

Ryan Kim

224-517-7069 | ryankim17920@gmail.com | linkedin.com/in/ryankim17920 | github.com/RyanKim17920 | ryankim.me

EDUCATION

University of Illinois Urbana-Champaign

B.S. Computer Science (Junior Standing) | GPA: 4.0/4.0 | Chancellor's Scholar | AIME Qualifier (4x)

Urbana, IL

Expected May 2029

EXPERIENCE

RL Intern (Founding Team)

Aug 2026–Present

Delightful AI (a16z speedrun-backed)

Remote

- Built bioinformatics **agent RL environments** end to end (Chimera): task design, Docker/Harbor execution, reward grading.
- Designed a synthetic generator of verified-solvable tasks: GPT-6 Astra, Claude Opus 5.5, Fable 5.1 score <**0.5/1.0 reward**.
- Analyzed agent trajectories with an **LLM-as-a-judge** that checks each run against expected milestones.

Undergraduate Researcher

Aug 2026–Present

BLENDER Lab, UIUC (with Alexi Gladstone)

Urbana, IL

- Led research extending **Energy-Based Transformers to multi-token prediction** through Explorative Modeling (XM).
- Implemented the full training stack and compared candidate energy formulations against block-diffusion LLMs.

ML Research Contributor

Apr 2026–Present

MedARC

Remote

- Built an LLM autoresearch agent whose dual-branch search with anti-overfitting cycles drove **37% of the gain over baseline** on the NanoPath v1 benchmark (**3rd-author paper**, accepted at a NeurIPS 2026 workshop).
- Designed agent infrastructure (knowledge graphs, SLURM, worktree isolation) that cut config errors 80% and token use 15x.
- Introduced SPECTRA, label-free contrastive LoRA fine-tuning that raised mean PathoROB robustness **0.726→0.891** and improved downstream AUC across 7 pathology foundation models (**first-author paper**, accepted at a NeurIPS 2026 workshop).

Pathology AI Research Intern

Oct 2024–Jul 2026

AI in Medicine & Computational Biology Lab, Weill Cornell Medicine

New York, NY

- Benchmarked 12 vision foundation models for tumor-purity regression on TCGA and internal cohorts with a PyTorch evaluation pipeline; **error doubled on out-of-distribution data** (MAE ~0.10→~0.20).

Computational Oncology Research Intern

Dec 2024–Jan 2026

Advanced Computing & Oncology Lab, Memorial Sloan Kettering

New York, NY

- Engineered a **CPU-only LLM pipeline** that classified 370 clinically annotated lung-cancer CT reports, optimizing prompts for Llama 3.1-8B and Qwen3-4B with DSPy MIPROv2 (**3rd-author abstract**, accepted at ACRO 2026).
- Linked percent thymic tissue to NSCLC outcomes via **Cox survival analysis** of 230 patients' clinical and demographic records.

Simons Summer Research Fellow (Computational Genomics)

Jun 2025–Aug 2026

Yurovsky Lab, Stony Brook University (Simons Foundation)

Stony Brook, NY

- Predicted survival of 185 glioblastoma patients with a **DNABERT-based Cox model** on non-coding DNA mutations.
- Identified significant **age-related survival signals** and racial disparities in mutational scores via a Polars pipeline.

PROJECTS

Which State Should Prediction Read? | Sole-author paper, under review at ICLR 2027

Aug 2026–Sep 2026

- Trained **State–Prediction Separation** (SPS) language models with separate-weight streams; head ablations showed they specialize, with prediction's induction-style retrieval relying on state's previous-token heads.
- Used these findings to design Sequential, a final-state-read architecture that reaches a FLOP-matched Transformer's loss on **2.1x fewer tokens** and ends **0.064 nats/token lower** after 20B-token training (92M non-embedding params).

Qwen3.6-27B ExpThink | LLM Post-Training, GRPO

Jul 2026–Aug 2026

- Post-trained Qwen3.6-27B with a length-adjusted GRPO reward, cutting thinking tokens **52% in-domain** / **34% OOD** and truncation 6.0%→2.6% while mean OOD accuracy rose 84%→86%.
- Trained with PyTorch FSDP on 8xH100 in verl, with rollouts from tensor-parallel SGLang.

speedLM | OpenAI-Compatible LLM Serving, Speculative Decoding

Jul 2026–Aug 2026

- Built a custom vLLM server that **fine-tunes EAGLE-3 draft models** on captured local traffic during idle GPU time.
- Raised tokens accepted per verification step **13.4%** over the stock draft on held-out Qwen3-8B data.

TECHNICAL SKILLS

LLM Stack: PyTorch, FSDP, verl, SGLang, vLLM, Hugging Face, DSPy, EAGLE-3.

Machine Learning: Large Language Models (LLMs), Deep Learning, Transformers, Reinforcement Learning, Distributed Training.

Methods: LLM Architecture & Interpretability, Post-Training (GRPO), Speculative Decoding, RL Environments, LLM Agents.

Languages: Python, C/C++, Java, SQL (PostgreSQL), JavaScript (React, Node.js).

Tools: SLURM, Docker, Harbor, Linux, Git, FastAPI, Scikit-learn, Polars, Pandas, NumPy.