

Bian, Kai

Email: motary59@gmail.com | Tel: +86 18961774459

Address: HITSZ, University Town of Shenzhen, Nanshan District, Shenzhen, Guangdong 518055 China

EDUCATION

Harbin Institute of Technology, Shenzhen (HITSZ)

09/2023 – 06/2027

- Bachelor of Engineering in Computer Science and Technology

- CGPA: 3.751 /4.0 (89.796 /100)

- Core Courses: *Computer Design and Practice* (100/100), *Introduction to Artificial Intelligence* (100/100), *Software Engineering* (100/100), *Discrete Mathematics* (97/100), *Digital Logic Design* (95/100), *Computer Architecture* (93/100), etc.

RESEARCH INTERESTS

- LLM Vocabulary Optimization; Multi-Agent Systems (MAS); Long-Context Inference; Efficient LLM Inference

PUBLICATION

- Bian, K., the first author, AgentVocab: Structure-Aware Vocabulary Adaptation for Efficient LLM Agents. ICML 2026 (Accepted).

- Introduced AgentVocab, a structure-aware vocabulary adaptation framework for efficient large language model (LLM) agents that derives specialized vocabulary entries from real tool-calling traces and adapts the model vocabulary to better reflect structural and semantic regularities without task-specific schema engineering, thereby addressing inefficiencies caused by mismatched tokenization in tool-calling scenarios;

- Demonstrated an approximately 15-25% reduction in decoding latency relative to the vanilla baseline while preserving tool-calling performance;

- The approach is orthogonal to existing fine-tuning and agent-training methods and integrates seamlessly into standard agent pipelines. Source code and models will be available at <https://github.com/Starry-159/AgentVocab>.

PATENT

- Bian, K., Cheng, N., and Li, J., Method and System for Resource-Efficient and Trustworthy Text Analysis Based on Large Language Models. CN Patent, Application No. 202610513322.9, Filed 2026 (Pending).

RESEARCH EXPERIENCE

- **Research Assistant, Research on LLM Vocabulary Optimization Mechanisms for Single-Agents** 09/2025 - Present

Advisor: Prof. Xuebo Liu, Harbin Institute of Technology, Shenzhen

- Proposed a tokenization/vocabulary-based optimization method by reconstructing the vocabulary based on action spaces and instruction characteristics to tackle the high computational cost and limited inference speed of large language model (LLM) agents in long-context and multi-turn interactions at the vocabulary level;

- Independently conducted literature review, algorithm design, code implementation, as well as in-depth analysis of the token consumption and latency bottlenecks of existing LLMs when implementing agent tasks;

- Designed and implemented a customized vocabulary optimization strategy for single-agent interaction frameworks, built an evaluation pipeline, and conducted comprehensive ablation experiments to quantitatively analyze the accuracy, inference latency, and resource consumption before and after optimization;

- Developed a prototype algorithm for task-specific vocabulary optimization and verified through preliminary experiments that the method significantly reduces token consumption and inference latency while maintaining or improving the accuracy in task implementation compared with baseline models.

PROJECT EXPERIENCE

- **Technical Leader, Design and Implementation of a Resource-Efficient and Trustworthy Text Analysis System Based on LLMs (National College Student Innovation and Entrepreneurship Training Program)** 05/2025 – 05/2026

Advisor: Prof. Jingjing Li, Harbin Institute of Technology, Shenzhen

 - Designed a four-stage LLM processing pipeline (data loading, dynamic binning, cloud inference, and three-stage analysis) to address high API cost and hallucination in large-scale financial text analysis (10k+ M&A announcements);
 - Achieved fully asynchronous processing with Alibaba Cloud Model Studio Batch API, conducted full-scale evaluation and cost/accuracy ablation studies on 11.6k real-world data, and reduced total processing time from 33.8 hours to 0.5 hours, improving efficiency by 68 times;
 - Developed a token-budget-aware dynamic binning algorithm with Python, using the tiktoken library for precise token estimation and packing fragmented texts through greedy strategy, thereby reducing token redundancy and lowering per-request cost by over 30% (to RMB 0.004/request);
 - Designed a fault-tolerant strategy based on Abstract Syntax Trees (AST, demjson3) to address parsing failures caused by JSON format inconsistencies from LLM outputs and maximize effective payload utilization;
 - Designed a three-stage anti-hallucination mechanism (screening-review-adjudication), integrated multi-path reasoning, chain-of-thought (CoT), and self-consistency, and achieved an accuracy of 99.74%;
 - Completed a technical paper as the first author and filed a national invention patent as the first inventor.
- **Project Leader, A Multi-Agent System for Automated Digital Human Video Generation (AI+ Track of the “Challenge Cup” National College Student Extracurricular Academic and Technological Works Competition)** 05/2025 – 09/2025

Advisor: Prof. Xuebo Liu, Harbin Institute of Technology, Shenzhen

 - Designed an MAS collaboration framework with Qwen as the central controller and constructed five sub-agents for task planning, text generation, text-to-speech (TTS), visual rendering, and composition to enable dynamic decomposition of complex video generation tasks;
 - Developed a graph-based context-aware agent memory module with CoT reasoning to ensure logical and cultural consistency of digital human identity, voice style, and background elements in cross-modal generation;
 - Constructed a domain-specific cultural tourism knowledge base to address cross-modal inconsistencies and improve feature alignment;
 - Integrated GPT-SoVITS, diffusion models, and FFmpeg into a highly modular Web-based pipeline for text-image-audio-video generation;
 - Achieved fully autonomous zero-shot video generation, reducing the production cycle from 7-10 days to within 1 hour, and improving efficiency by over 10 times.
- **Project Leader, An End-to-End LLM-Driven Automated Digital Human Generation System** 03/2025 – 07/2025

Advisor: Prof. Xuebo Liu, Harbin Institute of Technology, Shenzhen

 - Designed a text-speech-visual-subtitle-music generation architecture integrating Qwen, GPT-SoVITS, and SadTalker/MuseTalk through a microservices framework;
 - Developed a highly interactive, user-friendly Web UI, enabling database management (model/asset import), dynamic node configuration, and one-click end-to-end generation;
 - Implemented FFmpeg-based audio-visual synchronization for long-video rendering;
 - Optimized system inference efficiency through model quantization and asynchronous processing, achieving near-linear scaling that generates a 30-second video within 5 minutes;
 - Reduced the video production cycle from 3-5 days to within 30 minutes, improving efficiency by over 10

times and reducing the cost per video to less than RMB 10;

- Proposed a node-intervenable agentic workflow that enables controllable and scalable generation.

- **Project Leader, AI Anchor (Freshman-Year Research and Innovation Project)** 12/2023 – 12/2024

Advisor: Prof. Xuebo Liu, Harbin Institute of Technology, Shenzhen

- Developed an end-to-end AI anchor system that generates high-definition videos with two AI anchors delivering content alternately based on long-form text inputs and two user-selected roles;

- Designed long-text segmentation and multi-role alternating allocation algorithms to improve coherence and system stability in generation through LLM-based semantic segmentation and task decomposition;

- Led system architecture design, model selection, local deployment, and API integration to implement a fully automated multimodal pipeline;

- Evaluated semantic coherence, speech stability, and lip sync quality and developed iteration solutions.

AWARDS & HONORS

- Second Prize of Guangdong Province in China Undergraduate Mathematical Contest in Modeling 10/2025
- Second Prize of HITSZ in National Collegiate Statistical Modeling Contest 05/2025
- Third Prize and Excellent Award of HITSZ in “Challenge Cup” National College Student Extracurricular Academic and Technological Works Competition 03/2025

SKILLS

- **Languages:** Mandarin (native), English (fluent)
- **Technical Skills:**
 - **Programming:** Python; C/C++; Java; SQL; Shell/Bash
 - **Deep Learning & LLMs:** PyTorch; Hugging Face (Transformers, Tokenizers, Datasets) for LLM deployment, fine-tuning, and evaluation; LLM computation mechanisms and memory optimization techniques
 - **LLM Agents & MAS:** Agent framework development and deployment (e.g., OpenClaw); LLM API integration; MAS architectures and interaction design
 - **Systems & Tools:** Linux/Ubuntu; Git; Docker; NumPy; Pandas