# Yuan Shan

M.S. in Statistical Science · Duke University · cedricshan.github.io

## EDUCATION

**Duke University** Aug 2024 – May 2026
*M.S. in Statistical Science*

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

## EXPERIENCE

**Wenge Tech — LLM Algorithm Engineer Intern** May 2025 – Oct 2025
*[Paper] [Code & Benchmark] · Accepted at IEEE ICASSP 2026*

- Developed **MEOW**, the first metadata-driven, end-to-end LLM framework for academic survey outline generation, combining supervised fine-tuning (SFT) and reinforcement learning (GRPO) within the RLHF paradigm.
- Designed and implemented a benchmark and multi-dimensional evaluation metrics (structure, content, pragmatics) for outline quality assessment; curated and distilled ∼20k+ survey datasets from arXiv, bioRxiv, and medRxiv.
- Constructed Chain-of-Thought (CoT) annotations to guide structured reasoning, materially improving logical taxonomy construction and overall coherence of generated survey outlines.
- Optimized **Qwen3-8B** model (full & LoRA) with SFT and reinforcement learning; designed reward functions (structural distance & format compliance) to align generation with human writing preferences.
- Leveraged **vLLM** to accelerate inference and RL training, achieving state-of-the-art performance, surpassing SurveyX and outperforming strong LLMs (e.g., GPT-5 Nano, DeepSeek-R1) on evaluation benchmarks.

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

- Contributed to the development of the 3D electron microscopy processing toolchain at BAAI, with a focus on advanced neuron segmentation methodologies such as **PyTorch Connectomics (PyTC)** and **Local Shape Descriptors (LSD)**.
- Optimized neuron segmentation models via hyperparameter tuning using **Grid Search** and **Bayesian Optimization**.

**Tencent — Machine Learning Engineering Intern** Jun 2022 – Aug 2022

- Independently undertook a project to identify potential video categories on the Tencent Video platform, managing the entire process from data extraction, cleaning, and algorithm implementation through to the reporting of results.
- Established pipelines for task scheduling and dependency management on the internal task-scheduling platform.
- Developed machine learning models based on **XGBoost** to construct quantitative indicators for video potential; used SQL and Python for data retrieval and ML model implementation.
- Employed **Spark** for distributed computing to enhance pipeline efficiency with data volumes up to **7,500,000 records**.
- Authored two project reports providing insights for Tencent Video operations.

## SELECTED PROJECTS

**Music Chat Recommender** Apr 2026
*[Live Demo] [GitHub]*

- Built an interactive music-recommendation chatbot with **Gradio** as the front-end and **Groq Llama 3.3 70B** as the LLM backend, with multi-turn conversation memory so the assistant tracks user preferences across turns.
- Connected the chatbot to the **iTunes Search API** for track and artist metadata and the **YouTube Data API** for playable video links, returning recommended songs together with album art and direct YouTube embeds inside the chat.

**Bradley–Terry Analysis of the English Premier League** Mar 2026
*[Slides] [GitHub]*

- Fit a **Bradley–Terry** paired-comparison model — reformulated as logistic regression to enable standard MLE inference — to 21 seasons of English Premier League match data (2005/06–2025/26).
- Extended the base model with the **Davidson** formulation for draws plus season-specific and team-specific home-advantage parameters; validated each extension via Wald and likelihood-ratio tests against nested baselines.
- Found home-field advantage highly significant overall, with a clear secular decline across the study period and a near-complete collapse during the 2020/21 COVID season — direct empirical support for the causal role of spectators.

**Football Transfer Market Visualization** Oct 2023
*[Live Demo] [GitHub]*

- Aimed to provide insights into the evolution trends and key influencing factors of the global football transfer market via Python-based data cleaning and processing of multi-season transfer data.
- Built interactive data visualizations and UI design in **D3.js**; final dashboard deployed publicly online.

## HONORS

Dean's List with Distinction (DKU) · Finalist Prize, MCM/ICM · First Prize, National Olympiad in Informatics (NOI) in Provinces

ys328@duke.edu · (919) 697-2488 · github.com/cedricshan · linkedin.com/in/yuanshan20011218