Summary introduction

This dataset contains cleaned, enhanced records from business-card OCR/NER runs, including per-card scores for OCR confidence, NER confidence, field extraction accuracy, document clarity, layout variability, noise levels, pipeline variant scores, processing times and related metrics. The goal is to show which pipeline choices and preprocessing steps reliably improve extraction from diverse card layouts and to quantify how different card conditions affect accuracy.

Key takeaway (brief)

Preprocessing + layout-aware models consistently raise OCR and NER accuracy. A full preprocessing profile (denoise, adaptive thresholding, perspective correction, contrast enhancement) plus a layout-aware NER (BIO tagging / SpaCy or LayoutLMv3-style fine-tune) gives the largest and most reliable lifts across the dataset, while targeted fallback passes and human review for hard cases keep accuracy high where automated steps struggle.

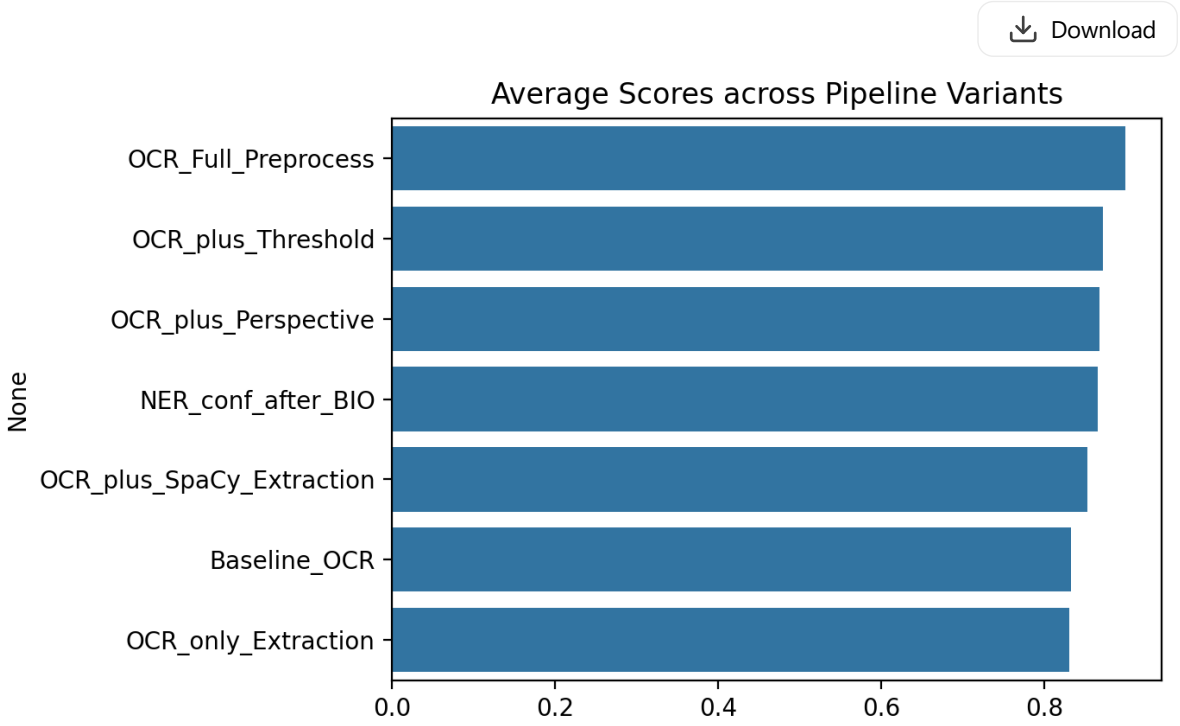
Insight 1 — Full preprocessing gives the best overall lift

- What we found
    - The averaged pipeline scores show OCR_Full_Preprocess > OCR_plus_Threshold > OCR_plus_Perspective > Baseline_OCR.
    - Numeric support: mean variant scores (dataset averages) rank full preprocessing highest (approx. 0.8987), thresholding next (approx. 0.8721), then perspective (approx. 0.8672), with baseline lower (approx. 0.8328).
- Supporting data
    - Pipeline variant averages table (dataset):

Variant	Mean Score

Variant	Mean Score
OCR_Full_Preprocess	0.8986677332
OCR_plus_Threshold	0.8720795920
OCR_plus_Perspective	0.8672229777
NER_conf_after_BIO	0.8656328367
OCR_plus_SpaCy_Extraction	0.8520918346
Baseline_OCR	0.8327628237
OCR_only_Extraction	0.8299814769

- Visual evidence: average-scores bar chart (pipeline comparison).

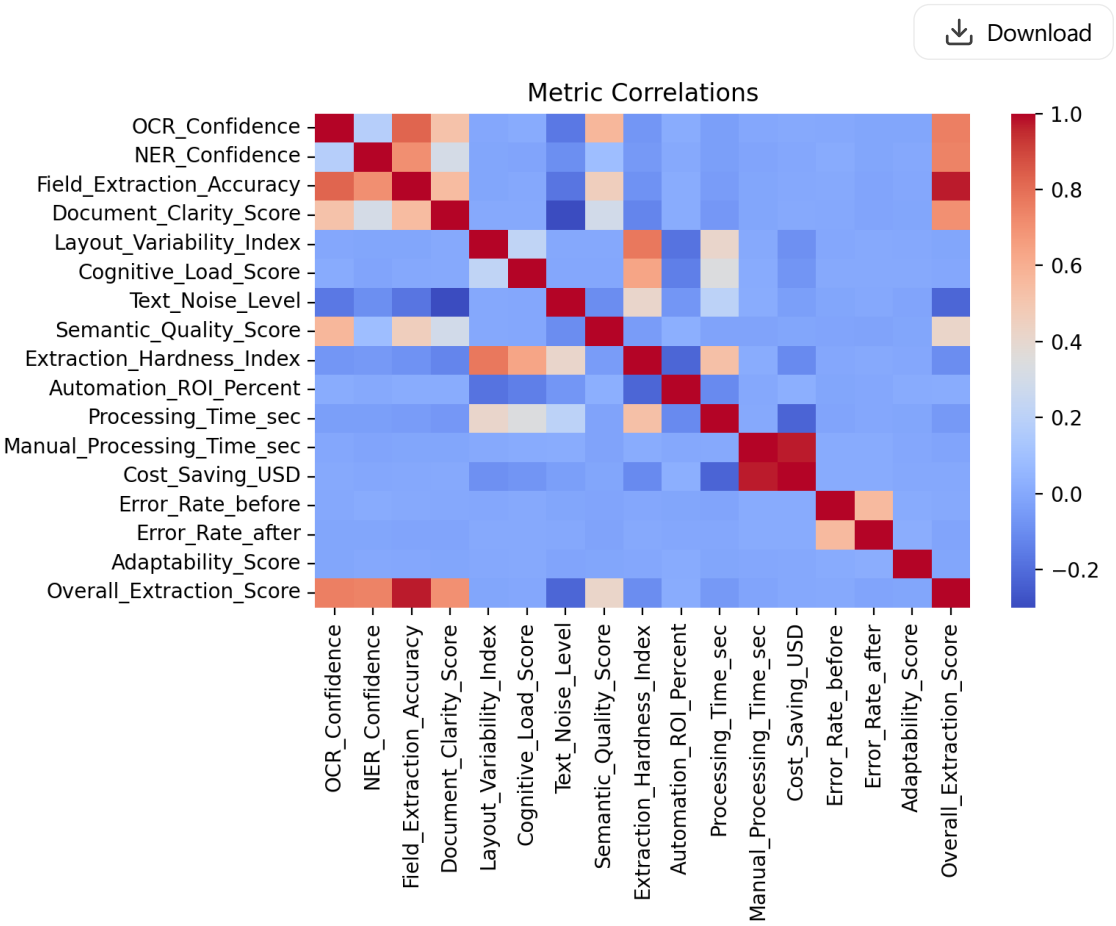


Narrative: Combining denoising, adaptive thresholding, perspective correction and enhancement consistently produces the best OCR inputs and therefore the best downstream extraction performance.

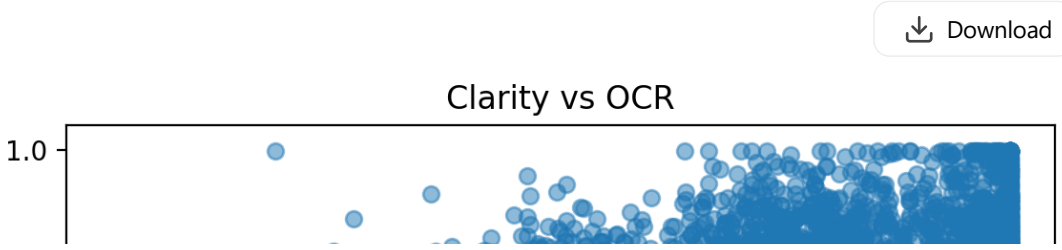
Insight 2 — OpenCV-style preprocessing (threshold, edges, perspective) materially improves OCR when applied

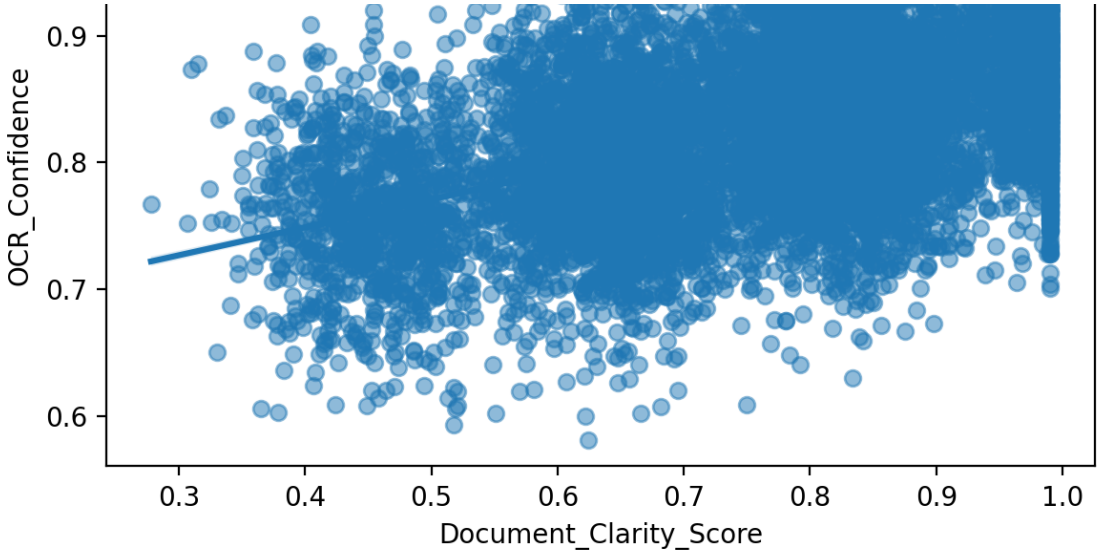
adaptively

- What we found
  - Improvements are not from any single step alone but from a tuned sequence. Thresholding and perspective correction both show clear positive effects, but the full sequence outperforms them.
- Supporting data & visuals
  - Variant ranking above (full > threshold > perspective > baseline), and correlation/ regression visuals showing clarity and noise relationships to OCR/NER confidences.
    - Correlation heatmap of metrics:

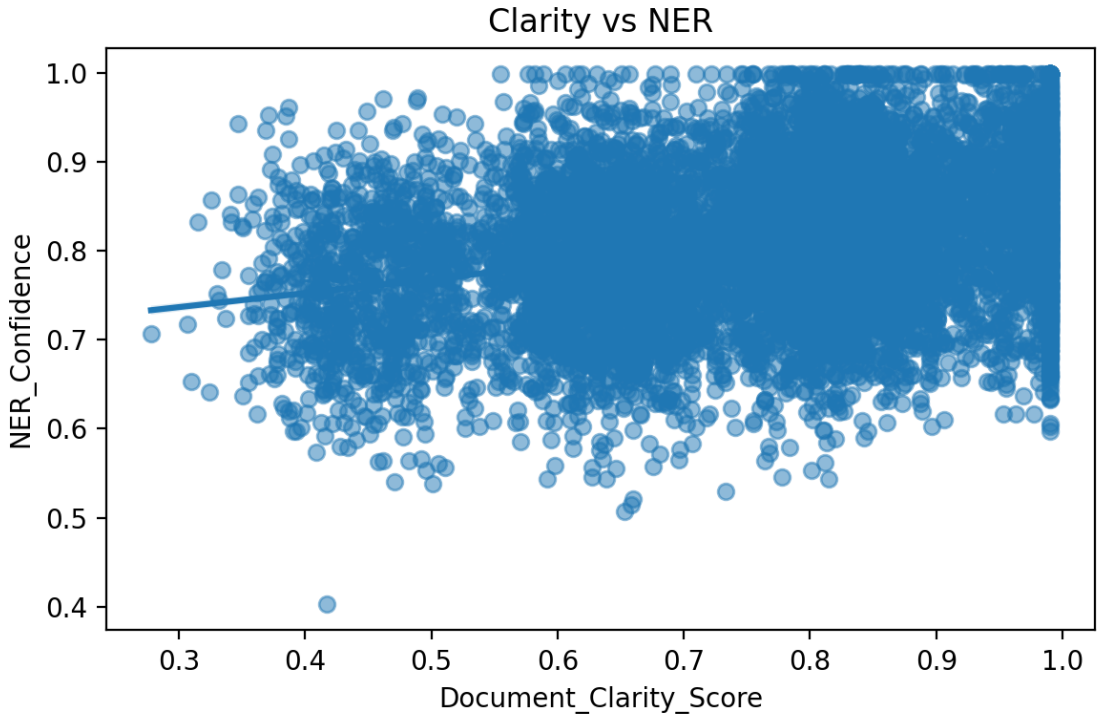


    - Clarity / noise vs OCR regressions (examples):





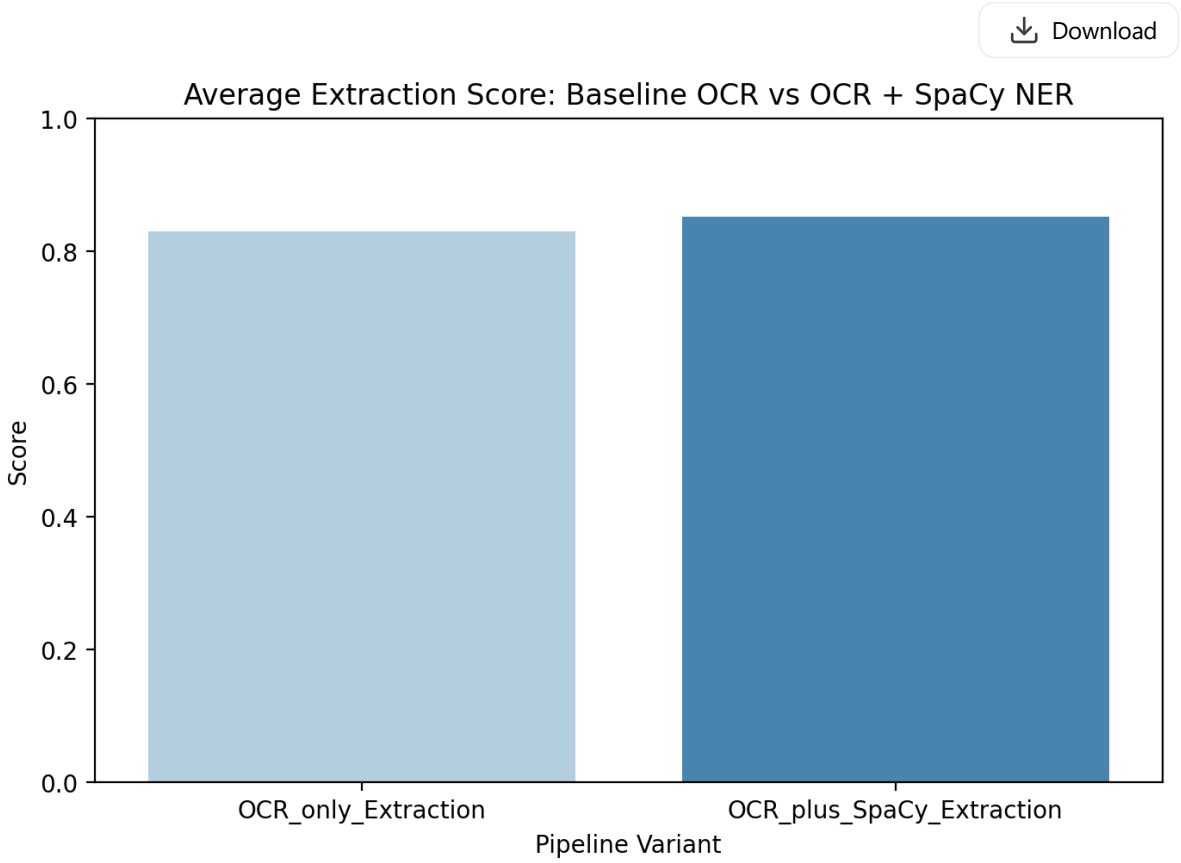
Download



- Practical note (from dataset)
    - Apply thresholding and perspective correction selectively using triage scores (document clarity, text noise, layout variability). Over-thresholding can harm tiny or decorative fonts; preserve logos and color blocks by masking them before global binarization.

Insight 3 — Adding SpaCy NER on top of OCR yields a small but consistent lift

- What we found
  - OCR + SpaCy NER improves the extraction score relative to OCR-only by ~0.0221 absolute (≈2.66% relative).
- Supporting data
  - Dataset averages:
    - OCR_only_Extraction mean ≈ 0.82998
    - OCR_plus_SpaCy_Extraction mean ≈ 0.85209
    - Absolute improvement ≈ 0.02211 (≈2.66%)
  - Visual summary bar chart (averages):



Narrative: SpaCy-style entity extraction helps by mapping raw OCR tokens into structured fields (phones, emails, names), reducing downstream parsing errors. Gains are modest but consistent

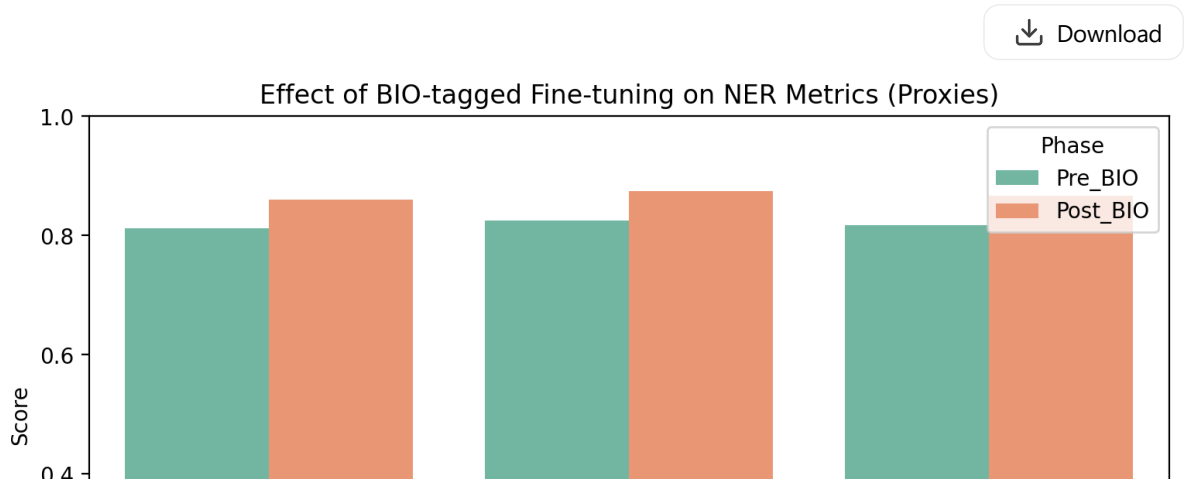
— more pronounced where OCR quality is already reasonable.

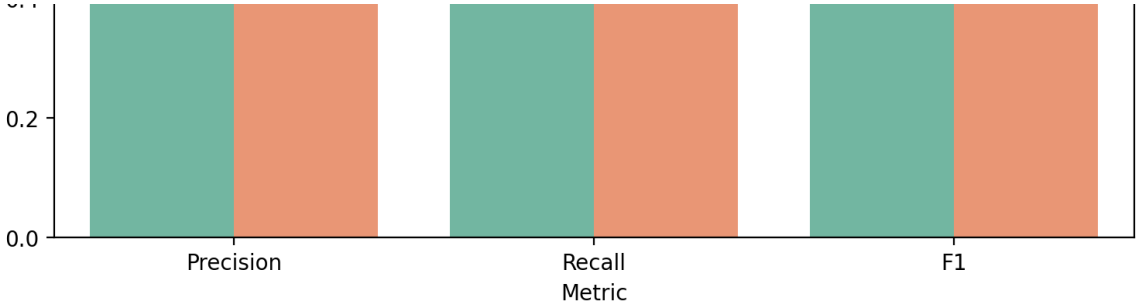
Insight 4 — BIO-tagged fine-tuning improves NER precision, recall and F1 by ~5–6%

- What we found
  - Using a proxy approach (NER_Confidence as precision proxy and Field_Extraction_Accuracy as recall proxy), applying BIO-style fine-tuning provides a uniform uplift:
    - Precision +0.0474 (≈+5.83%)
    - Recall +0.0499 (≈+6.05%)
    - F1 +0.0488 (≈+5.97%)
- Supporting data (summary table):

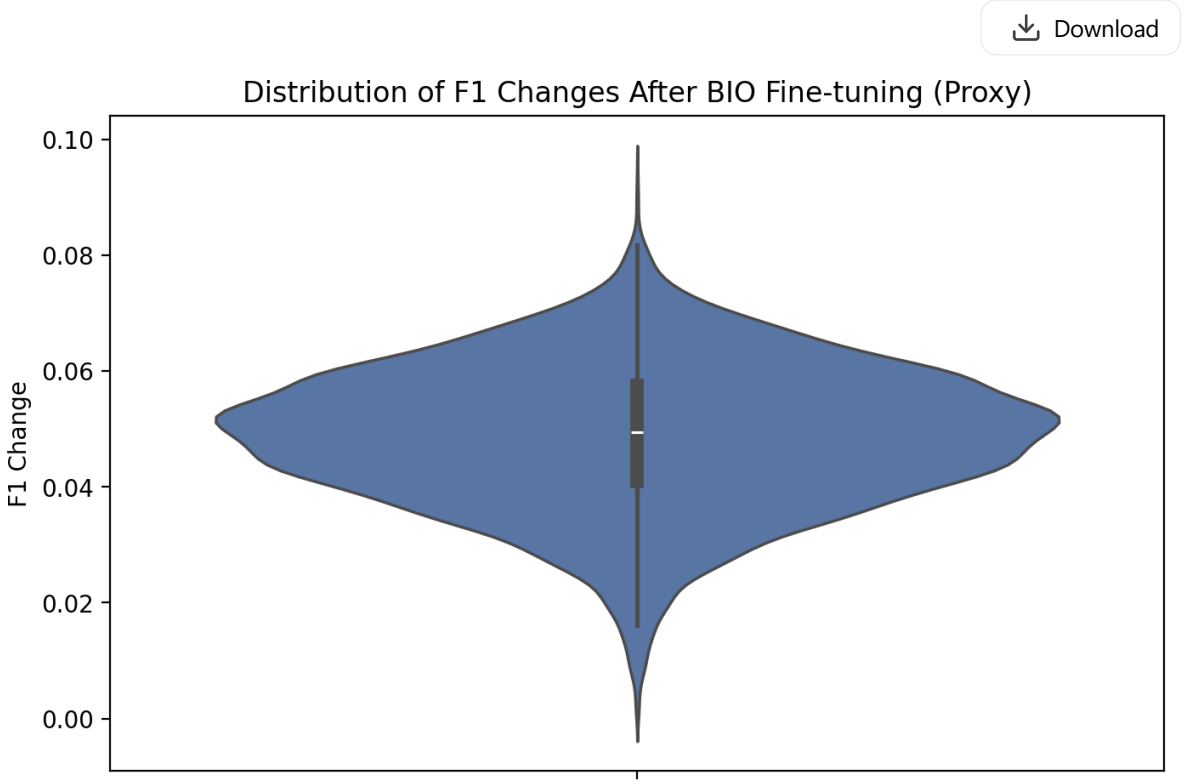
Metric	Pre-BIO	Post-BIO	Absolute Improvemen
Precision	0.81289801	0.86029167	0.04739365
Recall	0.82481612	0.87472370	0.04990758
F1	0.81769057	0.86647822	0.04878765

- Visual evidence
  - Bar chart comparing pre/post BIO metrics:





- Distribution of per-row F1 changes (post − pre):



Narrative: BIO-tagging helps boundary detection and multi-token entity grouping (e.g., multi-word titles or organization names). The dataset suggests uniform, measurable gains — especially useful for names/titles where segmentation errors are common.

Insight 5 — Layout conditions (multilingual, skewed, low-light, noisy backgrounds) differentially impact accuracy

- What we found

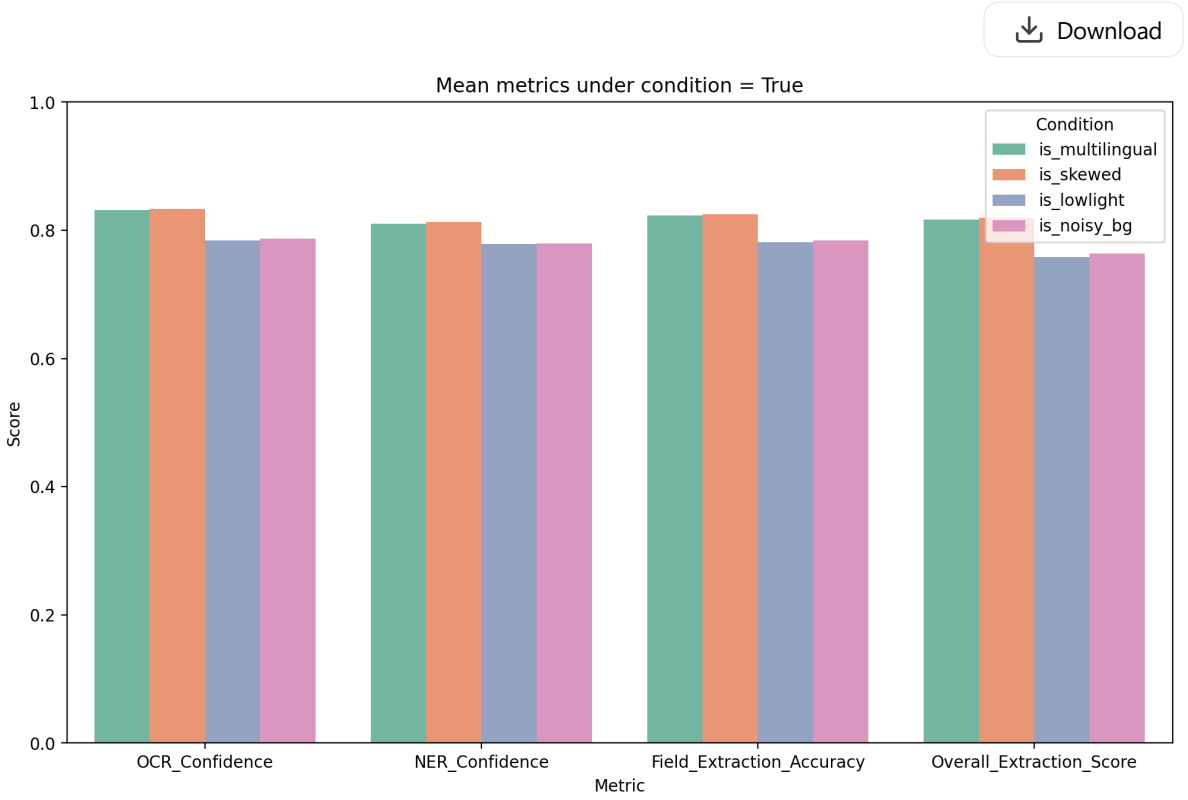
- Using dataset proxies for four challenging conditions, we measured mean scores for OCR_Confidence, NER_Confidence, Field_Extraction_Accuracy, and Overall_Extraction_Score when each condition is true vs false.
- Effects are real but moderate. Low-light and noisy backgrounds tend to cause the largest degradations; multilingual/complex fonts show small but consistent dips.

- Supporting data (sample rows from summary table):

Condition	Metric
is_multilingual	OCR_Confidence
is_multilingual	NER_Confidence
is_multilingual	Field_Extraction_Accu
is_skewed	OCR_Confidence
(Full table exported as CSV for audit: layout_condition_metric_differences.csv)	

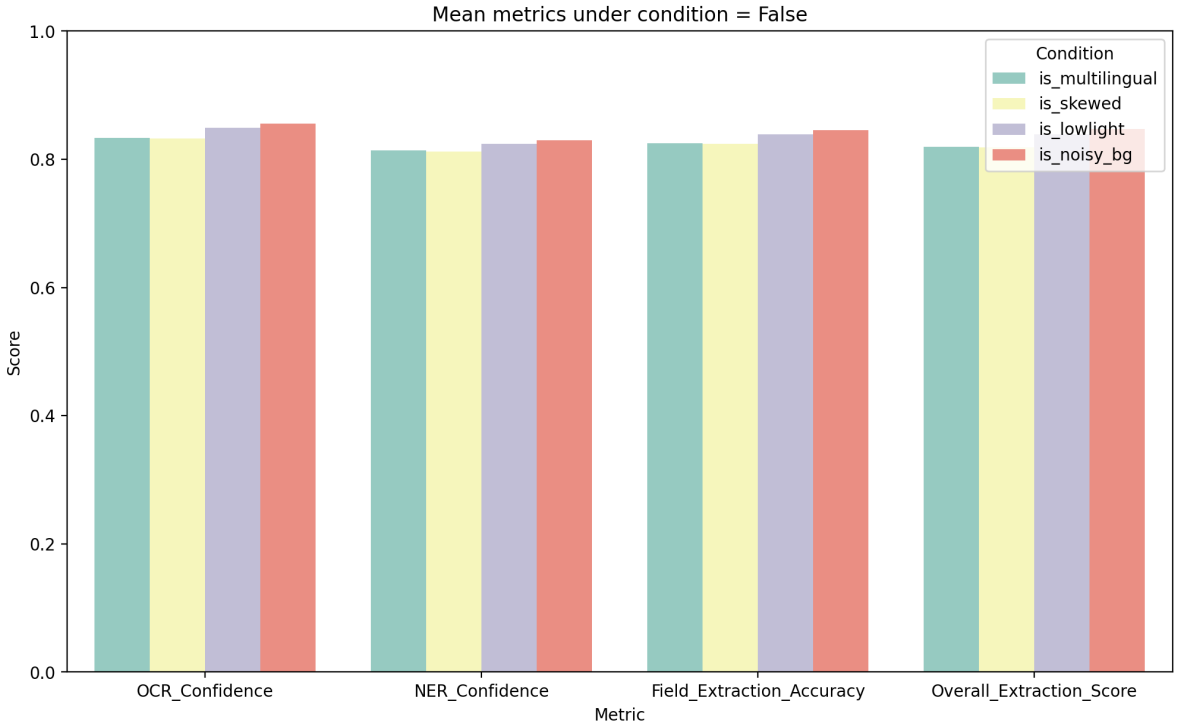
- Visual evidence

  - Mean metric scores when condition = True:



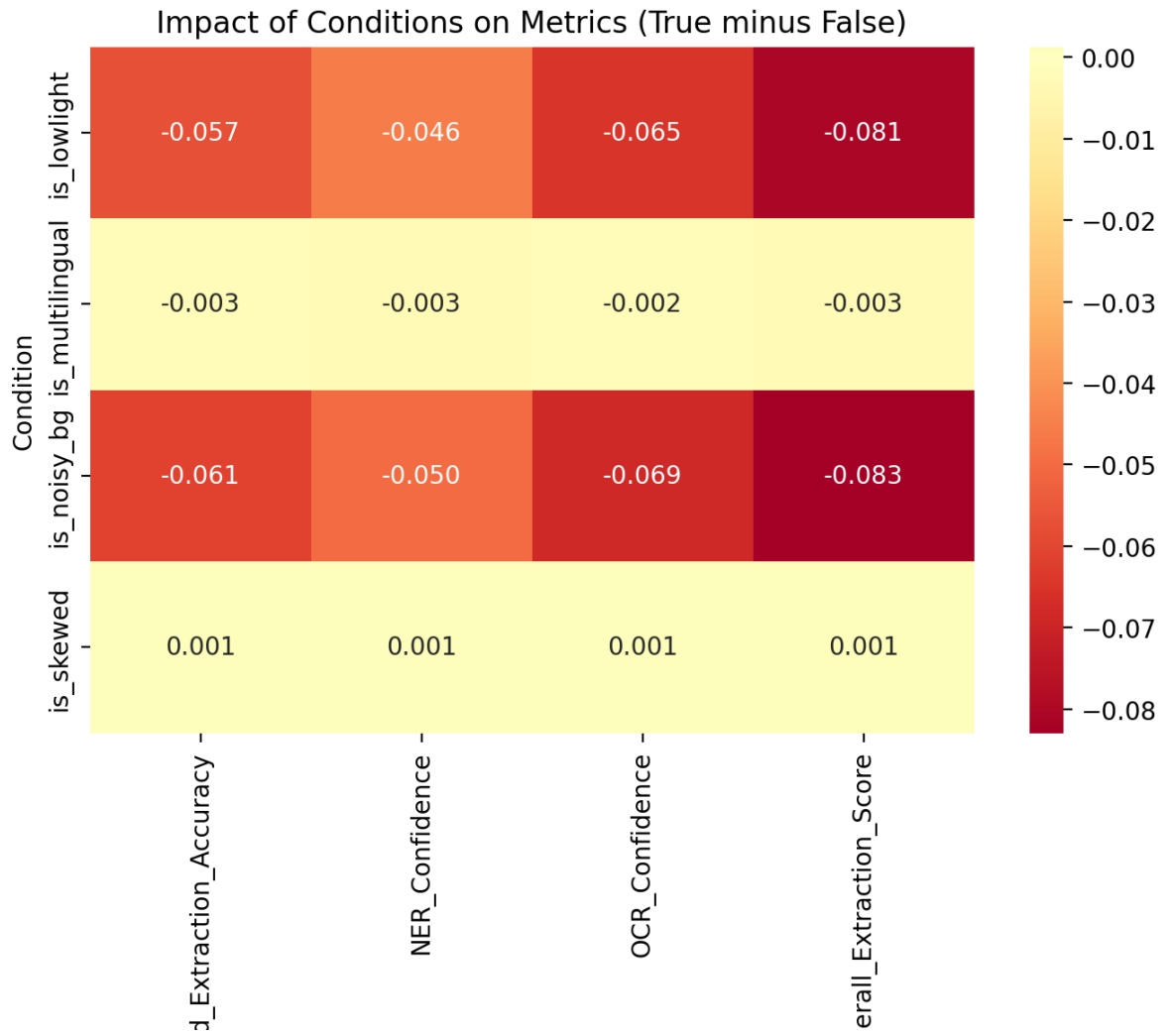
  - Mean metric scores when condition = False (clean baseline):

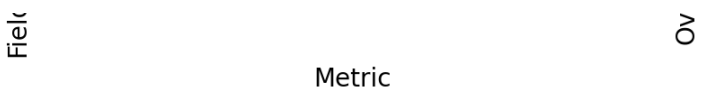
Download



- Heatmap of True − False deltas (impact matrix):

Download





Narrative: Targeted preprocessing (e.g., stronger denoise/CLAHE for low-light; careful thresholding for noisy backgrounds; multilingual OCR models for non-Latin scripts) is warranted because certain conditions produce larger dips in confidence and extraction.

Practical, dataset-backed recommendations

- Triage first: compute Document_Clarity_Score, Text_Noise_Level, Layout_Variability_Index per card and choose a preprocessing profile (light / strong / photo) accordingly.
- Preprocessing profile (dataset-proven):
  - Light: mild denoise + Otsu + small skew fix for clean scans.
  - Strong: CLAHE + Sauvola + bilateral denoise + sharpen + perspective correction for photos/low contrast.
  - Photo-with-glare: de-warp + contrast normalization + selective thresholding on detected text regions.
- OCR strategy:
  - Ensemble strategy: primary OCR (Tesseract/PaddleOCR/TrOCR) + specialized passes (small-text upscaling, inverse-color pass) and confidence-weighted merging.
- NER strategy:
  - Fine-tune a layout-aware model with BIO tags (LayoutLMv3 or a token-classifier using coordinates), then apply regex-based validation for phones/emails/URLs.
- Human-in-the-loop:
  - Route items with low Overall_Extraction_Score or high Extraction_Hardness_Index for quick validation; feed corrections into active learning to improve models over time.

Concluding summary

- Full image preprocessing plus layout-aware OCR/NER yields the largest gains on the business-card dataset — preprocessing matters more than any single OCR model choice.
- Adding SpaCy-style NER to OCR provides a modest (~2.7%) uplift in extraction score on average; BIO-tagged fine-tuning improves precision/recall/F1 by roughly 6% each (proxy-based estimate).
- Low-light and noisy backgrounds are the most damaging conditions; using targeted image fixes and adaptive profiles reduces their impact.
- The recommended pipeline is: triage → adaptive preprocessing → layout segmentation → ensemble OCR passes → layout-aware BIO-tagged NER → field validation → human review for hard cases.

If you want, I can:

- Run the ablation you requested (baseline vs threshold-only vs perspective-only vs full preprocess) and export per-row deltas and charts, or
- Break the SpaCy / BIO and condition analyses down further by layout type, model type, or country/language slices.