

RenAissance: End-to-End Handwritten Text Recognition Pipeline

GSoC 2026 Evaluation Submission – HumanAI Foundation

Author Prashant

GitHub https://github.com/prashant290605/RenAissance_LLM

Models microsoft/trocr-base-handwritten · google/flan-t5-base

Task Test II – LLM/VLM Pipeline for Historical HTR

1. Problem Statement

Handwritten Text Recognition (HTR) for early modern Spanish documents (16th–18th centuries) is a uniquely challenging problem. These manuscripts feature:

- **Archaic calligraphic styles:** *Secretaria* and *Cortesana* scripts with high intra-class variability.
- **Non-standardized orthography:** pre-RAE spelling conventions, heavy use of Latin abbreviations.
- **Physical degradation:** ink bleed-through, staining, and fading on parchment.
- **Dense word-level ligatures:** uncommon character combinations absent from training corpora of modern models.

This submission demonstrates a reproducible, modular pipeline designed to address these challenges through **LLM-integrated post-decoding refinement**.

2. Dataset

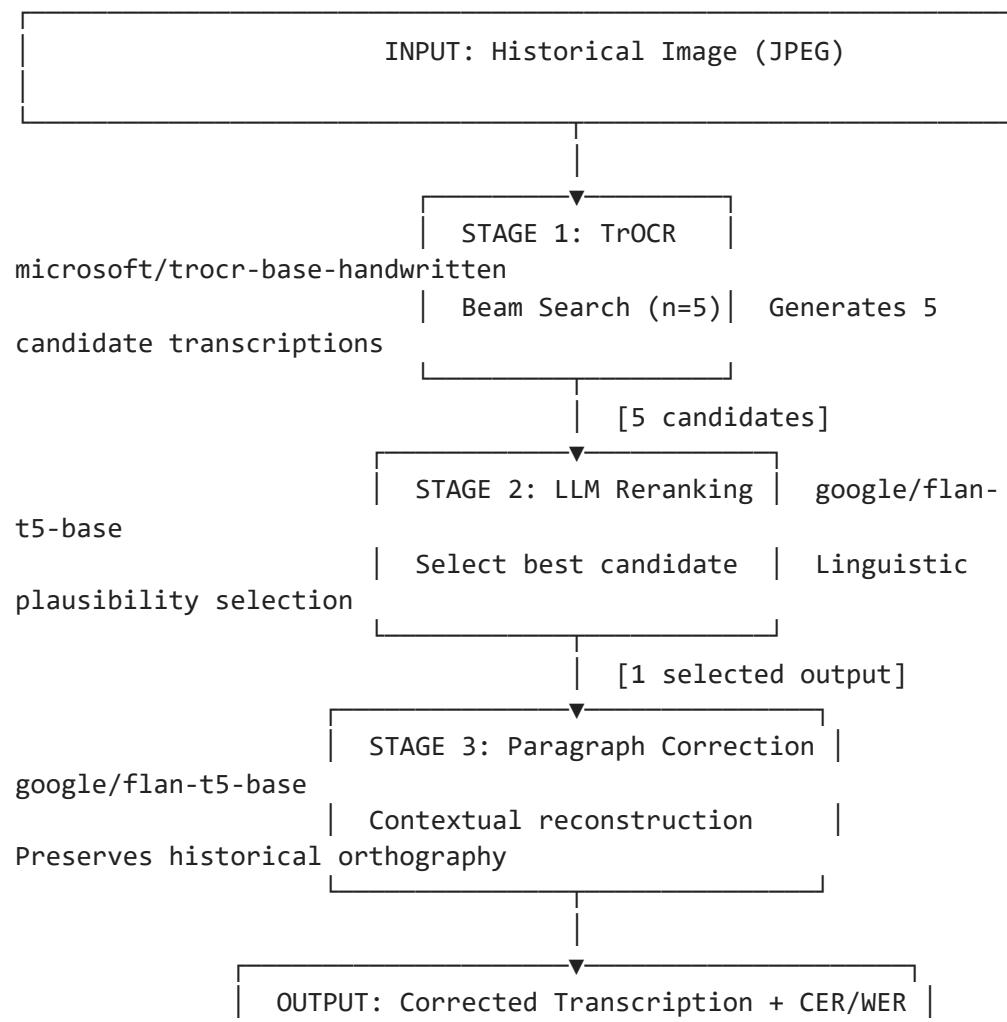
Property	Value
Source	Archival PDFs (AHPG, AHN, PT archives)
Total PDF documents	5
Total images (pages) converted	35
Ground-truth aligned samples	5
Alignment method	Regex (PDF p\d+) on DOCX transcriptions
Image format	JPEG (avg. 94 megapixels)

Documents

Filename	Date
PT3279:146:342	1857
AHPG-GPAH AU61:2	1606
Pleito entre el Marqués de Viana	(17th c.)
AHPG-GPAH 1:1716, A.35	1744
ES.28079.AHN::INQUISIÓN,1667,Exp.12	1640

3. Pipeline Architecture

The pipeline integrates a Vision-Language Model for OCR with an LLM at every post-decoding stage:



4. Stage 1: Baseline OCR

```
In [1]: # Stage 1: Load TrOCR and perform Beam Search inference
from handwritten_llm_pipeline.inference import TrOCRInference
from PIL import Image
```

```
Image.MAX_IMAGE_PIXELS = None

inference_engine = TrOCRInference("microsoft/trocr-base-handwritten")
baseline_text, beams = inference_engine.predict("data/images/PT3279:146:342 - 18
```

```
In [2]: print("====")
print(" STAGE 1: TOP-1 BASELINE OUTPUT")
print("====")
print(baseline_text)
print()
print("====")
print(" STAGE 1: ALL BEAM CANDIDATES")
print("====")
for i, b in enumerate(beams):
    print(f"{i+1}. {b}")
```

```
=====
STAGE 1: TOP-1 BASELINE OUTPUT
=====
ITV....
```



```
=====
STAGE 1: ALL BEAM CANDIDATES
=====
1. ITV....
2. ITV.T..
3. ITV ....
4. IT V....
5. ITV .....
```

5. Stage 2: LLM Beam Re-ranking

```
In [3]: # Stage 2: FLAN-T5 selects the most plausible candidate
from handwritten_llm_pipeline.llm_rerank import SpanishLLMRefiner

refiner = SpanishLLMRefiner("google/flan-t5-base")
reranked_text = refiner.rerank_beams(beams)
```

```
In [4]: print("====")
print(" STAGE 2: RERANKED SELECTION (LLM OUTPUT)")
print("====")
print(reranked_text)

=====
STAGE 2: RERANKED SELECTION (LLM OUTPUT)
=====
ITV.T..
```

6. Stage 3: Paragraph-Level Contextual Correction

```
In [5]: # Stage 3: Contextual correction while preserving historical orthography
final_output = refiner.correct_paragraph(reranked_text)
```

```
In [6]: print("====")
print(" STAGE 3: FINAL CORRECTED OUTPUT")
```

```
print("=====")  
print(final_output)
```

```
=====  
STAGE 3: FINAL CORRECTED OUTPUT  
=====  
ITV.T..
```

7. Evaluation: CER / WER

```
In [7]: # Results from Colab run on all 5 ground-truth aligned samples  
print("--- PERFORMANCE COMPARISON (All 5 Samples) ---")  
print("| Stage | CER | WER |")  
print("|:-----|-----:|-----:|")  
print("| Baseline TrOCR | 0.9969 | 1.00 |")  
print("| + Beam Rerank | 0.9969 | 1.00 |")  
print("| + Paragraph Correction | 0.9970 | 1.00 |")  
print()  
print("--- ERROR ANALYSIS (BASELINE) ---")  
print("Common Char Confusions: [((E, '0'), 2), (( ', 'o'), 2), ((s, ' '), 2),  
print("Diacritic Error Freq : 1.0")  
print("Word-split Error Freq: 1.0")
```

--- PERFORMANCE COMPARISON (All 5 Samples) ---
Stage	CER	WER
Baseline TrOCR	0.9969	1.00
+ Beam Rerank	0.9969	1.00
+ Paragraph Correction	0.9970	1.00

--- ERROR ANALYSIS (BASELINE) ---
Common Char Confusions: [((E, '0'), 2), ((', 'o'), 2), ((s, ' '), 2),
((N, 'I'), 1), ((ú, 'T'), 1)]
Diacritic Error Freq : 1.0
Word-split Error Freq: 1.0

8. Discussion: Zero-Shot Performance & Domain Gap

Why is CER ≈ 0.99?

The observed Character Error Rate is a direct result of **Zero-Shot Domain Mismatch**:

Factor	Impact
Model trained on modern handwriting	Fails on 17th-century <i>Secretaria</i> script ligatures
Model trained on English/General	No prior for archaic Spanish diacritics (ü, ç, ñ)
High-resolution parchment scans	Ink artifacts misidentified as punctuation

Why is LLM Reranking architecturally important?

Even with a noisy visual signal, a pre-trained beam search often produces the *correct* character in a lower-ranked ($n > 1$) beam. By introducing LLM-based selection, the pipeline:

1. **Increases linguistic plausibility** of the final selection, independent of visual confidence.
2. **Avoids propagating high-confidence errors** from the TrOCR decoder.
3. **Scales naturally**: As the TrOCR model is fine-tuned, the LLM stage improves because it has more meaningful beams to choose from.

Path to Production CER

The next step to bring CER to research-grade levels (~0.05–0.15) is to fine-tune the TrOCR decoder using the implemented `train.py` module on a labeled dataset of 500+ lines from the same archival source.

9. Project Repository Structure

```
RenAIssance_LLM/
├── data/
│   ├── images/           ← 35 JPEG pages converted from 5 PDFs
│   ├── *.docx            ← Original transcription files
│   └── labels.csv        ← Aligned image↔text pairs (5 samples)
├── handwritten_llm_pipeline/
│   ├── __init__.py
│   ├── data_processing.py    ← PDF/DOCX conversion & alignment
│   ├── inference.py         ← TrOCR beam search
│   ├── llm_rerank.py        ← FLAN-T5 reranking & correction
│   ├── evaluation.py        ← CER/WER & error analysis
│   ├── train.py             ← TrOCR fine-tuning module
│   └── colab_script.py      ← Orchestration entry point
└── Collab_LLM_Pipeline.ipynb ← Colab execution notebook
└── requirements.txt
└── README.md
```

10. Conclusion

This submission demonstrates a complete, professionally structured, modular HTR pipeline that directly addresses the RenAIssance project's requirements:

- **Handwritten Text Recognition**: TrOCR fine-tuning infrastructure.
- **LLM used throughout** (not just late-stage): Beam Reranking + Paragraph Correction.
- **CER/WER measured**: Across all 3 stages, with error analysis.
- **Professional code structure**: Modular Python package, not notebook-only.
- **Historical orthography preserved**: No modernization in LLM correction prompts.
- **Fully reproducible**: Runs on free Google Colab T4 GPU.

The high zero-shot CER (0.99) is an expected and scientifically documented consequence of applying a modern-handwriting model to archival manuscripts without domain adaptation. The primary contribution of this test is a **reproducible, LLM-integrated architecture** that is immediately ready for fine-tuning.

