Ryan Kim

ryankim17920@gmail.com · ryankim.me
github.com/RyanKim17920 · linkedin.com/in/ryankim17920

Research Interests

LLM architecture and mechanistic interpretability; LLM post-training with reinforcement learning and agent RL environments; efficient inference; energy-based language models; robust foundation models for computational pathology.

Education

B.S. in Computer Science Expected May 2029
University of Illinois Urbana-Champaign Urbana, IL
Junior standing. GPA: 4.0/4.0. Chancellor's Scholar.

Publications

Under Review

[1] **Ryan Kim**. *Which State Should Prediction Read? A Mechanistic Analysis of State–Prediction Separation.* Under review. [Code] [Models]

Workshop Papers

[2] **Ryan Kim**, Hassan Muhammad, Anish Dulal, Paul Steven Scotti, Tanishq Mathew Abraham. *Contrastive fine-tuning against stain and scanner bias improves robustness and downstream performance across pathology foundation models.* NeurIPS 2026 Workshop on AI at Scale for Clinical Impact (ASCI). [OpenReview] [Code] [Models]

[3] Paul Steven Scotti, Hassan Muhammad, **Ryan Kim**, Nima Ashjaee, Ratna Sagari Grandhi, Saman Rahbar, Raymond Biju, Anish Dulal, Chao Feng, Connor Lane, Benjamin Warner, Tanishq Mathew Abraham. *NanoPath: Fast, Fair, Open Experimentation for Pathology Foundation Models.* NeurIPS 2026 Workshop on AI at Scale for Clinical Impact (ASCI). [OpenReview] [Code]

Conference Abstracts

[4] Kwadwo Amoako-Boadu, Alexander Kim, **Ryan Kim**, David Miller, Rami Naaman, Abraham J. Book, Robert Lee, Narek Shaverdian, Daniel Gomez, Tafadzwa L. Chaunzwa. *From Text to Treatment: Automated Identification of Airway Events and Referral Patterns in Lung Cancer Imaging Using Locally Deployed CPU-Only LLMs.* ACRO 2026 Annual Meeting.

Research Experience

Undergraduate Researcher Aug 2026 – Present
BLENDER Lab, University of Illinois Urbana-Champaign Urbana, IL
With Alexi Gladstone.

- 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 language models.

Independent Researcher Aug 2026 – Sep 2026
Sole author of paper [1]

- Trained State–Prediction Separation (SPS) language models, which gave the state and prediction streams separate weights; head ablations showed that the two streams specialize, and that prediction's induction-style retrieval relied on previous-token heads in the state stream.

- Used these findings to design Sequential, an architecture in which prediction read the final state; it reached the loss of a FLOP-matched Transformer on 2.1× fewer training tokens (92M non-embedding parameters, 20B-token runs).

ML Research Contributor Apr 2026 – Present
MedARC Remote

- Developed SPECTRA, a label-free contrastive fine-tuning method (LoRA) that made pathology foundation models less sensitive to stain and scanner differences; it raised the mean PathoROB robustness index from 0.726 to 0.891 across 7 models (first author of [2]).
- Showed that the robustness gains carried over to downstream tasks: SPECTRA improved CPTAC AUC (38 tasks) and HEST correlation on all 7 models, with 3 seeds per backbone.
- Built an LLM-based autoresearch agent whose dual-branch search, with anti-overfitting cycles, accounted for 37% of the gain over baseline on the NanoPath v1 benchmark (third author of [3]).
- Created the agent's research scaffolding (knowledge graphs, git-worktree isolation, preemption-robust SLURM jobs, experiment tracking), cutting configuration errors by 80%.

Simons Summer Research Fellow (Computational Genomics) Jun 2025 – Aug 2026
Yurovsky Lab, Stony Brook University Stony Brook, NY
Funded by the Simons Foundation.

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

Computational Oncology Research Intern Dec 2024 – Jan 2026
Advanced Computing & Oncology Lab, Memorial Sloan Kettering New York, NY

- Engineered a locally deployed, CPU-only LLM pipeline that classified 370 clinically annotated lung-cancer CT reports, using DSPy MIPROv2 to optimize prompts for Llama 3.1-8B and Qwen3-4B (abstract [4]).
- Linked percent thymic tissue to non-small cell lung cancer (NSCLC) outcomes through Cox survival analysis of 230 patients' clinical and demographic records.

Pathology AI Research Intern Oct 2024 – Jul 2026
AI in Medicine & Computational Biology Lab, Weill Cornell Medicine New York, NY

- Benchmarked 12 pathology foundation models on tumor-purity regression across TCGA and internal cohorts with a PyTorch harness for measuring distribution shift.
- Exposed a clinical deployment risk: prediction error doubled on out-of-distribution data.

Industry Experience

RL Intern (Founding Team) Aug 2026 – Present
Delightful AI (a16z speedrun-backed) Remote

- Owned agent RL environments for bioinformatics (Chimera) end to end, covering task design, Docker/Harbor execution, and reward grading.
- Designed a synthetic generator of verified-solvable tasks with a controlled task distribution and injected errors, filtered by rejection sampling; GPT-6 Astra, Claude Opus 5.5 and Fable 5.1 all scored below 0.5 on the 0–1 reward scale.
- Analyzed agent trajectories with an LLM-as-a-judge using milestone matching.

Lead Technical Developer Nov 2025 – Apr 2026
MathLinks.org Remote

- Directed engineering for a full-stack competition platform that served 150,000+ users across 11 partner organizations.
- Shipped React/Node.js authentication, submission, and leaderboard workflows for high-concurrency contests.
- Secured $20,000 in funding with the team (Hudson River Trading; Innovate901 1st Place).

Projects

Qwen3.6-27B ExpThink Jul 2026 – Aug 2026
LLM post-training with GRPO

- Post-trained Qwen3.6-27B with a length-adjusted GRPO reward that cut thinking tokens by 52% in-domain and 34% out-of-distribution (OOD); mean OOD accuracy rose from 84% to 86%.
- Ran training with PyTorch FSDP on 8×H100 in verl, with rollouts from tensor-parallel SGLang.

speedLM Jul 2026 – Aug 2026
OpenAI-compatible LLM serving with speculative decoding

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

PapersToCode Mar 2024 – Mar 2026
Data pipeline, full-stack

- Matched 500,000+ ML/AI papers to GitHub repositories with an automated data pipeline, surfacing research that lacked public code.

Honors & Awards

Simons Summer Research Fellowship, Simons Foundation 2025
Chancellor's Scholar, University of Illinois Urbana-Champaign
AIME Qualifier (4×)

Technical Skills

LLM Stack	PyTorch, FSDP, verl, SGLang, vLLM, Hugging Face, DSPy, EAGLE-3
Methods	LLM architecture & interpretability, post-training (GRPO), RL environments, LLM agents, speculative decoding, contrastive learning, pathology foundation models, OOD robustness, Cox survival analysis, clinical NLP
Languages	Python, C/C++, Java, SQL (PostgreSQL), R, JavaScript (React, Node.js, Next.js), HTML/CSS
Tools	SLURM, Docker, Harbor, Linux, Git, FastAPI, Flask, SQLAlchemy, MongoDB, Scikit-learn, Polars, Pandas, NumPy