Yihang Jiao

yihangj3@illinois.edu | 217-926-4101 | yhj3.github.io | github.com/yhj3 | linkedin.com/in/yihang-jiao-587885221

Research Interests

Machine learning for biomedical data: medical image segmentation, biomedical and scientific literature retrieval, and reliable evaluation of deep models under class imbalance and limited supervision.

Education

University of Illinois Urbana-Champaign Expected Dec 2027
Master of Computer Science — GPA 3.91/4.0 **University of Illinois Urbana-Champaign** May 2025
B.S. in Mathematics and Computer Science — GPA 3.83/4.0

Research Experience

FOCAL Lab, UIUC — Prof. Gagandeep Singh Aug 2024 – Dec 2024
Research Assistant — Reliability of Automated Red-Teaming

- Probed text classifiers by having LLaMA-2 rewrite whole sentences rather than swapping words, then asked what such a pipeline measures: across 32 configurations, rewrites of already-successful attacks still fool the classifier only 44.5% of the time, buying fluency (median perplexity 22.9 vs. 522) at the price of half the attacks.
- **Label validity.** A prompt ablation on identical data (119/180 vs. 58/180, $p = 1.7 \times 10^{-10}$) showed the winning prompt wins by instructing the model to change the label: an independently trained referee reproduces 86–87% of those "successes", and requiring the flip to be target-specific **reverses the ranking**, 12/180 against 25/180. The cosine-similarity constraint standard in this literature catches this on topic classification and misses 56% of it on sentiment.
- **Input validity.** On sentence-pair tasks most probes collapse the two segments and are not well-formed inputs; since malformed probes are scored as failures, leaving them in *understates* attack success (12.5% vs. 41.7%) — the two errors run opposite ways and do not cancel.
- **Generator and measurement validity.** The generating model refuses 6.2% of requests at a rate tracking the prompt under study, and its refusals are scored as attacks; my own earlier version of this study never ran the classifier on the generated text at all. Closes with five numbers any red-teaming report should carry.

Code: `https://github.com/yhj3/counterfactual-textattack`

Projects

Brain Tumor Segmentation on Multi-Modal MRI: U-Net vs. TransUNet Aug 2025 – Dec 2025
Independent project

- Built a unified 2D segmentation pipeline over four co-registered MRI modalities (FLAIR, T1, T1-CE, T2) in which a U-Net baseline, a configurable U-Net, and a ResNet50 + ViT TransUNet train through one entrypoint, so architectural effects separate from preprocessing and loss design. The data path is end to end: brain-region cropping from non-zero tissue, slice indexing that discards empty slices and flags tumor presence, per-slice z-score normalization, and foreground oversampling against severe class imbalance.
- Found that redesigning the objective alone — present-class foreground Dice with warmup, class weighting, foreground oversampling — recovered roughly +7 Dice points on an unchanged architecture, showing loss and sampling design matter about as much as backbone choice.
- Reached 0.7575 foreground Dice with TransUNet against 0.7326 for the tuned U-Net under a fixed 15-epoch budget; class-wise analysis localized the transformer's advantage to the diffuse tumor subregion (0.7652 vs. 0.6800) rather than to compact structures.

Code: `https://github.com/yhj3/brats-unet-vs-transunet`
Weights: `https://huggingface.co/yihangj3/brats-transunet`

VibeRepair+: Retrieval-Augmented Program Repair for Java and Python Jan 2026 – May 2026
CS Software Engineering for GenAI — co-author

- Designed a pluggable retriever framework over a 97,402-example bug-fix corpus in which adding a retrieval strategy takes one Python file and no changes to the serving endpoint, evaluation harness, or metrics; reproducibility enforced by a corpus SHA-256 fingerprint and a locked holdout split.
- Diagnosed the incumbent leave-one-out evaluation as a self-match artifact — every retriever scored a perfect recall@1 while agreeing on only 34.1% of top-1 results — and replaced it with an external-query protocol on 521 real Defects4J bugs, with an operator-aware token-overlap proxy ground truth and a two-stage embedding plus edit-distance leak check.
- Under the corrected protocol, lexical BM25 beat the dense-embedding baseline by 27.8% relative recall@10, inverting the usual dense-beats-lexical assumption; a Reciprocal Rank Fusion hybrid led five of six metrics and recovered 15 top-10 hits missed by both sub-retrievers.

Code: `https://github.com/yhj3/Vibe_Repair_Plus`

LLM RAG Reranker for Scientific Literature with Citation Context Aug 2024 – Dec 2024
Text mining, registered as CS 397 (Prof. Jiawei Han's section) — co-author

- Built a literature retrieval system that reranks candidate papers through LLM-adjudicated pairwise comparisons, treating the noisy judgments as a dueling-bandit problem and applying UCB Elimination to cut roughly 190 exhaustive comparisons at $K = 20$ down to about 44.
- Curated a corpus of 10,395 papers and 497,210 citation contexts by breadth-first traversal of LaTeX sources; augmenting candidates with in-link citation context — how a paper is used by others, rather than how it describes itself — raised MRR from 0.485 to 0.572 at $K = 5$.

Write-up: `https://yhj3.github.io/projects/scilit-rag.html`

Skills

ML: PyTorch, CUDA, HuggingFace Transformers, mixed-precision and multi-GPU training, 4-bit quantization, FAISS, scikit-learn, NumPy, pandas. **Biomedical/scientific data:** multi-modal MRI handling and normalization, 2D/3D segmentation, Dice evaluation under class imbalance, large-scale corpus construction. **Programming:** Python, C++, C#, Java, Git, LaTeX.