

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

M.S. in Biostatistics, Data Science Pathway

Nanjing University of Posts and Telecommunications

B.Eng. in Data Science and Big Data Technology

New Haven, CT

Sep 2026 – May 2028 (Expected)

Nanjing, China

Sep 2022 – Jun 2026

Industry Experience

Baidu

Algorithm Intern - Search Retrieval

- 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 that produced semantically similar but unusable results.
- Designed hard-negative mining with higher loss weights on confusing false recalls plus staged optimization, strengthening decision boundaries and improving average Recall by 10.5% across multiple Recall cutoffs and evaluation sets.
- Built four LLM-driven query-URL hard-case routes: answer-preserving paraphrases, condition/term perturbations that break answerability, confusion-pair reconstruction from bad URLs, and URL-to-query generation for exact/high-answerability matches; added +3.2% sequential average Recall.
- Introduced offline 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.

Beijing, China

Jan 2026 – Jun 2026

HONOR

Algorithm Intern - Generative Engine Optimization

- Independently built the core GEO optimization system to increase HONOR product recommendations in major connected LLMs: mined user concerns from Weibo/Xiaohongshu, allocated weekly article volume by keyword demand and generated product content through an iterative Python workflow.
- Developed an automated posterior evaluator that queried target prompts after publication, tracked whether articles entered retrieved-source lists and whether HONOR products appeared in final recommendations, and produced weekly reports for strategy updates.
- Built a pre-publication evaluator by LoRA+contrastive fine-tuning an 8B model; 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.
- Published 67 initial GEO articles through 13 online outlets for HONOR 400 Pro/Magic V5, reaching >70% visibility on 14 core prompts; later supported 1,000+ articles and sustained >65% visibility for six months. A targeted five-article strategy moved visibility from 0.8% to 63.5% (83.3% on DeepSeek); the workflow was estimated to save RMB 4M/year and cut generation time 40%.

Nanjing, China

May 2025 – Dec 2025

Bosch

Algorithm Intern - Automotive Ethernet / Developer Tools

- Designed a rule-driven gateway generator converting structured Excel requirements into complete C++ message parsing, transformation/port adaptation and test code, reducing repetitive implementation from days or weeks to under 30 minutes.
- Built automated differential validation for a lightweight Protocol Buffer implementation required under embedded-resource constraints: compared against Google's reference library on real vehicle signals and randomized legal-range cases, traced cross-language rounding differences and raised output consistency from 98.4% to 100%.
- Implemented RTP/RTSP/RTCP components in Python and studied/implemented IEEE 1588 and gPTP synchronization logic, documenting protocol behavior and integration details for the team.

Suzhou, China

Jul 2024 – Dec 2024

Industrial and Commercial Bank of China (ICBC)

Algorithm Intern - Credit Risk Modeling

- Cleaned, integrated and standardized multi-source customer data, resolved schema inconsistencies and engineered model-ready features for an internal credit-risk model-comparison pipeline.
- Implemented KNN, Naive Bayes and MLP loan-risk baselines with parameter/feature tuning, with the best lightweight model reaching 72% accuracy.
- Developed an LSTM-Transformer temporal model over three months of behavioral data, reaching 0.89 AUC and identifying 85% of high-risk customers two weeks in advance as an experimental early-warning model.

Changzhou, China

Jan 2024 – Feb 2024

Selected Research

Full list: personal website

K/V Restoration: A Training-Free Repair for Uncertainty-Behavior Drift in Quantized Language Models *1st*

Author · AAAI 2027 · Under Review

- Proposed training-free mixed-precision K/V repair for quantization-induced behavioral drift; across 10 GPTQ models, 9 NLP benchmarks and 6 calibration bases, reduced candidate-distribution divergence by up to 16.9% and Prediction Flip by 11.4% at comparable accuracy.

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

· EMNLP 2026 Findings

- Developed target-aware calibration-data selection for uncertainty-preserving quantization across 8 language models, 9 NLP datasets and 22 comparison methods without modifying GPTQ/AWQ or inference.