Yinting (James) Chen

2177786565 • yintingchen762@gmail.com • cyt1ng.github.io • linkedin.com/in/yintingchen/

EDUCATION

University of Illinois Urbana-Champaign, Champaign, IL *Aug 2025 - Jan 2028 (Expected)*
Bachelor of Science in Computer Science *GPA: 4.0/4.0*

TECHINICAL SKILLS

- **Programming Languages:** Python, C++, Java
- **AI/LLM:** Distributed System, Agent Skills, Agent Harness, Agent Memory, RAG, Prompt Engineering
- **Tools/Frameworks:** Hermes Agent, Claude Code, OpenClaw, LangChain, Vector Databases, Git, Docker

WORKING EXPERIENCE

Software Development Intern, **Ooin Media, New York, NY** *May 2026 - Jul 2026*

- Designed and implemented an agentic voice-of-customer (VoC) analyzer that aggregates user reviews across X, Instagram, TikTok, Reddit, and YouTube and synthesizes them into structured reports.
- Engineered an iterative agent loop driving an end-to-end pipeline (input → search → scrape → integrate → analyze → report), with DDGS handling scraping and OpenRouter managing cross-model LLM routing.
- Utilized NLP-driven sentiment analysis to extract product insights from unstructured, cross-platform user reviews.
- Built a frontend that visualizes the system's workflow and converts text-only report data into clear, intuitive graphs.

Research Intern at U Lab, UIUC (Advised by Prof. Jiaxuan You) *Dec 2025 - Present*

- Paper currently under double-blinded review at EMNLP 2026.
- Designed a GUI-agent error taxonomy with 4 major categories and 29 subtypes, covering perception, interaction localization, task reasoning, and external system failures across mainstream OSWorld/CUA.
- Built a self-evolving VLM-agent debugging framework that automatically identifies root causes from failed trajectories, distills them into reusable debugging skills, and enables agents to learn autonomously from historical failures.
- Designed a dual-layer memory architecture with episodic and semantic memory, together with an intent-aware RAG retrieval mechanism that injects skills from similar-intent trajectories to improve diagnosis and re-rollout accuracy.

Software Development Intern, **BytesBoost, San Francisco, CA** *Sept 2025 - Nov 2025*

- Developed and maintained React/Next.js front-end interfaces for flowmuse.ai, a visual AI workflow platform for multimodal content generation.
- Built a canvas-style prompt-to-video/photo workspace, streamlining prompt iteration and user-side optimization to improve UX and interaction quality across LLM features.
- Integrated model and media-generation APIs (e.g., Sora 2, Nano Banana 2, Veo 3.1) into the product workflow.

PROJECTS HIGHLIGHTS

Multi-Agent LLM System *Oct 2025 - Nov 2025*

- Developed autonomous research agent using LangChain for web scraping, information synthesis, and knowledge graph construction.
- Built multi-agent marketing strategy system with CrewAI, demonstrating practical LLM applications in business automation.
- Implemented agent orchestration and task delegation for complex, multi-step reasoning workflows with chain-of-thought prompting.

Retrieval-Augmented Generation (RAG) System *Aug 2025 - Sept 2025*

- Architected end-to-end RAG pipeline for document-based QA using GPT-4o, comparing against Llama and Mistral models.
- Implemented advanced retrieval techniques including HyDE (Hypothetical Document Embeddings), improving context relevance by 15%.
- Optimized chunking strategies and embedding generation for 10K+ document corpus using semantic similarity algorithms.
- Enhanced prompt engineering techniques to reduce hallucinations and improve factual accuracy through systematic evaluation.