Curriculum Vitae

Jiale Zhao

- Location: Beijing, China
- Phone: +86 182-5673-1893
- Email: jialeuuz@gmail.com
- GitHub: github.com/jialeuuz

Professional Summary

I have two years of LLM algorithm internship experience, with strong coding and research skills in LLM and multimodal domains. Initiated ML research in a university lab (Freshman), conducted CV research (Sophomore Fall), transitioned to NLP research (Junior Spring), and gained industry experience as an LLM Algorithm Intern at Li Auto (Junior Year).

Key interests: 1) Agent-based LLMs; 2) Data & self-improving systems; 3) Human–LLM interaction; 4) Interpretability & analysis; 5) Benchmarks & evaluation; 6) Bridging research and real applications.

Experience

LLM Algorithm Intern — Li Auto (Sep 2023 – present · Beijing)

- Data Flywheel for Code LLM: Iterative cycle centered on evaluation (SFT $\rightarrow$ evaluation $\rightarrow$ data generation $\rightarrow$ filtering $\rightarrow$ back to SFT) to mass-produce high-quality training and evaluation data.
- Multi-step Reasoning + Tool Use Agent: Constructed SFT data for LLM Q&A, implemented API function calls, and solved complex reasoning problems through multi-step processes.
- MindGPTo: End-to-end multimodal app inspired by GPT-4o with paralinguistic features; built from scratch with modular FE/BE, large-scale audio data pipelines, and SFT to enhance conversational capabilities.

Education

Chongqing Univ. of Posts and Telecommunications — B.Eng., Algorithm Engineering (2021–2025)

Initiated ML research in a university lab (Freshman), conducted CV research (Sophomore Fall), transitioned to NLP research (Junior Spring), and gained industry experience as an LLM Algorithm Intern at Li Auto (Junior Year).

Publications (Under Review)

- ThinkPilot (under review): Submitted to AACL 2025 via ARR (plan to resubmit to ACL 2025 after revision based on reviews). Contributions: experiment design and implementation; explainability integration; appendix and parts of main text.
- Decoding the Ear (under review): arXiv:2510.20513.
- ExpressiveSpeech Dataset: ~51h bilingual expressive speech (~14k utt.).

Research Directions

Agent-based LLMs — Efficient Complex Problem Solving

Many real-world problems require multi-source reasoning and dynamic user feedback. I focus on: 1) decomposing complex tasks; 2) time-frame mechanisms for updates and reflection; 3) coordinating several parallel smaller models (e.g., multiple 7B LLMs) to improve efficiency/accuracy at the same compute.

Self-Evaluation → Self-Improvement

Three failure modes in code tasks: (1) inconsistently solved problems; (2) problems rarely solved spontaneously but solvable with simple hints; (3) problems needing complex guidance. Address via sampling, heuristics/self-guided strategies, and iterative evolution. Continuous improvement needs diverse fresh data and strong filtering; diverse, accurate test queries are critical.

Human–LLM Interaction — Enabling LLMs to Ask

Enable models to challenge incorrect or contradictory information, detect and request missing conditions for underspecified problems, and proactively seek user assistance on hard tasks.

Interpretability & Analysis — Insight → Control → Safety

Use deeper interpretability to develop more effective control over reasoning. Explore lightweight control modules trained on large response corpora to steer reasoning without modifying LLM weights.

Benchmarks & Evaluation — Making Benchmarks "Dumber"

Transform single-turn benchmarks into simulated multi-turn interactions featuring omissions, errors, and colloquial phrasing; for knowledge tasks, intentionally vague queries requiring elicitation. Better matches real user inputs.

Bridging Research & Application

Keep academic research grounded in real-world needs; pursue a PhD and future faculty path while collaborating with industry to overcome practical constraints like compute.