

OCR-Based Student Verification and Local LLM Assistant for Indonesian Students in Taiwan

專題編號：114-2-CSIE-S018

執行期限：114 年第 1 學期至 114 年第 2 學期

指導教授：陳香君

專題參與人員：112590036 李岱磐

112590040 施孝健

112590041 鐘美萍

一、摘要

This project develops PengYou, a web-based platform integrating PaddleOCR and a locally deployed LLaMA model, to address administrative challenges faced by Indonesian Student Associations in Taiwan. Using RAG sourced from the PERPITA guidebook [3], features include automated ARC verification, AI-driven student guidance, an announcement board, and an admin management panel.

Keywords: OCR, LLM, ARC Verification, RAG, PaddleOCR, SQLite, Privacy-Preserving AI

二、緣由與目的

The growth of Indonesian students in Taiwan has created bottlenecks in manual ARC verification and information distribution, worsened by language barriers and privacy risks from cloud-based AI. This project aims to:

1. Automate ARC verification using PaddleOCR-based document processing [1].
2. Provide context-aware student guidance via a local LLM with RAG [4][5][6].
3. Give administrators a centralized platform for student management and announcements.

三、研究範圍

This project focuses on three main components:

(a) OCR-Based Document Verification

The system automatically extracts and validates key fields from ARC documents, supporting both Traditional Chinese and English text recognition with image preprocessing to improve accuracy.

(b) Local LLM Guidance Assistant

The system provides context-aware student support through a locally deployed language model integrated with a retrieval-based knowledge source, ensuring factual accuracy and data privacy compliance.

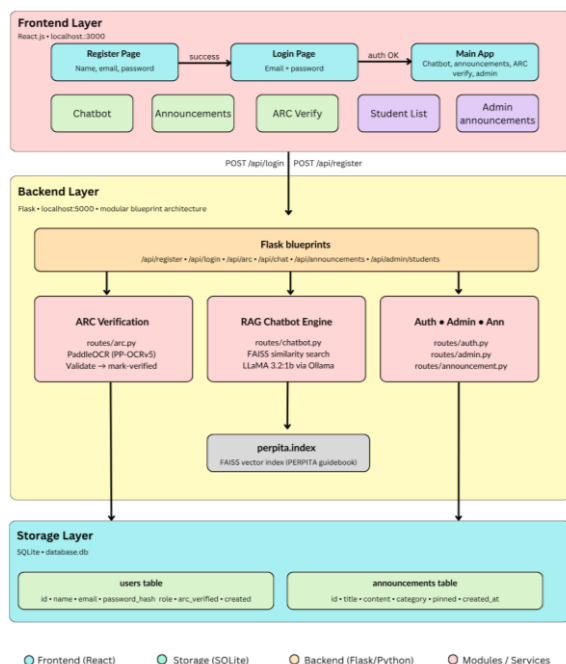
(c) Student Management and Announcements

Administrators can view registered students and their ARC verification status, and publish categorized announcements to students via a dedicated board.

四、架構流程

The system uses a three-layer architecture (Frontend: React.js, Backend: Flask, Storage: SQLite). Students register with name, email, and password, then log in to access the platform. ARC verification can be submitted at any time after login. The backend is implemented using modular Flask blueprints, with dedicated endpoints for authentication,

ARC verification, chatbot, announcements, and student management.



五、使用技術方法

(a) Optical Character Recognition (OCR)

Image preprocessing using OpenCV [2] is applied before text extraction using PaddleOCR (PP-OCRv5 server pipeline, lang="ch") [1]. The OCR engine simultaneously recognizes Traditional Chinese and English text found on ARC documents. Extracted text is structured using regular expression pattern matching to identify key fields: Name, Nationality, Status, School, and Expiry Date.

(b) Large Language Model (LLM)

A locally deployed LLaMA 3.2 1B model running via Ollama [5] serves as the core language generation engine. A FAISS vector index built from the PERPITA Student Guidebook 2025 [3] is used for Retrieval-Augmented Generation (RAG) [6]. The system encodes queries using the paraphrase-multilingual-MiniLM-L12-v2 embedding model and retrieves the most relevant chunks before passing them to the LLM, ensuring

domain-restricted and factually grounded responses.

六、實驗結果

PaddleOCR achieves high field extraction accuracy on ARC documents under standard conditions. The decoupled registration flow reduces onboarding friction for newly arrived students without an ARC. The local LLM delivers consistent, domain-restricted responses from the PERPITA guidebook without data leakage. The admin panel reflects ARC verification status in real time.

七、結論

PengYou demonstrates that PaddleOCR and a local LLM can effectively automate Indonesian student onboarding in Taiwan while maintaining data privacy. The flexible registration flow, announcement board, and admin panel reduce administrative workload. The modular architecture allows straightforward feature expansion.

八、參考文獻

- [1] Baidu PaddlePaddle, "PaddleOCR: Awesome Multilingual OCR Toolkits," GitHub. [Online]. Available: <https://github.com/PaddlePaddle/PaddleOCR>
- [2] OpenCV Team, "OpenCV Documentation," [Online]. Available: <https://docs.opencv.org>
- [3] PERPITA, *Student Essential Guidebook 2025*. Taiwan: PERPITA, 2025.
- [4] Meta AI, "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.
- [5] Ollama, "Ollama Documentation," [Online]. Available: <https://ollama.com>
- [6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proc. NeurIPS*, 2020.