

Minghui Pan

Mobile: (+86) 137 0145 9590
Email: panmingh@outlook.com
Website: www.panmingh.com



Education

Beijing University of Posts and Telecommunications

Beijing, China

B.Eng. in Telecommunications Engineering and Management; GPA: 3.76/4.00

2023–2027

Research Experience

Tsinghua University

Research Assistant, advised by [Zhenpeng Chen](#); produced AAI 2027 submission

Mar. 2026 – Present

Peking University, OpenDCAI

Research Intern, advised by [Wentao Zhang](#); produced [One-Eval](#) open-source system

Dec. 2025 – Mar. 2026

Beijing University of Posts and Telecommunications

Undergraduate Researcher, advised by [Guoshun Nan](#) and [Kun Wang \(NTU\)](#); produced ACL 2026 and CVPR 2026 papers

Jul. 2025 – Jan. 2026

Institute of Information Engineering, Chinese Academy of Sciences

Research Assistant, advised by [Xiaoyan Gu](#); produced ICASSP 2026 paper

Jul. 2025 – Sept. 2025

Publications

- **Tool Specifications Matter: Uncovering and Mitigating Safety Risks in AI Agents**

AAAI 2027 (CCF-A) · Under Review · First Author, Solo Work

Mar. 2026 – Jul. 2026

We identify schema-formatted tool specifications as a key cause of safety degradation in LLM agents and propose SafeKeep, an inference-time framework that improves agent safety while preserving function-calling capability.

- **One-Eval: An Agentic System for Automated and Traceable LLM Evaluation**

EMNLP 2026 Industry Track (CCF-B) · Under Review · Co-First Author

Dec. 2025 – Mar. 2026

We build an agentic evaluation system (under PKU DCAI Lab’s DataFlow project) that converts natural-language requests into executable evaluation workflows via NL2Bench, BenchResolve, and task-aware metric reporting with human-in-the-loop checkpoints.

- **Streaming Hallucination Detection for Long Chain-of-Thought Reasoning**

ACL 2026 (CCF-A) · Findings · Co-First Author

Sept. 2025 – Jan. 2026

We propose a streaming detector that identifies both step-level and prefix-level hallucination states online during long CoT reasoning in large reasoning models, using relative-position-aware gating to improve late-stage detection under long-context accumulation.

- **Reallocating Attention Across Layers to Reduce Multimodal Hallucination**

CVPR 2026 (CCF-A) · Poster · Fifth Author

Jul. 2025 – Nov. 2025

We study hallucination in multimodal large reasoning models through staged functional specialization of attention heads, identifying perceptual bias and reasoning drift, and evaluate a lightweight, model-agnostic plugin based on functional head identification and class-conditioned rescaling.

- **CorrEctor: An Execute-to-Correct Paradigm for Efficient LLM Secure Inference**

ICASSP 2026 (CCF-B) · Poster · Fourth Author

Jul. 2025 – Sept. 2025

We propose an execute-to-correct paradigm that offloads privacy-obfuscated inputs to untrusted GPUs with parallel verification and correction in trusted environments, achieving $1.36\times$ – $3.55\times$ speedup over CPU-based TEE baselines with accuracy comparable to plaintext inference.

Research Interests

My work to date has centered on **LLM safety** and **Agent safety**. Across these projects, I have been particularly interested in understanding why models become unsafe or hallucinate, and in identifying the underlying mechanisms within their representations and reasoning processes. This perspective motivates my future research on **Agentic RL** and **Agent Alignment**, with the goal of developing post-training methods that make models **Trustworthy** by construction, rather than relying primarily on external guardrails imposed at inference time.

Projects

• BUPT Innovation Platform: Maintenance and Operation *Mar. 2025 – Present*

Maintain and operate a university-wide platform in continuous production use by 3000+ students and faculty across multiple departments. Own day-to-day reliability of the live system — triaging and resolving user-reported issues, monitoring service health, and shipping incremental fixes and improvements against real usage without disrupting active users. Sustained the codebase through successive semesters of changing requirements, and extended it with intelligent search and advisor-matching built on an LLM-based RAG system.

• LoopLM: A Self-Iterative Reasoning Framework for LLMs *Oct. 2025 – Present*

Designed a modular framework that improves safety and reliability of locally deployed LLMs via continuous hallucination monitoring and feedback-driven adaptation. Implemented step-level and cumulative hallucination detection for long chain-of-thought reasoning, enabling streaming safety assessment during generation, and built a closed-loop pipeline that supports lightweight adaptation through probes and LoRA-style modules, deriving its training signal from the model's own monitored generations without requiring any external annotated dataset.

Achievements and Awards

First Prize (National) — The Chinese Mathematics Competition (CMC)

First Prize (Beijing) — National College Students Engineering Innovation Competition, Intelligent Rescue Track

Merit Award (National) — National College Students Engineering Robot Competition

Third Prize (Beijing) — Challenge Cup College Students' Extracurricular Academic Science and Technology Competition

Third Prize (National) — National College Students Network and Information Technology Competition

Skills

English — IELTS: 6.5; CET4: 601; CET6: 475

Programming — Python, LaTeX, JavaScript/TypeScript

Machine Learning — PyTorch, Transformers, LoRA, vLLM, Agent pipelines, Common RL Methods

Tools — Git, Linux, Multi-GPU training