

Kehao Chen

18177110580 | kkehaochen@gmail.com | Nanning

[Personal Website](#) | [LinkedIn](#) | [GitHub](#)

AI Application Engineer | RAG · Multimodal Document Intelligence · LLM Evaluation and Workflows

EDUCATION

The University of Sydney | Master of Data Science · Full-time

Sydney | Jul 2022 - Dec 2023

Beijing University of Posts and Telecommunications | Bachelor's Degree in Telecommunications Engineering and Management · Full-time

Beijing | Sep 2017 - Jun 2021

TECHNICAL SKILLS

- **LLM Applications:** LLMs, VLMs, RAG, prompt engineering, multi-agent workflows, structured output, SFT, LoRA, LLM-as-a-Judge, offline evaluation.
- **Models and Retrieval:** QwenVL, Qwen2.5, OneChart, LayoutLM, BERT, CLIP, DINOv2; vector retrieval, HyDE, Multi-Query.
- **Development and Data:** Python, SQL, R, FastAPI, PyTorch, TensorFlow, Pandas, Scikit-learn, Docker, Linux, Git, SwiftUI, Supabase, Power BI, Tableau.
- **Languages:** CET-6; IELTS 6.5; able to read and write English technical documentation and communicate in English in daily settings.

WORK EXPERIENCE

Matrix Origin | Algorithm Engineer · AI Platform

Shanghai | Sep 2024 - Jan 2026

Multimodal Document Intelligence and RAG

- Contributed to migrating a document parsing system from a multi-model pipeline combining LayoutLM, OCR, and table parsing to an end-to-end VLM solution based on QwenVL, reducing pipeline maintenance costs and improving parsing stability for complex-layout documents.
- Built fine-tuning datasets, designed structured output schemas, and established an offline evaluation framework; achieved 94.6% text recognition accuracy measured by 1-Edit Distance and improved table parsing TEDS from 74% to 87%, enhancing downstream RAG performance on complex documents.
- Contributed to an internal enterprise knowledge-base question-answering system, handling Markdown document chunking, vectorization and indexing, and retrieval strategy optimization; compared baseline vector retrieval, HyDE, and Multi-Query, and used an offline QA set to evaluate retrieved passages, generated answers, and response latency.

Multi-Agent Medical Consultation Data Generation and Model Evaluation

- Built a persona-based multi-agent data generation framework that used medical case logic to simulate multi-turn doctor-patient interactions, synthesizing 20k+ multi-turn consultation records for SFT.
- Introduced LLM-as-a-Judge to build an automated data-cleaning pipeline and fine-tuned Qwen2.5-7B with LoRA, improving the model's BERTScore from 65% to 88%.

Wind Information | Algorithm Engineer · Intelligent Applications Department

Shanghai | Jan 2024 - Aug 2024

Financial Chart Understanding and Structured Extraction

- Led domain-specific SFT for OneChart to extract structured JSON from complex financial charts, maintaining chart parsing precision above 95% and reducing hallucinations in numerical extraction from unstructured charts.

- Designed a chart-specific annotation tool and semi-automated data production pipeline to reduce manual labeling costs; combined VLM and structured extraction capabilities to build a chart question-answering system for financial research reports, supporting numerical reasoning over charts.

Intelligent Video and Documents

- Designed a video keyframe extraction algorithm that combined frame differencing and edge detection to identify PPT transition points in meeting videos; fine-tuned LayoutLMv3 to locate PPT titles across different templates, achieving an F1 score of 98%.

INDEPENDENT PROJECT

PawMood | AI Pet Meme Generator iOS App

Independent Developer | [Product Website](#)

- Independently completed full-stack development of the SwiftUI iOS client and FastAPI backend and designed the end-to-end product architecture; used a VLM to understand pet poses and visual context, generating English meme captions and shareable cards under style and tone constraints.
- Implemented core flows including photo upload, mood and style selection, caption generation and regeneration, local saving, and sharing, while continuously improving the mobile interaction and generation experience.
- Completed user login, subscription payments, Supabase data synchronization, and FastAPI cloud deployment, connecting the client, backend services, and monetization capabilities into a complete workflow.

RESEARCH EXPERIENCE

Multimodal Sentiment Analysis | The University of Sydney

Jul 2023 - Dec 2023

- The paper [Enhancing Sentiment Analysis Through Multimodal Fusion: A BERT-DINOv2 Approach](#) was accepted by ICCS 2025 (International Conference on Computational Science, CCF-C).
- Proposed a BERT + DINOv2 dual-tower multimodal fusion architecture and used Cross-Attention to improve text-image feature alignment; conducted ablation experiments on three public datasets and used SMOTE and Focal Loss to address class imbalance.