Hailin Liu

+7(916)000-31-55 | hereishailin@outlook.com | GitHub | Huggingface | Google Scholar

Research Interests

My research interests lie in the **runtime governance and security control of autonomous LLM-based agent systems**, with a focus on runtime protection, dynamic permission control, privacy-preserving coordination, resource-aware execution, and verifiable accountability in trustworthy multi-agent infrastructures.

Education

Lomonosov Moscow State University — Moscow, Russia
M.Sc. in Applied Mathematics and Computer Science ***(with Honours);*** — *09.2024 – 06.2026*
Specialization: Artificial Intelligence in Cybersecurity; — *GPA:* ***4.90/5.00***

Relevant Coursework: Robust Machine Learning, Interpretable Machine Learning, Attacks on AI Systems, Reinforcement Learning, Modern NLP and Large Language Models, Development of AI Agents

Lomonosov Moscow State University — Moscow, Russia
B.Sc. in Fundamental Informatics and Information Technology; — *09.2020 – 06.2024*

Research Experience

Runtime Protection Architecture for LLM-based Agentic Systems — Moscow, Russia
Researcher (Advisor: Eugene Ilyushin) — *09.2024 – Present*

- Developed SafeAgent, a runtime protection architecture for LLM-based agents against prompt-injection, tool-use, and workflow-level attacks.
- Framed agent security as a stateful decision-making problem over multi-step interaction trajectories, rather than a local input-output filtering task.
- Designed the overall architecture separating execution governance from semantic risk reasoning through a runtime controller and a context-aware decision core.
- Formalized decision operators for risk encoding, advantage-cost evaluation, consequence modeling, session policy, and state synchronization.
- Evaluated SafeAgent on Agent Security Bench and InjecAgent against baseline agents and text-level safeguard methods, demonstrating improved robustness while preserving benign-task performance.
- Led the integration of undergraduate subprojects on risk encoding, evidence extraction, and automated safety evaluation into the broader SafeAgent research program.

Machine Learning-Based SQL Injection Detection — Moscow, Russia
Undergraduate Researcher — *11.2022 – 06.2024*

- Investigated machine learning approaches for detecting anomalous SQL queries and SQL injection attacks.
- Reproduced and extended the SQLiGoT framework based on token-graph representations of SQL queries.
- Conducted comparative experiments across multiple classifiers, including Random Forest, Linear SVC, Logistic Regression, and KNN.
- Explored ensemble voting and feature selection strategies to improve detection accuracy and computational efficiency.
- Produced a full research manuscript and presented the work at a university student research conference.

Research Mentorship

SafeAgent Research Team — Moscow, Russia
Graduate Student Research Mentor — *09.2025 – Present*

- Mentored two undergraduate researchers, Jie Ni and Ming Zhu, on research subprojects related to runtime security, automated red-teaming, and risk-aware protection for LLM-based agents.
- Guided the decomposition of the SafeAgent research agenda into independent undergraduate thesis topics, including risk encoding with structured evidence extraction and automated multi-step safety evaluation.
- Provided technical guidance on literature review, system design, experimental methodology, implementation planning, and manuscript preparation.
- Coordinated the integration of undergraduate subproject results into the broader SafeAgent research program.
- Co-authored an arXiv preprint with the mentored students; the extended version is currently under review at an EI-indexed conference.

Publications & Preprints

H. Liu, E. Ilyushin, J. Ni, M. Zhu. *SafeAgent: A Runtime Protection Architecture for Agentic Systems.* arXiv preprint arXiv:2604.17562 [cs.AI], 2026. Under review at the XXVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2026).

Conference Abstracts

E. Ilyushin, **H. Liu**. *Architecture for Runtime Protection of Agentic Systems Against Vulnerabilities and Attacks.* Lomonosov Readings 2026: Scientific Conference, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia, 2026. Conference Abstract (in Russian), pp. 143–144. DOI: 10.29003/m5056.978-5-317-07565-1.

H. Liu. *Anomaly Detection Using Query Modeling and Machine Learning.* International Youth Scientific Forum **Lomonosov-2024**, Lomonosov Moscow State University, Moscow, Russia, 2024. Conference Abstract (in Russian), pp. 28–29. Published in *Materials of the International Youth Scientific Forum "Lomonosov-2024"*. ISBN: 978-5-901-64042-5.

Selected Research & Systems Projects

SafeAgent-Coder: Runtime-Controlled Coding Agent for Tool-Use Safety | *GitHub*

- Built a research prototype of a runtime-controlled coding agent that applies the SafeAgent controller protocol to coding-agent workflows, intercepting agent execution and tool calls before high-impact actions such as file modification, shell command execution, repository cloning, and environment setup are performed.
- Implemented a controller-mediated Vibe Coding application layer with MCP-based tools for file-system operations, terminal command execution, repository handling, Python environment setup, and project inspection, routing tool calls through safety decisions such as allow, block, override, and human-in-the-loop approval.
- Designed the system as a practical artifact for studying runtime safety in LLM-based agents, emphasizing execution-time control, policy-driven decisions, call-budget enforcement, human supervision, and compatibility with SafeAgent-style safety cores without modifying the underlying language model.

SmallLM-Forge: Modular Toolkit for Small Language Model Training and Alignment | *GitHub*

- Developed a modular PyTorch toolkit for small language model experimentation, covering byte-level BPE tokenizer training, decoder-only Transformer pretraining, supervised fine-tuning with PEFT adapters, and DPO-based preference alignment.
- Implemented reusable pretraining components including a GPT-2-style byte-level BPE tokenizer, MiniLlama-style causal language model with RoPE, GQA, SwiGLU, and RMSNorm, masked causal language-modeling loss, validation utilities, and autoregressive text generation.
- Built PEFT and alignment modules with LoRA/DoRA adapter injection, adapter save/load utilities, supervised fine-tuning data pipelines, DPO loss computation, preference-batch collation, reward accuracy/margin tracking, and a pytest suite covering tokenizer, model, training, PEFT, SFT, generation, and alignment functionality.

White-Box Adversarial Robustness Evaluation of Qwen-VL Models

- Designed a white-box robustness evaluation project for modern vision-language models, focusing on how small visual perturbations affect visual grounding, OCR behavior, VQA responses, and safety-relevant multimodal instruction following.
- Formulated adversarial objectives for autoregressive VLM outputs by backpropagating token-level generation losses through the language model, visual-token interface, and vision encoder to construct image-space perturbations.
- Planned comparative evaluation of FGSM, PGD, and Universal Adversarial Perturbations under constrained visual distortion, measuring attack success rate, accuracy degradation, semantic preservation, and perceptual similarity.

Industry Experience

Beijing Waiyan Online Digital Technology Co., Ltd. Wuhan, China

LLM Algorithm Specialist Intern, AIGC / Digital Human Team *07.2025 – 09.2025, Internship*

- Built an end-to-end speech data construction pipeline for multi-speaker emotional TTS, covering multi-source audio collection, automatic denoising, speech segmentation, speaker clustering with voice embeddings and cosine similarity, and emotion/speaker annotation.
- Fine-tuned CosyVoice-based speech generation models with text-audio alignment, Qwen2-based autoregressive speech sequence generation, and LoRA-based speaker adaptation for timbre and emotional expressiveness.
- Designed preference-based emotional adaptation experiments using expressive original speech as positive samples and neutral SFT-generated speech as negative samples, improving emotional expressiveness while preserving speaker identity consistency.
- Developed automation tools for batch multi-speaker data processing, embedding export/loading, LoRA adapter training, dynamic adapter switching during inference, and a web UI for one-click inference and rapid evaluation.

Technical Skills

Research Areas: LLM Security, Agentic AI Systems, Runtime Protection, Prompt Injection Defense, Trustworthy AI, Multimodal Robustness

LLM & Agent Systems: PyTorch, Hugging Face Transformers, PEFT, LoRA, DPO, PPO, LangChain, ReAct Agents, RAG, BM25/Vector Retrieval, vLLM

Security & Evaluation: Adversarial Attacks, Red Teaming, Prompt Injection Evaluation, Agent Security Benchmarks, Robustness Testing

Programming & Tools: Python, C++, CUDA, Git, Linux, Docker, LaTeX, Weights & Biases