikan Asia Pacific University
Bachelor's Degree in Business Administration | CGPA: 3.65 Sep 2019 – Sep 2023

Technical Skills

Languages:	Python, TypeScript, JavaScript
ML/AI:	PyTorch, NumPy, HuggingFace, transformer architectures, Seq2Seq, Mixture of Experts
Web/Frontend:	React, Next.js, Astro
Tools:	Git, Vim, Tmux

Language Proficiency

Japanese:	JLPT N2 - Professional working proficiency
English:	Fluent