Yuan Huang 黄远

huangyuan4629@gmail.com
Homepage · Google Scholar · GitHub
Northeastern University, Shenyang, China

RESEARCH INTERESTS

Natural language processing, with a focus on **measurement validity** — whether the benchmarks and automatic judges used to rank models measure what they claim to. One result recurs across my projects: **surface competence ≠ underlying competence**. I want to measure that gap, explain it mechanistically, and close it. My current project is on the **interpretability of bias in multimodal models**: identifying the internal semantics that carry a bias, and mechanisms to correct it.

EDUCATION

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) — Visiting Student

Oct 2026 – Jan 2027

- Hosted by Xiuying Chen (Assistant Professor, MBZUAI), my collaborator since 2025

Northeastern University, China — B.Eng., Computer Science and Technology

2023 – 2027 (expected)

- CGPA **4.0481 / 5.00** (90 / 100) · **Rank 5 / 110** (top 4.55%) in major cohort
- Advised by Shiqi Zhao (Assistant Professor, Northeastern University) since 2024

PUBLICATIONS & RESEARCH

[1] Do MLLM Judges Judge the Edit? Auditing Bias in Image Editing Evaluation with Verified Quality Preservation

Yuan Huang*, Zirui Song*, Xiuying Chen *(* equal contribution)*

EditJudgeBias · *Under review at ICLR 2027* · co-first author · arXiv:2610.01670, Oct 2026

- **Design.** 1,196 real editing samples and 13 cues across four evaluation sites. A shift counts as bias only if the cue is verified to preserve edit quality, and is read against the judge's own zero-dose and re-query noise floors, not against zero.
- **Findings.** Quality-preserving cues move all five MLLM judges beyond their own noise; swapping candidate order alone reverses up to **60.9%** of pairwise decisions, against 1.0–10.5% on re-query.

[2] Two Routes to the Middle: Placement Search and Brain Readouts Converge on Where Continual Learners Should Specialize

Yuan Huang, Zihan Chen, Runbin Zhang, Hongwei Ding, Changzeng Fu, Shiqi Zhao

LS-B · *Under review at ICLR 2027* · first author · arXiv:2610.01590, Oct 2026

- Placement search over ViT depth traces an **inverted U** (up to 3.5 pp); LS-B reads early tasks through an fMRI encoding model of twelve human visual areas and picks blocks overlapping the peak — on Split ImageNet-R, within **1.5 pp** of full-BiLoRA at **60%** of the storage, with no labels or backpropagation.

[3] Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies

Zirui Song*, **Yuan Huang***, Junchang Liu*, et al. *(* equal contribution)*

WereBench · *EMNLP 2026, Main Conference* · co-first author · arXiv:2510.11389, Oct 2025 · Dataset

- Benchmark built from **100+ hours** of professional human social-deduction gameplay (**32.4M utterance tokens**, 30 roles, 15 rule variants), scored against **winning-side MVP trajectories** rather than LLM self-play; 5 verbal dimensions and 2 decision tasks.
- **Half of the evaluated models score below 0.50** (best 0.720). Deception reasoning and counterfactual trade-off are weakest while persuasive generation stays strong — **fluency ≠ strategic correctness**.

[4] Geolocation with Real Human Gameplay Data: A Large-Scale Dataset and Human-Like Reasoning Framework

Zirui Song, Jingpu Yang, **Yuan Huang**, et al.

Geolocation · *Preprint* · arXiv:2502.13759, Feb 2025 · Dataset · Code

- Dataset from a public geo-guessing platform — **740K users, 2.7M locations, 25M human response records** — with human accuracy calibrating item difficulty. Multi-step geographic reasoning (GeoCoT) improved classic geolocation metrics by up to **25%**, and reasoning quality as scored by GeoEval by **9%**.

HONORS & AWARDS

- **China National Scholarship** — national level · 2023–2024
- Northeastern University Comprehensive Scholarship · 2024–2026
- **Third Prize, National Finals**, Lanqiao Cup Programming Competition · 2024

SKILLS & LANGUAGES

- **Programming.** Python, C++, Java · PyTorch, Hugging Face Transformers · backend engineering intern, 2024 (Java / Spring Boot)
- **Languages.** English — IELTS Academic 7.0 (Oct 2026) · Chinese (native)