Zhen Yang

New Haven, CT | zhen.yang.zy372@yale.edu | Website | GitHub | LinkedIn

Summer 2027: Eligible to apply for CPT authorization through Yale

Education

Yale University — **New Haven, CT**
M.S. in Biostatistics, Data Science Pathway — *Sep 2026 – May 2028 (Expected)*

Nanjing University of Posts and Telecommunications — **Nanjing, China**
B.Eng. in Data Science and Big Data Technology — *Sep 2022 – Jun 2026*

Selected Research

Full publication list: personal website

K/V Restoration: A Training-Free Repair for Uncertainty-Behavior Drift in Quantized Language Models *1st Author · AAAI 2027 · Under Review*

- Diagnosed quantization-induced uncertainty/behavior drift via candidate-distribution JS, prediction/correctness flips, confidence/margin/entropy/NLL and Realized Attention-Relation Mismatch; FP attention-output interventions localized the dominant tested repair opportunity to the attention path and K/V projections.
- Proposed training-free KVR-Soft/Hard mixed-precision K/V repair with no inference-time FP forward/cached activations; across 10 GPTQ models, 9 NLP benchmarks and 6 calibration bases, reduced candidate-distribution divergence by 16.2/16.9% and Prediction Flip by 11.1/11.4% at comparable accuracy.

Accuracy Is Not Fidelity: Decision Churn in Low-Rank Compressed Language Models *1st Author · AAAI 2027 · Under Review*

- Reframed low-rank compression as dense-reference decision fidelity; introduced Prediction Flip, Damage/Repair/Wrong-to-Wrong decomposition and HiddenChurn, and derived BAR, a parameter-free boundary analysis linking reconstruction-residual direction, score sensitivity and dense-model margin.
- Audited 235 compressed checkpoints from 6 dense LLMs, 14 variants and 7 benchmarks: average accuracy fell 19.07 pp while 39.22% of predictions changed and HiddenChurn averaged 20.15%; 216 accuracy-matched cross-method pairs still differed in PredFlip by up to 6.97 pp.

Target-Aware Calibration Data Selection for Preserving Uncertainty in Quantized Language Models *1st Author · EMNLP 2026 Findings*

- Proposed DPQ, treating calibration-data selection as deployment-target-specific preservation of full-precision uncertainty behavior; selected high-doubt/low-margin samples plus generic anchors without modifying GPTQ/AWQ, model parameters or inference paths.
- Evaluated 8 language models, 9 NLP datasets and 22 comparison methods; boundary-heavy mixtures best preserved SQuAD2 answerability boundaries, while milder/single-signal variants generalized better across broad MCQA.

EviE: Evidential Exploration for Value-Based Reinforcement Learning *1st Author · NeurIPS 2026 · Under Review*

- Converted Dirichlet vacuity into a state-adaptive exploration rate; proved tabular $1/N(s)$ decay and GLIE/Q-learning convergence while adding only a lightweight evidential head in deep RL.
- Validated across Chain/Cliff/RiverSwim, DeepSea, 4 Classic-Control and 5 MinAtar tasks against DQN, NoisyNet, Bootstrapped DQN, RND and SUNRISE-UCB, with strongest gains in hard-exploration settings.

Industry Experience

Baidu — **Beijing, China**
Algorithm Intern - Search Retrieval — *Jan 2026 – Jun 2026*

- Owned post-training, data construction and evaluation for a dual-encoder query-URL retrieval pipeline; diagnosed term omission/mismatch, semantic misinterpretation and relevant-but-not-answerable failures.
- Designed hard-negative mining, loss reweighting and staged optimization to strengthen retrieval boundaries, improving average Recall by 10.5% across multiple cutoffs and evaluation sets.
- Built four LLM-driven hard-case generation routes for answer-preserving paraphrase, condition perturbation, confusion-pair reconstruction and URL-to-query generation, adding +3.2% sequential average Recall.
- Introduced LLM+rule teacher scoring over candidate URLs and listwise post-training/distillation for finer ranking supervision, delivering a further +6.0% sequential average Recall improvement.

HONOR — **Nanjing, China**
Algorithm Intern - Generative Engine Optimization — *May 2025 – Dec 2025*

- Independently built the core GEO system to increase HONOR recommendations in connected LLMs: mined user concerns from Weibo/Xiaohongshu, allocated weekly content by keyword heat and generated product articles through an iterative generation workflow.
- Developed an automated posterior evaluator that queried target prompts after publication and separately tracked source-list inclusion and final product recommendation, turning live LLM behavior into the main optimization signal.
- Built an 8B LoRA+contrastive prior evaluator; cold-started each query with 45 high-ranking SEO pages not selected by the LLM as negatives and 5 LLM-selected reference pages as positives, then refreshed training data from posterior observations.
- Reached >70% product visibility on 14 core prompts with 67 initial articles across 13 online outlets; later supported 1,000+ articles and sustained >65% visibility for six months, with the Python workflow estimated to save RMB 4M/year and cut generation time 40%.

Technical

Python, PyTorch, Hugging Face; LLM quantization/compression, retrieval/ranking, uncertainty estimation, reinforcement learning, large-scale evaluation