Guijia Zhang

Phone: (+86) 188-2018-3845 | Email: 2022040399@email.szu.edu.cn

Research Interests: Agent Security and Safety, Trustworthy AI, LLM Reasoning, Agent Systems, Code LLM Security

Academic Profile My research focuses on trustworthy agent systems, agent security, and LLM reasoning safety. I have built a complete research loop from problem formulation, benchmark construction, and risk modeling to system validation and paper writing, with current first-author/co-first-author submissions to EMNLP 2026, CoLM 2026, and NeurIPS 2026.

Research Potential I am especially interested in turning real system failures into measurable research problems, including request-conditioned safety auditing, procedural knowledge representation, and runtime regulation. I bring cross-institution collaboration experience, English academic writing practice, and reproducible experiment organization.

Education

Shenzhen University | B.Eng. in Elite Innovation Program in Electronic Information Engineering 2022.09 -- 2026.06

- Core coursework: Machine Learning, Artificial Intelligence, Computer Networks, Data Structures, Mathematical Analysis, Advanced Algebra, and Stochastic Processes.
- Honors: Academician Scholarship and Star of Culture and Sports Award; intensive honors-program training strengthened mathematical foundations, English literature reading, and technical writing.

Research Experience

HKUST Academy of Interdisciplinary Studies | Remote Research Intern · Advisors: Huamin Qu, Yuzhe Shi 2026.02 -- 2026.05

- Contributed to the co-first-author paper Agent skill regulation as a procedural knowledge representation problem, modeling execution commitments, gate conditions, and recovery actions in human-written skills.

KAUST | Remote Research Collaboration · Collaborators: Di Wang, Shu Yang 2026.01 -- Present

- Produced STARS by formulating request-conditioned invocation safety, constructing SIA-Bench, designing a risk-auditing framework, and writing the paper; further collaboration is expected on agent applications and runtime safety.

HKUST (Guangzhou) | Exchange Student · Advisor: Lei Chen 2025.12 -- 2026.06 (Expected)

Research Focus: Agent Memory, Databases

- Surveyed MemOS, MIRIX, AgentFold, Recursive Language Models, and ReasoningBank to identify bottlenecks in long-context memory retrieval, state updates, and long-horizon dependency modeling.
- Built SFT data with BrowserAgent and converted multi-turn trajectories from turn-level to task-level supervision; studied trade-offs among memory write policies, state compression, and retrieval latency.

Tsinghua University HCI Lab | Research Assistant (Remote) · Advisors: Chun Yu, Yuanchun Shi 2024.12 -- 2025.08

- Contributed to SynergyAdvisor by modeling expert-avatar co-evolution as a human-AI alignment problem with memory-aware decision making and expert-feedback-enhanced Chain-of-Thought optimization.
- Worked on long-context cross-topic Agentic RAG, building a benchmark and retrieval-enhancement pipeline to analyze information recall failures and context drift in complex dialogue.

Selected Publications

- **First Author**, Hallucination as Exploit: Evidence-Carrying Multimodal Agents. **(Planned Submission to EMNLP 2026)**
- **First Author**, STARS: Skill-Triggered Audit for Request-Conditioned Invocation Safety in Agent Systems. **(Planned Submission to CoLM 2026)**
- **Co-First Author (Second)**, Agent skill regulation as a procedural knowledge representation problem. **(NeurIPS 2026 Submission)**
- **Third Author**, Document image tampering localization based on foundation models. **(ICDAR 2025)**
- **Third Author**, A Diffusion-Based Approach for Restoring Face-swapped Images. **(APSIPA ASC 2024)**
- **Main Contributor**, Spatial Causal Prediction in Video. **(CVPR 2026 Findings)**
- **Main Contributor**, EchoRAG: Real-Time Retrieval Augmented Generation with Episodic Memory. **(arXiv 2025)**

Research Projects

Hallucination as Exploit: Evidence-Carrying Multimodal Agents 2026.05 -- Present

First Author; Planned Submission to EMNLP 2026

- Formalized hallucination-to-action conversion, where unsupported perceptual claims become authorization for privileged actions, and proposed Evidence-Carrying Agents (ECA).
- Decomposed tool calls into action-critical predicates with DOM/OCR/AX verifier certificates; reduced red-team gate bypass from 15% to 1.3% across 1,900 attacks and achieved 0% unsafe-action rate on 200 E2E tasks.

STARS: Skill-Triggered Audit for Request-Conditioned Invocation Safety in Agent Systems 2026.03 -- 2026.04

First Author; Planned Submission to CoLM 2026

- Advanced agent skill safety to request-conditioned runtime auditing; built SIA-Bench with 3,000 invocation records and achieved 0.439 high-risk AUPRC on held-out indirect prompt injection.

Agent skill regulation as a procedural knowledge representation problem 2026.02 -- 2026.05

Co-First Author (Second); NeurIPS 2026 Submission

- Reframed skill regulation as procedural knowledge representation, modeling execution commitments, state constraints, and recovery actions for inspectable skills.
- Organized the paper structure across syntax, semantics, and execution-level representations and defined runtime regulation targets.

Research Engineering and Projects

DeepWisdom (MetaGPT) | Algorithm Evaluation Engineer Intern 2025.10 -- 2025.11

- Contributed to multi-agent code-generation benchmarks and MGX product evaluation, designing process-level evaluation workflows and using failure-case analysis to identify reliability bottlenecks in production agents.
- Built Python automation for evaluation pipelines, gaining hands-on experience with LLMs, agents, benchmarks, and code evaluation; these failures informed later agent-safety research questions.

Projects

Legal Pilot 2.0: Intelligent Contract Review System 2026.02 -- 2026.04

- Built a Qwen3-235B dual-stage contract-review workflow (legal red-line scan → rule-by-rule audit), integrating rule constraints, risk localization, and revision suggestions into a verifiable professional workflow.
- Designed privacy and performance modules with LLM detection of sensitive entities, deterministic pre-masking, length-preserving replacement, and FastAPI asynchronous scheduling for upload 15 / review 4 / desensitization 2.
- Used the system as a real long-document agent workflow testbed, strengthening understanding of latency, privacy, runtime control, and human-in-the-loop review constraints.

IntelliSelect: AI-Assisted Faculty Recruitment Pre-Screening System (HKUST-GZ) 2026.02 -- 2026.04

- Built a full-stack "Smart Assistant, Human in Charge" pipeline (FastAPI async + SQLAlchemy 2.0 + React 19/TypeScript, Dockerized) that turns subjective resume pre-screening into a multi-source, auditable, explainable workflow.
- Designed a five-dimension LLM scoring engine (academic output, research fit, teaching, background, interdisciplinary potential) on a self-hosted vLLM (Qwen3-14B), decoupling LLM per-dimension scoring from backend weighted re-aggregation to keep final scores and recommendations controllable.
- Fused structured FRS data, PyMuPDF-parsed resumes, references, gradings, and a deterministic rule-based matching engine into the prompt, with robust JSON parsing, retry/backoff, and structured explainable output (highlights, risks, evidence, qualification checks).
- Added Google Scholar benchmarking (H-index/citations with 24h cache and anti-throttling) and peer-group quartile comparison to ground each candidate against an academic peer cohort.

Coursework and Research Methods

- Methods: Familiar with benchmark design, group-safe split, OOD evaluation, risk calibration, error analysis, ablation studies, and risk-oriented evaluation metrics.
- Coursework: Strong foundations in machine learning, artificial intelligence, probability, stochastic processes, mathematical analysis, advanced algebra, computer networks, and data structures.

Research Skills

- Programming and model development: Proficient in Python and familiar with PyTorch, Hugging Face, PEFT, and Unsloth for data construction, SFT, and model evaluation.
- Agents and systems: Familiar with RAG, multi-agent systems, agent memory, and tool-use pipelines; experienced with LangChain, MetaGPT, Faiss, and Milvus.
- AI safety and research practice: Focused on agent/Code LLM security, runtime defense, and mechanistic interpretation; experienced in literature review, paper writing, experiment reproduction, and Lean-based formalization.
- Quantitative modeling: Familiar with GARCH-family models, HAR models, EVT, and basic time-series evaluation methods for VaR forecasting.

Awards and Additional Information

- National Second Prize, Huashu Cup Mathematical Contest in Modeling (Team Leader).
- Guangdong Provincial Third Prize, National Undergraduate Mathematical Contest in Modeling (Team Leader).
- Guangdong Provincial Third Prize, China Collegiate Computing Contest, Network Technology Challenge (Graduate Track, Core Member).
- Completed the AI4MATH formal mathematics course organized by Professor Zaiwen Wen at Peking University, with training in Lean and mathematical formalization.
- Long-term interest in large-model safety, agent memory, LLM reasoning, and database-related systems, with strong ability for cross-domain problem abstraction, experiment organization, and research transfer.