Rupeng Zhang 张濡芃

18301003358 | zhangrupeng2023@iscas.ac.cn | Beijing
https://zrp.cool

SUMMARY

I am a Master's student in Software Engineering (2nd year) with proficiency in programming languages, including Python, Java, C++, and JavaScript. I have experience applying NLP models, training large language models, and developing LLM applications such as RAG and Tool Agent. I am familiar with the PyTorch framework and knowledgeable in Computer Organization, Operating Systems, Computer Networks, Artificial Intelligence, and Deep Learning. I am well-versed in standard algorithms and data structures, with expertise in NLP-related technologies.

EDUCATION

University of Chinese Academy of Sciences Sep 2023 - Jun 2026
Software Engineering Master Institute of Software, Chinese Academy of Sciences
GPA: 3.82 / 4.00
Awards: Second-class Scholarship (2024)

Beijing Jiaotong University Sep 2019 - Jun 2023
Software Engineering Bachelor
GPA: 3.62 / 4.00
Awards: Merit Student (2021), First-class Scholarship (2021 - 2023)

PROFESSIONAL EXPERIENCE

Youdao | NetEase, Inc. Apr 2022 - Apr 2023
Developer YoudaoAds
Led the design and implementation of the global overseas short-link system, ytiny.net, based on AWS, enabling fast redirection for users across different regions and accurate exposure data tracking. Achieved sub-200ms redirection times for major cities in Asia, Europe, and the Americas by using DynamoDB for cross-region instance incremental synchronization. Implemented auto-scaling for metrics reporting services with Kubernetes (K8s), supporting over 10,000 QPS. Developed a real-time ETL pipeline with Kafka and Druid to process user-agent and other access data, facilitating efficient data flow and analytics, and successfully supported overseas promotional campaigns for 20+ brands.

Siemens AG May 2025 - Present
AIGC Intern
Led the design and development of an LLM-based code quality review Agent, achieving deep, automated assessments for both full-repository and incremental code by integrating Abstract Syntax Tree (AST) parsing with Retrieval-Augmented Generation (RAG) technology. This approach leverages AST to precisely capture code context and dependencies, while the RAG pipeline provides the LLM with accurate information, effectively addressing the failure of traditional static tools to identify high-level errors. The Agent was successfully integrated into our GitLab CI/CD pipeline to automatically identify non-compliant code and generate modification suggestions, resulting in an approximate 30% increase in code review efficiency and a 20% reduction in related production defects. This tool has now become a standard automated quality gate for the team, significantly enhancing code maintainability and overall development productivity.

RESEARCH EXPERIENCE

From Allies to Adversaries: Manipulating the LLM Tool-Calling through Adversarial Injection Jul 2024 - Oct 2024
Co First Author [NAACL 2025 Main]
We introduce ToolCommander, a framework that exploits vulnerabilities in LLM tool-calling systems, enabling privacy theft, denial-of-service attacks, and business competition manipulation through adversarial tool injection.

Joint-GCG: Unified Gradient-Based Poisoning Attacks on Retrieval-Augmented Generation Systems
Co First Author [IEEE S&P under review]
We introduce Joint-GCG, a new attack framework that significantly improves poisoning attacks on Retrieval-Augmented Generation (RAG) systems. Joint-GCG overcomes the imitations of previous methods by simultaneously optimizing both the retrieval and generation stages using a unified gradient approach with innovations like cross-vocabulary alignment and adaptive weighting. Experiments show that Joint-GCG achieves significantly higher attack success rates than existing methods, highlighting serious security vulnerabilities in RAG systems.

PROJECT EXPERIENCE

R3Gen - RAG Powered Code Completion Assistant for Integration Test Jul 2024 - Oct 2024
R&D Cooperation Project with Huawei
I was responsible for the integrated testing code generation feature of Huawei's internal code completion tool, CodeMate. I designed and implemented the corresponding RAG pipeline, enabling the accurate generation of integrated test scripts based on user queries. This feature is now live at Huawei. By analyzing business feedback on badcases, I innovatively applied LLMs to perform sliding window reordering on retrieval results. This method leveraged the internal knowledge of the large model, significantly enhancing the understanding of retrieved code snippets. The reordering resulted in a 10%+ improvement in the retrieval results' top-1 precision and recall rates.

SKILLS, CERTIFICATIONS & OTHERS

- **Skills:** Code development (Python - Django/FastAPI, Java - SpringBoot, C++, Vue.js), databases (MySQL, ElasticSearch, Redis), middleware (Kafka, RabbitMQ), cloud services (AWS, k8s), data science/machine learning (Pandas, Spark, Scikit-learn, PyTorch), others (Linux, Git, Docker).
- **Languages:** English (fluent), IELTS (Overall Band 8)
- **Activities:** Hands-on Machine Learning for Finance and Python - 2022 (Oxford University Online Term), SegmentFault AIGC Hacker Marathon (GPT-based Automated HR Recruitment SAAS Platform) - 2023