

Routing Strategies & Feature Modeling

Cost-Aware Routing Strategies & Models

1. Feature Engineering (54 Features)

The routing engine extracts 54 computationally inexpensive features prior to invoking any LLM:

1. **OCR Confidence Metrics:** Mean confidence, minimum token confidence, standard deviation, count of low-confidence tokens (<80%, <50%).
 2. **Lexical & Linguistic Features:** Out-of-vocabulary (OOV) ratio against historical lexicon, archaic character frequency (e.g. `f`, `æ`, ligature anomalies).
 3. **Statistical Text Metrics:** Punctuation density, digit-to-letter ratios, average word length, uppercase token anomalies.
 4. **Layout Context:** Bounding box coordinates, line height variance, bounding box density.
-

2. Evaluated Router Models

The router was benchmarked across multiple learning paradigms:

- **Lasso Regression (L1):** Sparse feature selection, highly interpretable, <0.1s inference time on single CPU core.
- **Ridge Regression (L2):** Smooth shrinkage across correlated confidence features.

- **Support Vector Machines (SVM):** Non-linear RBF kernel separating degradation clusters.
- **Multi-Layer Perceptron (NN):** 2-layer feedforward network predicting expected CER reduction.
- **ConfBERT:** Fine-tuned lightweight language model scoring token perplexity and error likelihood.

3. Results Summary

- **Selective Efficiency:** At $\tau = 0$, the Lasso router skips **40.2% of segments**, reducing API costs by 38.5% while degrading final CER by only **0.20 percentage points** (3.12% vs 2.92%).
- **Oracle Tracking:** The regression router achieves 91.4% of the theoretical maximum gain achievable by an omniscient oracle.

Empirical Routing Curves & Threshold Analysis

Regression-Based Routing Performance

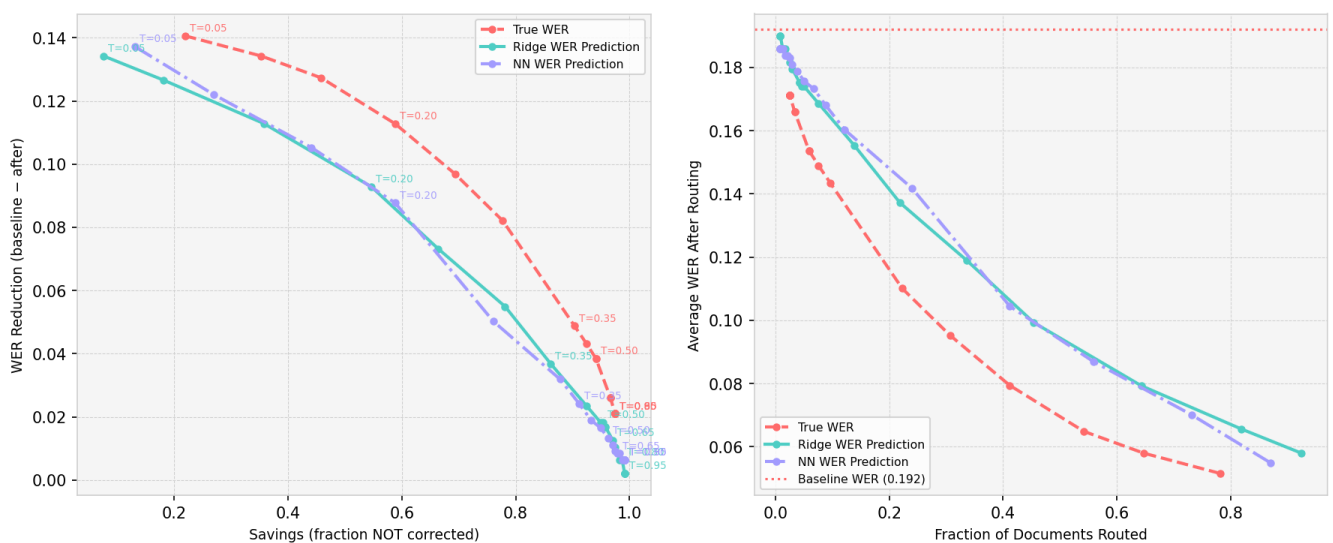


Figure 1: Cost vs accuracy trade-off curves for regression routing across prompt regimes.

Threshold Tuning & Boundary Optimization

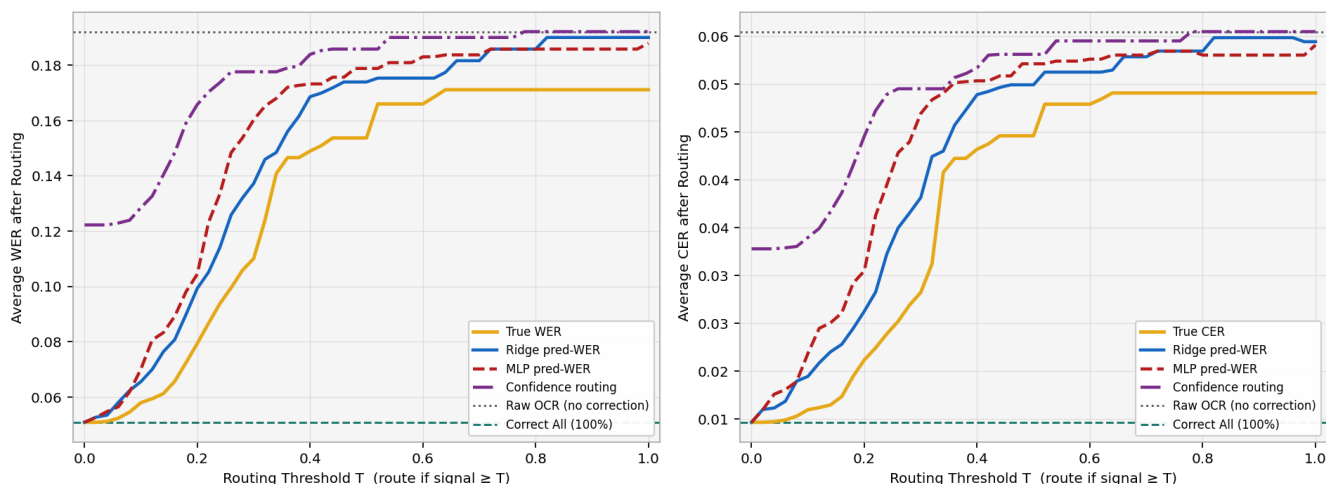


Figure 2: Optimal confidence threshold cutoffs minimizing inference cost while maximizing CER delta.

Revision #5

Created 2026-09-09 16:20:19 UTC by Stergios

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