From OCR to Multimodal LLMs

Building a Textbook Corpus with ALICE

Steven Denney
Leiden University

ALICE-SHARK User Meeting 2025
June 3, 2025

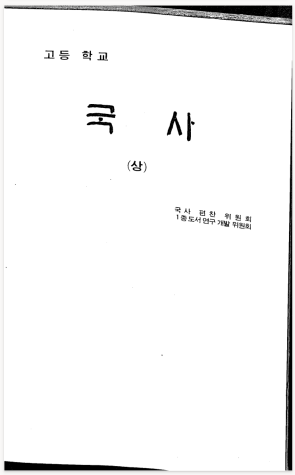
Why This Project?

- Textbooks are key to understanding civic education, national identity, and state discourse, especially over long periods of time.
- OCR quality on historical text in non-Latin scripts (e.g., Korean) is often poor, especially when using English-centric tools.
- By working with Korean text and a Korean-trained LLM (EXAONE), this project demonstrates the viability of non-English, script-specific LLM pipelines.
- ALICE enables local, high-performance language model inference – ideal for multilingual document workflows.

Corpus Construction Goals

- Build a clean, searchable corpus from scanned Korean textbooks.
- Compare and evaluate two workflows:
  1. Pipeline A: Traditional OCR + LLM cleaning
  2. Pipeline B: Vision-language model + LLM post-editing
- Optimize for ALICE: GPU-accelerated, fully local, efficient.

What We're Working With Here



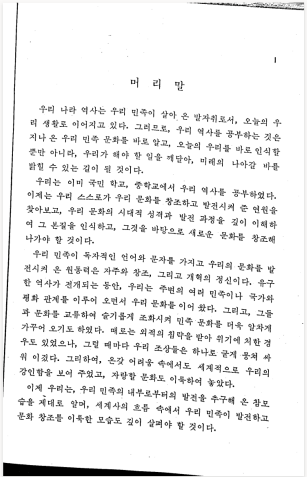
고등 학교

국 사

(상)

국사 편찬 위원회
1종도서연구개발 위원회

Cover page



머 리 말

우리 나라 역사는 우리 민족이 살아 온 발자취로서, 오늘의 우리 생활로 이어지고 있다. 그러므로, 우리 역사를 공부하는 것은 지나 온 우리 민족 문화를 바로 알고, 오늘의 우리를 바로 인식할 뿐만 아니라, 우리가 해야 할 일을 깨닫아, 미래의 나아갈 바를 밝힐 수 있는 길이 될 것이다.

우리는 이미 국민 학교, 중학교에서 우리 역사를 공부하였다. 이제는 우리 스스로가 우리 문화를 창조하고 발전시켜 온 연원을 찾아보고, 우리 문화의 시대적 성격과 발전 과정을 깊이 이해하여 그 문점을 인식하고, 그것을 바탕으로 새로운 문화를 창조해 나가야 할 것이다.

우리 민족이 독자적인 언어와 문자를 가지고 우리의 문화를 발전시켜 온 힘들력은 자주와 창조, 그리고 개혁의 정신이다. 우구한 역사가 전개되는 동안, 우리는 주변의 여러 민족이나 국가와 평화 관계를 이루어 오면서 우리 문화를 이어 왔다. 그리고, 그들과 문화를 교류하여 슬기롭게 조화시켜 민족 문화를 더욱 알차게 가꾸어 오기도 하였다. 때로는 외적의 침략을 받아 뒤기에 처한 경우도 있었으나, 그럴 때마다 우리 조상들은 하나로 굳게 뭉쳐 싸워 이겼다. 그리하여, 온갖 어려움 속에서도 세계적으로 우리의 강인함을 보여 주었고, 자랑할 문화도 이룩하여 놓았다.

이제 우리는, 우리 민족의 내부로부터의 발전을 추구해 온 참모습을 제대로 알며, 세계사의 흐름 속에서 우리 민족이 발전하고 문화 창조를 이룩한 모습도 깊이 살펴야 할 것이다.

Introduction page

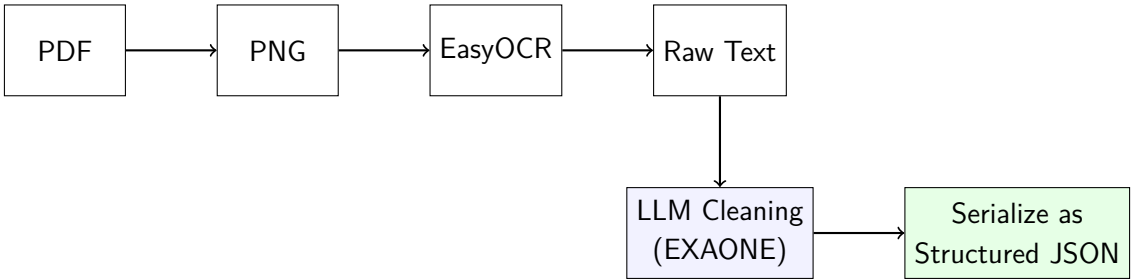
Language, Modularity, and Scale

- This pipeline is prototyped on Korean text — ideal for testing non-Latin scripts.
- EXAONE is used because it is trained on Korean — but other LLMs could be substituted.
- Pipeline is modular: Vision model, LLM, and post-processing components are interchangeable.
- Goal: develop a scalable, multilingual corpus-building framework for diverse scripts.

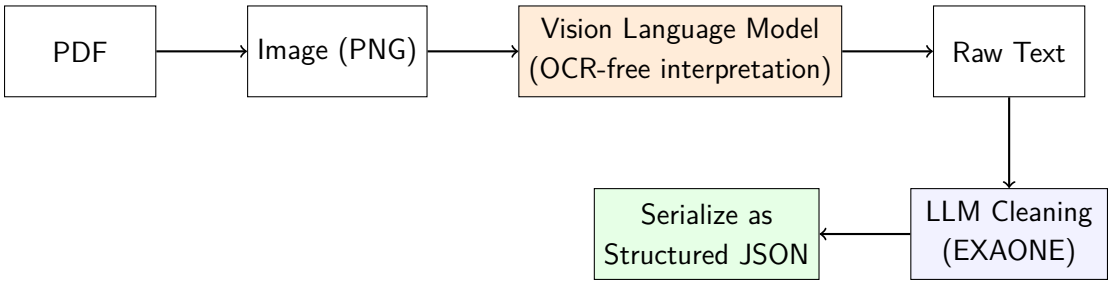
Pipeline Typical (Legacy Approach)

- The traditional pipeline relies entirely on OCR:
  1. PDF $\rightarrow$ Image conversion
  2. Image $\rightarrow$ OCR engine (e.g., Tesseract or EasyOCR)
  3. Output stored as raw or segmented plain text
  4. Optionally post-processed and structured manually or semi-manually into JSON or XML
- Works reasonably well for clean Latin-script sources, but performs poorly on layout-rich or non-Latin texts.

Pipeline A: OCR-Based Workflow



Pipeline B: Prototype Workflow (Vision Language Model)



Model Infrastructure: Ideal and Tested

Ideal setup for scalable, high-quality inference:

- EXAONE-Deep 70B (Q5_K_M) – Korean-specific LLM used for post-processing and normalization. High performance; fits on ALICE GPU nodes via quantization.
- LLaVA-1.5 + Mistral 7B – Open-source multimodal model for direct image-to-text interpretation. Supports OCR-free pipeline design, and runs efficiently with quantized weights on ALICE GPUs.

Tested alternatives via local prototyping:

- EXAONE 7B – Used for early-stage post-cleaning; faster but with slightly reduced performance.
- MiniGPT-4, LLaVA-1.5 – Lightweight VLMs tested for proof-of-concept inference. Run on consumer GPU, but slow(er) and less reliable than ideal models.

Comparing the Pipelines

Feature	**Pipeline A (OCR)**	**Pipeline B (VLM)**
Accuracy	Moderate, layout lost	Promising, preserves layout
Speed	Stable but multi-step	Faster per step, but newer
Robustness	OCR is (semi-)reliable	VLMs still evolving
Deployment	Simple batch processing	Needs better fallback/error handling

Pipeline B: Built for ALICE?

- Entire pipeline executes on ALICE — no cloud dependencies, full data control.
- Uses quantized models (e.g., EXAONE 70B Q5_K_M) for efficient GPU inference on single nodes.
- Vision-Language inference and LLM post-processing both benefit from ALICE's parallel architecture.
- Enables testing, tuning, and benchmarking of large multilingual models in-house.
- Future: add automated fallback from VLM to OCR — and batch-parallel scaling across textbook collections.

Next Steps and Broader Considerations

- Fine-tune VLM interpretation on Korean textbook samples.
- Release annotated corpus + pipeline documentation for other researchers/socializing.
- Assess tradeoffs: Is multimodal extraction truly necessary? Does it justify the extra compute and complexity?

Thank You

Project materials and code (work in progress):
https://github.com/scdenney/textbook-pipeline/tree/main

Contact: s.c.denney@hum.leidenuniv.nl