

Codebase & Experiments

Codebase Guide: Optimizing-LLM-based-OCR-Corrections

1. Directory Structure

- `code/experiments/` :
 - `confbert_router.py` : ConfBERT token-level scoring and routing pipeline.
 - `experiment_linear_regression.py` : Lasso and Ridge feature selection and cross-validation.
 - `experiment_svm_classifier.py` : Support Vector Machine routing.
 - `experiment_nn_regression.py` : Neural network router.
 - `lazy_clf_scan.py` : AutoML screening of 30+ regression algorithms.
- `code/plotting/` :
 - `plot_routing_frontier_paddle.py` : Generates CER vs Cost Pareto curves.
 - `plot_threshold_sweep.py` : Threshold sensitivity analysis.
 - `plot_error_confidence_cer.py` : Correlation between OCR confidence and CER.
- `code/evaluation/` :
 - `run_evaluations.py` : Batch evaluation harness across Tesseract, EasyOCR, and PaddleOCR.
- `data/` :
 - `evaluation_dataset/groundtruth.json` : Ground truth transcriptions for 609 segments.
 - `raw_ocr_results.json` : Output dumps from Tesseract, EasyOCR, and PaddleOCR.

Revision #3

Created 2026-09-09 16:20:20 UTC by Stergios

Updated 2026-09-17 12:21:53 UTC by Stergios