

LEE ZHI CHENG

(+86) 18621669483 | zhichenglee97@gmail.com | Website | LinkedIn

EDUCATION

Tsinghua University , Beijing, China	2023.09 - 2025.06
Master in Advanced Computing, GPA: 3.80 / 4.00	
Shanghai Jiao Tong University , Shanghai, China	2017.09 - 2021.06
Bachelor of Engineering in Automation, GPA: 3.44 / 4.3	

TECHNICAL SKILLS

- **Agent / Harness Engineering:** Claude Agent SDK and pi; multi-agent orchestration, tool integration, session and context management, and execution feedback loops.
- **Agentic Engineering:** AI-assisted development, requirements clarification with grill-me, project instructions in CLAUDE.md, Skills, parallel development with git worktrees, and automated regression testing.
- **AI / LLM Research:** Python; training data curation, supervised fine-tuning (SFT), and model evaluation. Led ReaRAG and DiaKoP research; co-first author of two papers (IJCNN and CIKM Demo).

PROFESSIONAL EXPERIENCE

Z.AI (Zhipu AI) Applied AI Engineer	2025.08 - Present
• Lead the design and delivery of enterprise agent applications for VC investment research, automotive software testing, and multimodal in-car entertainment.	
Z.AI (Zhipu AI) LLM Research Intern	2024.07 - 2025.05
• Agentic RAG: Developed a long-context vs. RAG evaluation framework (GLM-9B-128k), extending HotpotQA and NQ to 50k+ token contexts with added distractors. Built training data and fine-tuned a reasoning-enhanced RAG model for multi-hop QA with reflection and iterative retrieval; co-first author of ReaRAG.	
Wise AI Malaysia AI Engineer	2021.11 - 2023.08
• Led the R&D efforts for the development of a face anti-spoofing model. Achieved compliance with ISO-30107 Level 1 Presentation Attack Detection Standards. Managed the entire project lifecycle, including data collection, model design, training, evaluation, and enhancement.	

AGENT & HARNESS PROJECTS

Venture Buddy | VC Research Agent Platform

Led technical planning and Agent Harness design from initial concept through proof-of-concept (POC) delivery.

- **Application development:** Built a Claude Agent SDK-based investment research platform for a VC fund, including frontend and backend interfaces for multi-turn conversations, tool calls, and research outputs.
- **Tool integration:** Integrated search, knowledge-base retrieval, and read/write access to structured research data through Model Context Protocol (MCP); configured tools and plugins for individual sessions.
- **Research workflows and Skills:** Worked with domain experts to encode sector research and ongoing monitoring methods as Skills and plugins; refined instructions and execution constraints through expert feedback.

ASPICE Coding Agent | Automotive Unit Test Generation

Project lead; built multi-agent workflows on pi with feedback from the client's build environment.

- **Orchestration:** Coordinated code exploration, test generation, and review agents for a Kotlin codebase; provided relevant code, dependency information, testing conventions, and tools.
- **Execution feedback:** Integrated Gradle tests and Kover coverage checks via project CLIs, Skills, and remote execution scripts; fed build and test failures back to agents for correction and revalidation.
- **Context isolation:** Separated implementation and review contexts, using test results to guide revisions. Configured agent teams through pi extensions to share one test-generation workflow across three client scenarios.
- **Outcome:** Increased line coverage from **28.0% to 86.8%** and branch coverage from **22.3% to 71.3%** for the Abacus Kotlin module in an in-vehicle system; secured client acceptance of the POC.

AI Pet Video Generation | Automotive Client

- **Asynchronous architecture:** Built an asynchronous video generation service that animates uploaded pet photos, using Redis queues, workers, and PostgreSQL to handle long-running tasks.
- **Task processing:** Implemented workers that poll generation status, upload completed videos to object storage, and update PostgreSQL and Redis. Added cache-first task queries by task_id to reduce database reads.
- **Multimodal pipeline:** Implemented image-to-video model routing, evaluated models, tuned prompts, and added vision-language model (VLM) checks for generated content.
- **Client acceptance:** Secured client approval of generated video quality and written acceptance of the POC.

RESEARCH EXPERIENCE

ReaRAG | Training Data and SFT for Reasoning-Enhanced RAG

Co-first author; led training data development, SFT, and evaluation.

- **Training data:** Generated reasoning and retrieval trajectories with QwQ-32B-Preview from MuSiQue, HotpotQA, and NQ seed questions; curated approximately **20,000** training examples.
- **Supervised fine-tuning:** Fine-tuned GLM-4-9B on trajectory data mixed with general-purpose SFT data, applying training loss only to reasoning and action tokens to improve multi-step retrieval and question answering.
- **Evaluation and analysis:** Evaluated answer correctness on six QA benchmarks using LLM-as-a-Judge. Conducted ablation studies of reasoning capabilities and analyzed retrieval counts and failure cases to assess trade-offs between answer quality and retrieval efficiency.

PUBLICATIONS

ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation

Jinxin Liu*, **Zhicheng Lee***, Shulin Cao, Jiajie Zhang, Weichuan Liu, Xiaoyin Che, Lei Hou, Juanzi Li
International Joint Conference on Neural Networks (IJCNN), 2026

AtomR: Atomic Operator-Empowered Large Language Models for Heterogeneous Knowledge Reasoning

Amy Xin, Jinxin Liu, Zijun Yao, **Zhicheng Lee**, Shulin Cao, Lei Hou, Juanzi Li
31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Vol. 2, 2025

DiaKoP: Dialogue-based Knowledge-oriented Programming for Neural-symbolic Knowledge Base Question Answering

Zhicheng Lee*, Zhidian Huang*, Zijun Yao, Jinxin Liu, Amy Xin, Lei Hou, Juanzi Li
Conference on Information and Knowledge Management (CIKM Demo), 2024

Deep Semi-Supervised Learning via Dynamic Anchor Graph Embedding Learning

Zihao Wang, Enmei Tu, **Zhicheng Lee**
International Joint Conference on Neural Networks (IJCNN), 2021