Yuan Shan

Master of Statistical Science Student · Duke University

ys328@duke.edu | (919) 697-2488 (US) / +86 159-1080-1104 (CN) | github.com/cedricshan
linkedin.com/in/yuanshan20011218 | cedricshan.github.io

EDUCATION

Duke University Aug 2024 – May 2026
M.S. in Statistical Science *Durham, NC, USA*

Duke University Jul 2020 – Aug 2024
B.S. in Interdisciplinary Studies — Data Science *Durham, NC, USA*

Duke Kunshan University Jul 2020 – Aug 2024
B.S. in Data Science *Kunshan, China*

RESEARCH INTERESTS

Large language models, reinforcement learning (RLHF, GRPO), self-evolving language model agents, statistical modeling, causal machine learning, design of experiments.

PUBLICATIONS

Z. Ma*, **Yuan Shan***, J. Zhao, N. Xu, L. Wang. *"MEOW: End-to-End Outline Writing for Automatic Academic Survey."* **IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2026.** [IEEE Xplore] [arXiv] [Code & Benchmark] * *Equal contribution.*

Reformulates survey outline writing as an end-to-end task that maps paper metadata to a hierarchical outline. Curates a high-quality dataset distilled from 2.82M arXiv, 252k bioRxiv, and 72k medRxiv records, with Chain-of-Thought reasoning chains generated via DeepSeek-R1. Trains an 8B reasoning model (Qwen3-8B) in two stages: SFT cold-start followed by GRPO with a Tree-Edit-Distance structural reward and a format-compliance reward. Establishes an LLM-as-a-Judge benchmark covering five criteria (Structure Locate / Detail, Content Exclusion / Depth, Pragmatics Concise) plus a structural-distance metric. **Meow-8B-SFT-GRPO** surpasses SurveyX (35.62 vs. 29.60 LLM-Judge total) and outperforms GPT-5 Nano, DeepSeek-R1, and Gemini 2.5 Flash-Lite on the SurveyX test set.

WORKING PAPERS

Self-Evolving LLM Agent. *In preparation.*

Research project on reinforcement-learning-driven self-evolution for large language model agents.

INDUSTRY & RESEARCH EXPERIENCE

Wenge Tech — LLM Algorithm Engineer Intern May 2025 – Oct 2025
Beijing, China

Work accepted at IEEE ICASSP 2026 (co-first author)

- Co-designed and built **MEOW**, the first metadata-driven, end-to-end LLM framework that maps paper

metadata to hierarchical academic survey outlines in a single inference step, replacing prior multi-agent template workflows that produce shallow, rigid outlines.
- Built the data pipeline that curated and filtered **2.82M arXiv**, **252k bioRxiv**, and **72k medRxiv** records into a high-quality survey dataset; enriched references via vector retrieval (`all-MiniLM-L6-v2`) and distilled Chain-of-Thought reasoning chains using **DeepSeek-R1** to bridge metadata input and outline output.
- Trained **Qwen3-8B** in two stages: SFT cold-start on the CoT-annotated corpus, followed by Group Relative Policy Optimization (**GRPO**) with a Tree-Edit-Distance *structural similarity reward* and a binary *format-compliance reward*, weighted by λ.
- Established an LLM-as-a-Judge evaluation framework covering five criteria (Structure Locate / Detail, Content Exclusion / Depth, Pragmatics Concise) plus a structural-distance metric; released the benchmark, training corpus, and evaluation scripts publicly on Hugging Face and GitHub.
- Used **vLLM** to accelerate both inference and RL training; **Meow-8B-SFT-GRPO** surpassed SurveyX (35.62 vs. 29.60 LLM-Judge total) and beat GPT-5 Nano, DeepSeek-R1, and Gemini 2.5 Flash-Lite on the SurveyX test set.

Lakala Payment — Data Engineering Intern Jul 2024 – Aug 2024

Beijing, China

- Assisted in the development and maintenance of the company's internal data platform using Hadoop ecosystem tools.
- Designed and developed the internal data platform's data quality monitoring system.

Beijing Academy of Artificial Intelligence (BAAI) — Research Intern Jun 2023 – Oct 2023

Beijing, China

- Contributed to BAAI's 3D electron-microscopy (EM) processing toolchain, focusing on advanced neuron-segmentation methodologies including PyTorch Connectomics (PyTC) and Local Shape Descriptors (LSD).
- Built reproducible training pipelines for large-volume connectomics data, covering preprocessing, augmentation, and patch-based inference for cubic-micron EM volumes.
- Optimized segmentation models via systematic hyperparameter tuning combining Grid Search and Bayesian Optimization, materially improving segmentation quality on benchmark volumes.

Tencent — Machine Learning Engineering Intern Jun 2022 – Aug 2022

Tencent Video, Shenzhen, China

- Independently owned an end-to-end ML project to surface high-potential video categories on the Tencent Video platform, from data extraction and cleaning through algorithm design, evaluation, and stakeholder reporting.
- Built scalable data pipelines on Tencent's internal task-scheduling platform, handling dependency resolution across heterogeneous data sources.
- Engineered an XGBoost-based scoring model producing quantitative indicators of video-category potential; tuned the model with cross-validated hyperparameter search.
- Wrote production SQL and Python against large-scale internal data warehouses to extract, join, and transform features at scale.
- Deployed Spark for distributed feature engineering and model scoring, keeping the pipeline tractable on volumes of up to **7,500,000 records**.
- Synthesized findings into two project reports that informed downstream content-strategy and operational decisions on Tencent Video.

PROJECTS

Music Chat Recommender — LLM-mediated multi-turn recommender Apr 2026

Live Demo | *GitHub*

- Built a chat-first music recommender that parses free-form English requests (e.g. "sad songs for a rainy evening, but no Coldplay"), runs live searches, and returns an ordered short-list with embedded YouTube playback.
- Designed a **two-call LLM architecture**: an intent-parsing call extracts a Pydantic-typed `SearchQuery` (mood, genres, exclusions, year range, control flags) from the latest user turn plus a serialized memory; a rerank-and-reply call selects from the candidate pool and writes the human-facing reply, with both outputs validated against Pydantic schemas to make malformed LLM outputs unrecoverable failures rather than crashes.
- Implemented a **provider abstraction** over **Groq Llama 3.3 70B** (primary, JSON mode) and **Gemini 2.5 Flash-Lite** (fallback) to absorb mid-project free-tier policy changes from both Spotify (API deprecation, Nov 2024) and Google.
- Pivoted the music backend from Spotify to the keyless **iTunes Search API** after the Spotify Web API was restricted; reconstructed similarity reasoning from search alone via a fan-out strategy that issues 1–6 facet-specific queries per intent.
- Built per-session `ConversationState` (likes, dislikes, must-exclude artists, history) so follow-ups like "stop recommending Drake" and "more diverse next time" work without restating context; orchestrated everything in a single Gradio app deployed publicly on Hugging Face Spaces.

Age-Dependent Heterogeneity in the PA – Mental-Distress Association Apr 2026

Causal ML on BRFSS 2015–2024 | *GitHub*

- Pooled **ten consecutive annual waves** of the U.S. Behavioral Risk Factor Surveillance System (BRFSS, 2015–2024; **n = 3,242,218** adults after complete-case filtering) to test whether the protective association between leisure-time physical activity (PA) and frequent mental distress (FMD) varies systematically by age.
- Fit **survey-weighted logistic regression** with raked sampling weights, sandwich SEs clustered on PSU, and a formal PA × Age likelihood-ratio test for effect modification; ran age-stratified models within each of six age groups and a three-way PA × Age × Year interaction for temporal change.
- Independently estimated heterogeneous treatment effects via **Causal Forest under Double Machine Learning** (`CausalForestDML` in EconML), with HistGradientBoosting nuisance models and 2-fold cross-fitting, recovering the conditional ATE as a function of covariates without pre-specifying subgroups.
- **Found a striking, monotonic age gradient**: PA odds ratio for FMD ranges from **0.89** (ages 18–24) to **0.50** (ages 55–64); the Causal Forest independently identifies **age as the dominant heterogeneity driver** (feature importance = 0.39, $2.5\times$ the next predictor).
- **Documented a novel temporal finding**: the already-weak protective effect among 18–24-year-olds has been eroding over the decade, reaching the null in both **2018 and 2024** and paralleling the deepening youth mental-health crisis.
- Validated robustness with **E-values** for unmeasured confounding, propensity-score overlap diagnostics, a placebo test on a non-causal outcome, and a conditional-imputation sensitivity analysis on missing income.

Optimizing Turbine Blade Design under Operational Uncertainty Apr 2026

Robust DOE + Surrogate Modeling | *GitHub*

- Tackled a **constrained robust optimization** problem on a finite-element gas-turbine-blade simulator: minimize expected maximum von Mises stress subject to a strict displacement feasibility constraint ($d < 1.3 \times 10^{-3}$ m), under uncontrollable perturbations of cooling temperature ($\pm 2^{\circ}$C) and pressure load ($\pm 10^{4}$ Pa), all within a hard **300-evaluation simulation budget**.

- Designed a **four-phase pipeline**: (i) 100-point maximin Latin Hypercube Design over the 6-D unit hypercube; (ii) Gaussian Process surrogate with **Matérn-5/2 kernel and ARD length-scales**, achieving **10-fold CV** $R^2 > 0.994$ on both stress and displacement; (iii) 150 sequential runs driven by an **Expected Improvement with Constraints** acquisition, augmented by a Gaussian proximity penalty for batch diversification; (iv) 27-run robust grid validation (3×3 perturbation grid on 3 candidates) plus 23 random validation runs.
- Achieved a **41% reduction** in maximum stress relative to the LHD best, with the recommended design satisfying displacement feasibility on **100%** of all tested perturbations and a **45% safety margin** below the failure threshold; ARD length-scale analysis isolated **CTE** (Coefficient of Thermal Expansion) as the dominant design factor.
- Built a three-stage **SLURM pipeline** on the Duke Computing Cluster (input generation → array-job parallel simulation with `-array=1-N%40` → chained result merging via `-dependency=afterok`); 40-way parallelism delivered a ∼**33**× **wall-clock speedup** (4 min vs. 2.2 h for a 100-run batch), enabling rapid sequential iteration.

Paper Helicopter — Sequential DOE Apr 2026

STA 643 | GitHub

- End-to-end design optimization via fractional-factorial screening, confirmation runs, response-surface (RSM) optimization, and validation experiments.

Bradley–Terry Analysis of the English Premier League Mar 2026

Slides | GitHub

- Applied the **Bradley–Terry** paired-comparison framework to **7,860** EPL match results across 21 seasons (2005/06–2025/26, 44 clubs) to estimate relative team strengths and quantify home advantage; reformulated the base BT model as logistic regression to enable standard MLE inference and asymptotic Wald / likelihood-ratio tests.
- Extended the analysis with the **Davidson** formulation to absorb the 24% of matches ending in draws, plus **season-specific** and **team-specific** home-advantage intercepts; validated each extension via nested LRT and AIC comparison.
- Estimated the home-team odds boost at $\hat{\theta} \approx 1.64\times$ ($p < 10^{-15}$ under all model variants); team-strength estimates from BT and Davidson agreed at $r = 0.990$, confirming robustness.
- Detected a **significant secular decline** in home advantage over the study window and a **near-complete collapse** during the 2020/21 spectator-less COVID season ($\hat{\alpha} \approx -0.10$) — direct empirical support for the causal role of crowds.

HIV Self-Testing Study — Reproduction Jan 2026

Reproducibility in Public Health | GitHub

- Careful reproduction of a published HIV self-testing study, validating statistical methodology, effect estimates, and robustness checks.

Survey Outline Evaluation Benchmark Sep 2025

Companion code to ICASSP 2026 | GitHub

- End-to-end pipeline for generating and evaluating academic survey outlines with LLMs — multi-dimensional metrics, statistical analysis, and a public benchmark.

Football Transfer Market Visualization Sep 2023 – Oct 2023

Team Leader (3-person team) | Live Demo | GitHub

- Led a 3-person STATS 401 team end-to-end: scoped the research questions, defined the data model, and split engineering and design responsibilities.
- Wrote the full Python data-cleaning pipeline; designed the interactive UI and visual encoding from scratch

in D3.js, letting users drill down by club, season, and league.
- Shipped a publicly hosted visualization at transfer-market-vis.netlify.app.

Internal Data Quality Monitoring System Jul 2024

Independent project | GitHub

- Python-based data-quality monitoring system designed for large-scale data environments; uses statistical methods to evaluate data quality and triggers alerts when anomalies are detected.

HONORS & AWARDS

- Dean's List with Distinction, Duke Kunshan University.
- Finalist Prize, Interdisciplinary Contest in Modeling (MCM/ICM).
- First Prize, National Olympiad in Informatics in Provinces (NOIP).
- Outstanding Peer Mentor and Outstanding Peer Tutor, Duke Kunshan University (2022–2024).
- Best Volunteer, 7th China Buyout Fund Annual Conference, Chinese Museum of Funds.

LEADERSHIP & SERVICE

- **Technical Intern**, Beijing 2022 Olympic & Paralympic Winter Games — Organising Committee. Dec 2021 – Jan 2022
  Conducted statistical analysis of the IOC global technical partners' transportation demands during pre-Games operations; compiled briefing materials informing vehicle allocation and route planning.
- **Volunteer**, Duke China-U.S. Summit 2023 Committee. Jun 2023 – Jul 2023
- **Student Leader**, Kunshan Student Orientation Peers (KSOP) at Duke. Nov 2022 – May 2023

TECHNICAL SKILLS

Programming Languages: Python, R, SQL, JavaScript (D3.js), Bash.

Machine Learning & LLM: PyTorch, HuggingFace Transformers, vLLM, TRL, LoRA / PEFT, DeepSpeed, XGBoost, scikit-learn.

Data & Infrastructure: Hadoop, Spark, Causal Forests, Bayesian Optimization, Design of Experiments, Bradley–Terry models.

Tools: Git, LaTeX, Gradio, Linux, Jupyter.

Languages: English (professional working proficiency), Mandarin Chinese (native).