

Yufei Luo

◇ Email: luoyf@bupt.edu.cn

◇ Homepage: luoyufeiha.github.io

◇ Google Scholar

◇ GitHub

EDUCATION

M.S. in Cyberspace Security

Sep 2023 – Jun 2026

Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Overall Average: 86.9/100 | Rank: Top 10%

B.S. in Mathematics and Applied Mathematics

Sep 2018 – Jun 2022

Nanchang University, Nanchang, China

RESEARCH INTERESTS

My research focuses on **reliable execution** of LLM agents on long-horizon tasks. Reliability does not emerge automatically as base models grow stronger; it also depends on how execution information is organized and maintained. Guided by this premise, my research centers on the organization and maintenance of **execution state**, exploring how continuously maintained, inspectable, and traceable external structures can represent the goals, constraints, task progress, stage-level judgments, and supporting evidence that inform an agent's next action, so that the agent can decide and act on currently valid information amid ongoing interaction and evolving task state.

My current work, MemSIF, studies the structured organization of interaction histories and fact memory, focusing on the durable preservation and on-demand retrieval of execution information across long-horizon interactions. It provides an empirical starting point for moving from long-term memory organization toward execution state maintenance. Building on this foundation, I plan to further study state-driven verification and recovery—using execution state to detect execution deviations in a timely manner and make recovery decisions—and, longer term, the conditions under which cross-task experience can be reliably reused without negative transfer.

PUBLICATIONS

Manuscript Under Review

[M1] **Y. Luo**, X. Xu, Z. Yang. “MemSIF: From Structured Interactions to Dual-Track Fact Memory for LLM Agents,” arXiv preprint arXiv:2608.01742 (under review at AAAI 2027), 2026. [arXiv] [Code]

First- and Co-First-Author Publications

[P1] Z. Yang*, **Y. Luo***, J. Yang, et al. “Class-Aware Adversarial Unsupervised Domain Adaptation for Linguistic Steganalysis,” *IEEE Transactions on Information Forensics and Security (TIFS)*, 2025. * Equal contribution. [DOI] [Code]

[P2] **Y. Luo**, Z. Yang, X. Xu, et al. “Contrastive Hypersphere for One-Class Linguistic Steganalysis,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 13602–13606. [DOI] [Code]

[P3] **Y. Luo**, Z. Yang, X. Xu, et al. “Clustering-Driven Pseudo-Labeling in Source-Free Domain Adaptation for Linguistic Steganalysis,” in *Proc. ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2025, pp. 153–163. [DOI] [Code]

[P4] **Y. Luo**, Z. Yang, R. Zhang, et al. “Pseudo-Label Based Domain Adaptation for Zero-Shot Text Steganalysis,” in *Proc. International Conference on Communication and Electronics Systems (ICCES)*, 2024, pp. 128–142. [DOI]

Other Publications

[C1] Z. Yang, Y. Wang, **Y. Luo**, et al. “TASDF-Stega: High Capacity Secure Text-Audio Joint Steganography Using Diffusion Latent Space,” *IEEE Signal Processing Letters (SPL)*, 2026. [DOI]

[C2] Q. Niu, Z. Yang, **Y. Luo**, et al. “Domain-Assisted Few-Shot Linguistic Steganalysis in Imbalanced Class Scenarios,” *IEEE Signal Processing Letters (SPL)*, 2025. [DOI]

RESEARCH EXPERIENCE

Long-Term Agent Memory (MemSIF)

Mar 2026 – Jul 2026

Self-led research project, Beijing University of Posts and Telecommunications

- Identified two recurring mismatch patterns in long-term agent memory: Temporal-Structural Misalignment, which fragments cross-time evidence, and Delayed Utility Manifestation, which undermines write-time value estimation; constructed targeted diagnostic subsets to empirically analyze both patterns.
- Proposed MemSIF, an interaction-to-fact framework that organizes interaction histories around local topics and cross-time events using Topical Segments and Event Trajectories, then constructs Dual-Track Fact Memory with CoreFact for stable fact consolidation and ActiveFact for query-driven information activation.
- Validated on the LoCoMo and LongMemEval-S long-horizon conversation memory benchmarks across five LLM backbones including DeepSeek-V4-Pro, consistently outperforming the strongest baselines by 2.29–8.79% in accuracy across all backbone and benchmark settings; ablation studies support the contribution of each component, with a favorable accuracy–efficiency trade-off. [arXiv]

Multimodal & Multi-Turn Dialogue Intent Recognition

Jul 2025 – Jan 2026

Research Intern, Lenovo Research Institute

- Identified modality interference and perceptual misalignment as failure mechanisms in multimodal intent recognition, and built a policy-learning framework over Direct, Drop-Image, and Draw-Box reasoning strategies; with two-stage SFT and GRPO training, Qwen2.5-VL-3B raised accuracy from a 59.02% baseline to 80.90% on MMLA, surpassing the 80.12% of the larger Qwen2.5-VL-7B trained with SFT alone.
- To counter the redundancy and interference of fixed-concatenation dialogue history in multi-turn intent recognition, recast the task as sequential decision-making for Adaptive Context Acquisition, where the model selectively accesses dialogue history on demand and predicts the intent once sufficient information has been acquired; with two-stage SFT and GRPO training, a Qwen2.5-7B-Instruct backbone raised accuracy from a 93.30% baseline to 96.08% on an intent recognition task built from MultiWOZ 2.2.
- In multimodal intent recognition, investigated whether RL gains reflect expanded capability or probability redistribution: pass@k capability-boundary evaluation and layer-wise normalized parameter-change analysis show that SFT induces parameter changes on the order of 10^{-2} , whereas GRPO yields much smaller updates on the order of 10^{-3} ; GRPO training also redistributes success probability from examples with high initial pass@k toward those with low initial pass@k, markedly improving success rates on the harder low-pass@k cases.

Linguistic Steganalysis & Domain Adaptation

Sep 2023 – Jun 2025

Collaborative research with the Lab of New Generation Network Technology and Applications, Tsinghua University

- Focused on detecting distributional differences between LLM-generated steganographic text and natural text; to address class misalignment and class indistinguishability under domain shift, proposed Class-Aware Adversarial Unsupervised Domain Adaptation (CADA), which aligns same-class samples across domains and sharpens inter-class boundaries via adversarial training and progressive pseudo-label self-training; detection accuracy improved by 2.78%, 2.47%, and 0.38% on three steganographic algorithms, approaching 90% on both VLC and AC.
- For the stricter constraint that source data becomes unavailable in deployed cross-domain settings, proposed a Source-Free Domain Adaptation method that learns progressive pseudo-labels driven by clustering structure and semantic-similarity filtering, mitigating pseudo-label noise and prediction collapse to achieve cross-domain detection without access to source data or target labels.
- Extended to extreme low-supervision scenarios: proposed a domain-assisted few-shot detection framework for severe class imbalance; designed a contrastive-learning-based one-class detection method that identifies steganographic text without steganographic training samples; and further built a structure-aware contrastive learning model robust to mixed steganographic algorithms and embedding rates.

AWARDS AND HONORS

Beijing Outstanding Graduate, Beijing Municipal Education Commission	2026
National Scholarship for Graduate Students, Ministry of Education of China	2025
Open Source Security Award (Third Prize), China Cybersecurity Association	2025
Academic Scholarship (First Class), BUPT	2023 & 2024

ACADEMIC SERVICE AND PRESENTATIONS

Peer Review. IEEE ICASSP 2026, IEEE Signal Processing Letters, Aerospace Science and Technology.

Conference Presentations. Oral: ACM IH&MMSec 2025; Poster: IEEE ICASSP 2026, ICCES 2024.

TECHNICAL SKILLS

- **Programming & Deep Learning:** Python, PyTorch, Hugging Face Transformers
- **LLM Training & Inference:** DeepSpeed, OpenRLHF, TRL, ms-swift, vLLM
- **Agent, Memory & Experiment Tools:** LangChain, LangGraph, LlamaIndex, FAISS, ChromaDB, Weights & Biases