Page Stream Segmentation with LLMs: Challenges and Applications in Insurance Document Automation

Hunter Heidenreich [1] Ratish Dalvi [1]
Nikhil Verma [1] Yosheb Getachew [1]

[1]Roots Automation, New York, NY

Overview

Why does this matter? Page Stream Segmentation (PSS) is essential for automating document processing in industries like insurance, where unstructured document collections are common. This study evaluates **Large Language Models (LLMs)** in real-world insurance datasets, overcoming challenges in accuracy, calibration, and automation potential.

Abstract

This study investigates the use of **LLMs** for PSS, leveraging parameter-efficient fine-tuning on real-world insurance data. Key findings include:

- **Performance:** LLMs outperform baseline models like XGBoost in page- and stream-level segmentation accuracy.
- **Calibration:** Post-hoc calibration and Monte Carlo dropout yield limited improvements for stream-level confidence.
- **Automation Potential:** At high confidence thresholds, LLMs handle more streams automatically with minimal human intervention.

Key Contributions

- **Real-World Evaluation:** Confirmation of trends from synthetic evaluation, LLMs outperform XGBoost on real-world PSS tasks.
- **Calibration Assessment:** Evaluation of post-hoc methods for mitigating overconfidence.
- **Stream-Level Confidence:** Introduction of a confidence measure for automating streams while flagging uncertain cases.

Experimental Methods

- **Models:** Fine-tuned decoder-only LLMs, including Mistral-7B and Phi-3.5-mini, using Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning.
- **Baselines:** XGBoost with TF-IDF-based page representations as a traditional model for comparison.
- **Calibration Techniques:** Post-hoc recalibration (logistic regression) and Monte Carlo (MC) dropout to estimate uncertainties.
- **Evaluation Metrics:** Performance measured at page and stream levels using precision, recall, and F1. Calibration assessed via Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

Confidence Definition

Page-Level Confidence

For each page p_i, confidence is defined as:

$$C_i = p_i \cdot \mathbb{I}(p_i > 0.5) + (1 - p_i) \cdot \mathbb{I}(p_i \leq 0.5),$$

where p_i is the calibrated probability of the page being a new document start, and $\mathbb{I}(\cdot)$ is the indicator function.

Stream-Level Confidence

The product of page-level confidences for all N pages in the stream:

$$C = \prod_{i=1}^{N} C_i.$$

This determines the overall confidence in the segmentation accuracy of a document stream.

Results at a Glance

- **Model Performance:** Mistral achieves the highest stream-level F1 score (0.947).
- **Calibration:** Post-hoc recalibration reduces ECE to 0.009 for Mistral*.
- **Automation:** At 80% confidence, Mistral automates 70% of streams while maintaining high accuracy.

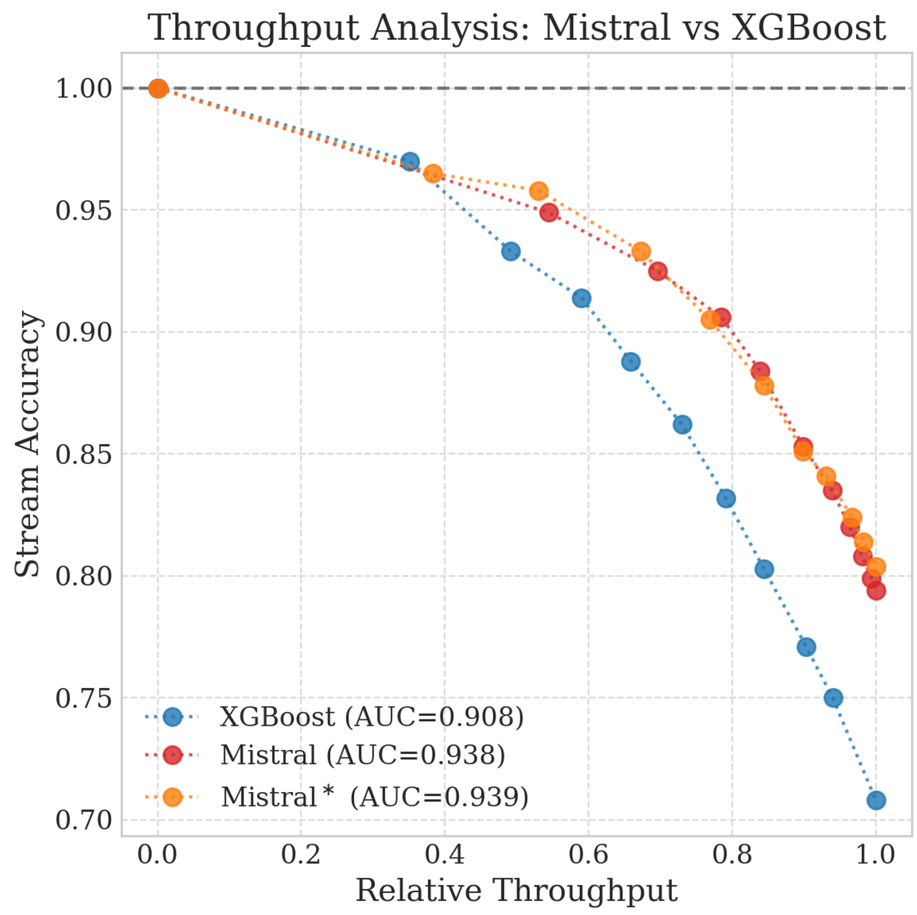
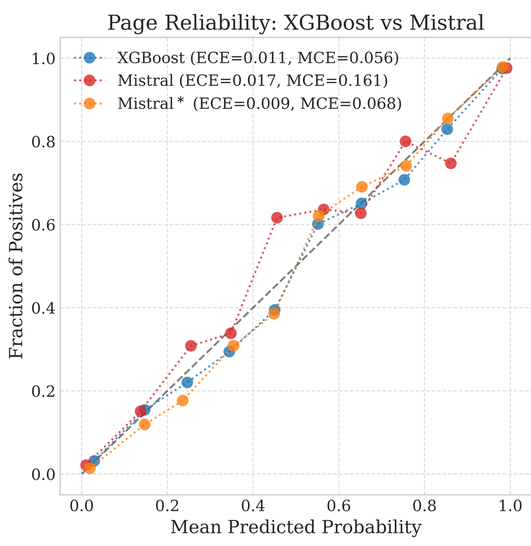
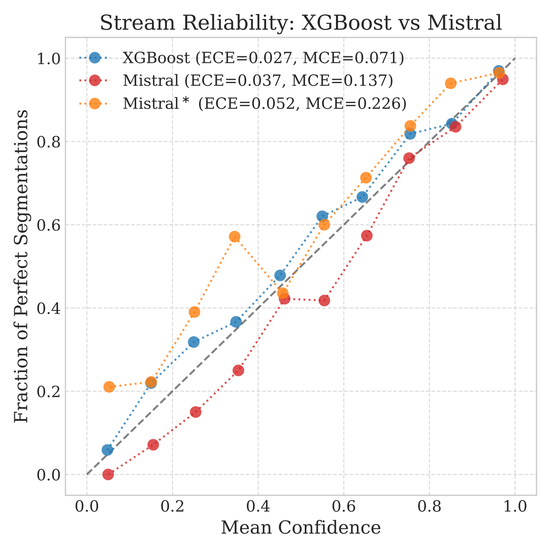


Figure 1. Accuracy vs. throughput for Mistral and XGBoost models.





Conclusions and Future Work

Conclusions

- LLMs significantly outperform baselines in both segmentation accuracy and automation potential.
- Calibration remains a challenge, particularly for stream-level confidence estimates.

Future Directions

- Explore active learning to iteratively improve calibration and accuracy.
- Develop synthetic benchmarks closely mimicking real-world data.