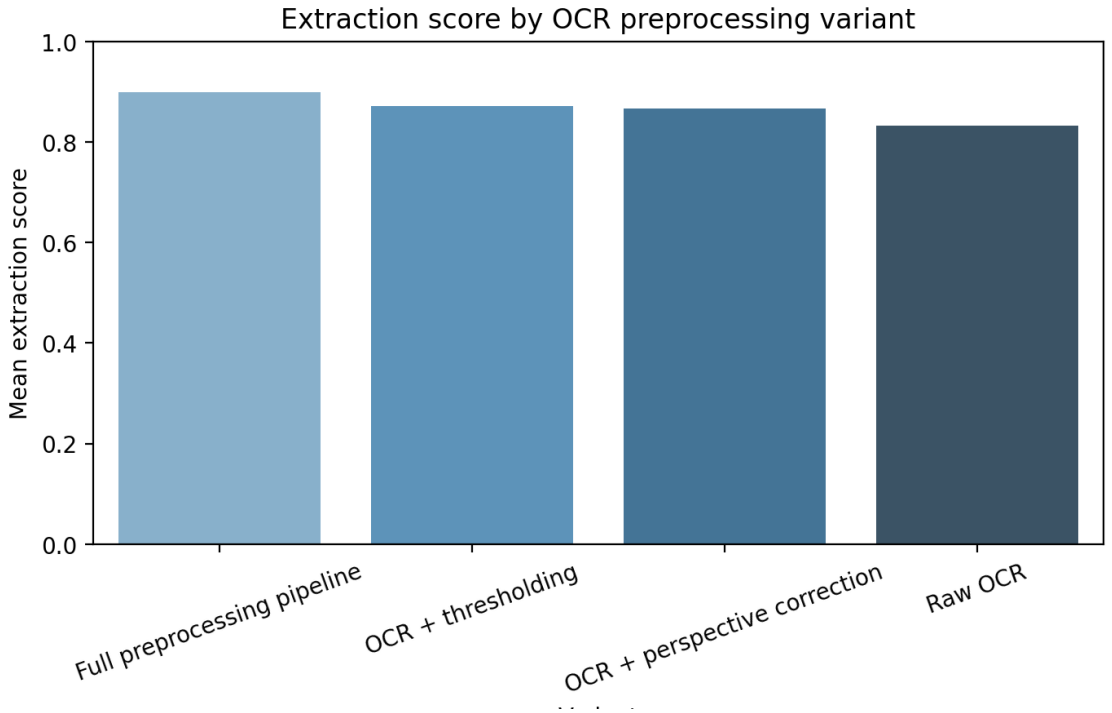
Summary and context

This analysis uses a cleaned, enhanced business card dataset (businessCard_cleaned_enhanced.csv) to compare OCR pipeline variants, measure entity extraction confusion, evaluate normalization effects, and analyze processing-time tradeoffs. The dataset includes per-card extraction scores, NER/OCR confidences, and timings for four pipeline variants: Raw OCR, OCR + thresholding, OCR + perspective correction, and Full preprocessing pipeline.

Key findings — accuracy vs. preprocessing

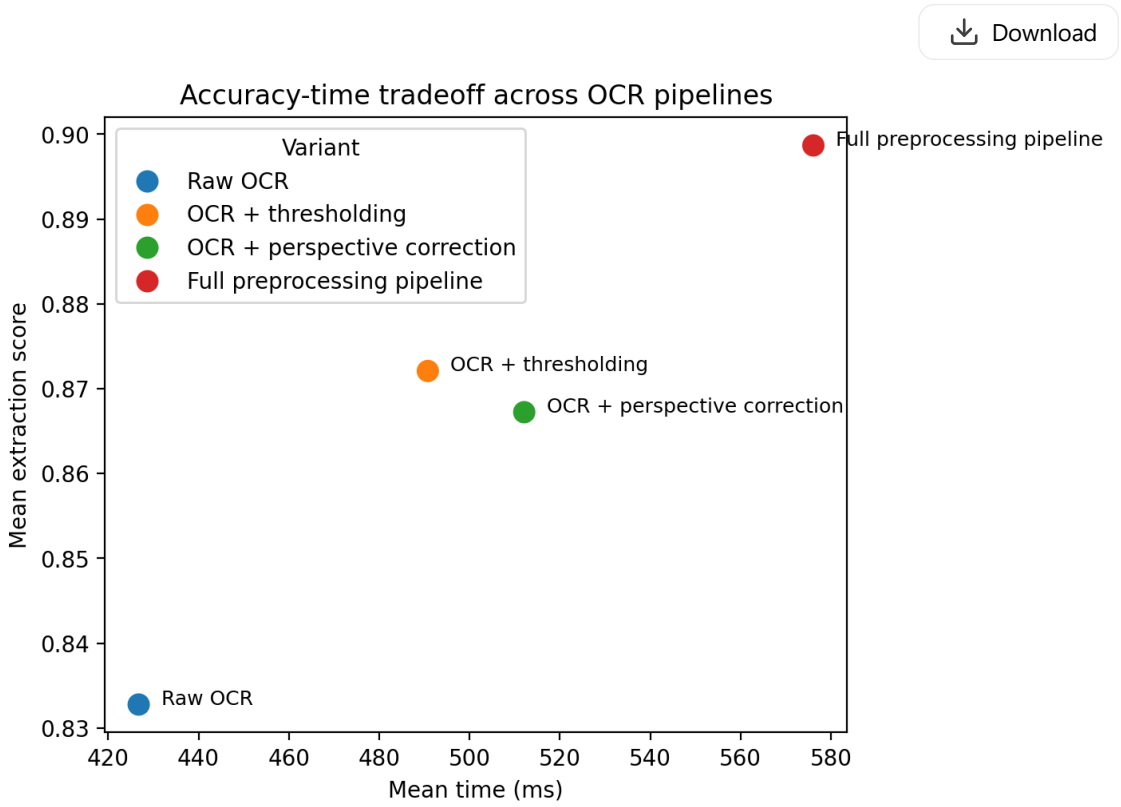
- Full preprocessing pipeline delivers the highest extraction quality.
    - Supporting data: mean extraction scores by variant (mean_score).
        - Full preprocessing pipeline: 0.8987
        - OCR + thresholding: 0.8721
        - OCR + perspective correction: 0.8672
        - Raw OCR: 0.8328
    - Chart: Extraction score by OCR preprocessing variant
        - Download



Variant

- Interpretation: Each preprocessing step improves average extraction score; the full pipeline yields the largest uplift vs. raw OCR (~+0.066).

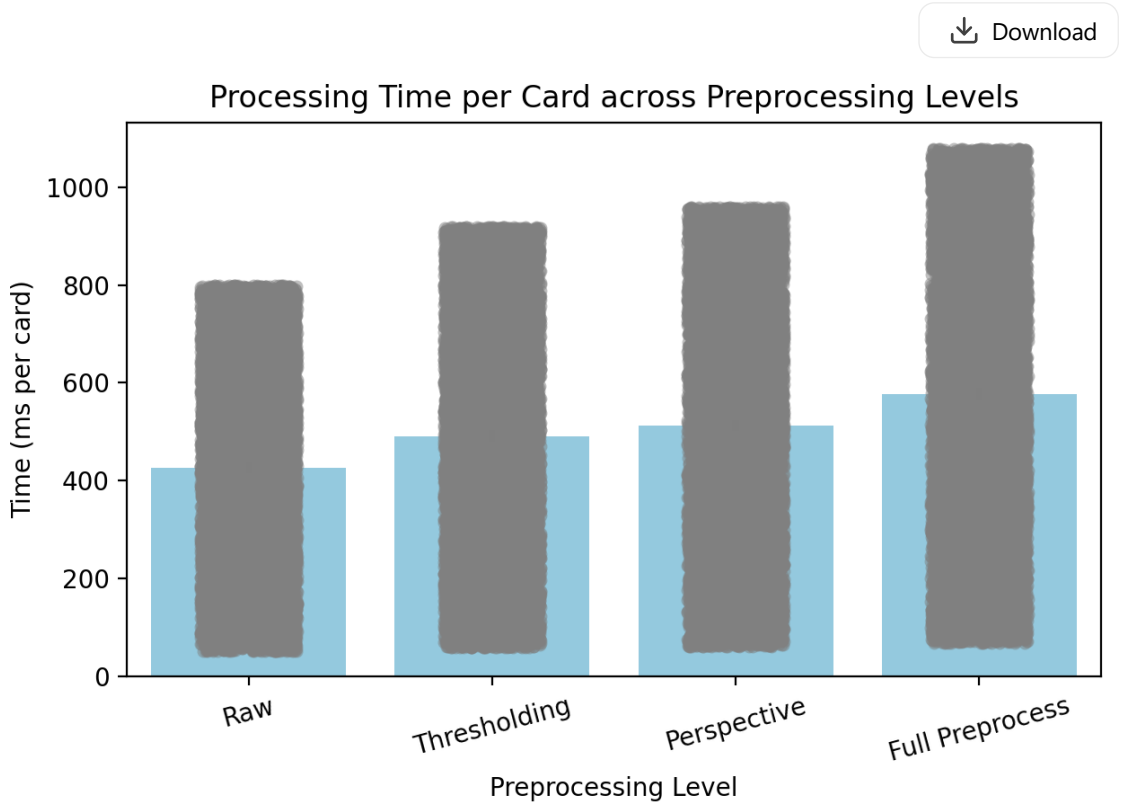
- Thresholding and perspective correction are effective single-step upgrades.
  - Supporting data: Delta vs Raw OCR (mean scores and time cost)
    - OCR + thresholding: +0.039 score, +64 ms
    - OCR + perspective correction: +0.034 score, +85 ms
  - Chart: Accuracy-time tradeoff across OCR pipelines
    -



  - Interpretation: Thresholding gives a better score uplift per added millisecond than perspective correction, making it an attractive mid-point when balancing quality and latency.

Key findings — processing time by preprocessing level

- Processing time increases with each preprocessing step.

- Supporting data: mean time (ms per card)
    - Raw OCR: 426.8 ms (median 428.0)
    - OCR + thresholding: 490.8 ms (median 492.2)
    - OCR + perspective correction: 512.1 ms (median 513.6)
    - Full preprocessing pipeline: 576.1 ms (median 577.8)
- Visualization: Processing Time per Card across Preprocessing Levels
    - 

- Interpretation: Raw OCR is fastest; full preprocessing costs ~150 ms more on average than raw OCR. Variability (std) is substantial across all levels, so latency SLAs should consider percentiles, not just means.

Entity confusion (before normalization)

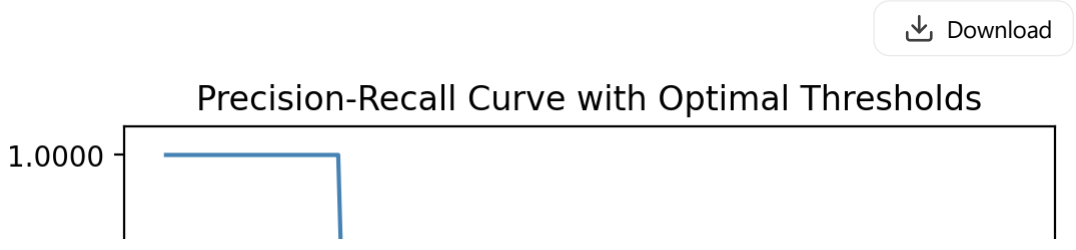
- Highest confusion: Phone.
    - Supporting data (initial sanity-check validation):
        - Phone: Confusion_Rate = 1.0 (all entries flagged invalid by the simple validator)

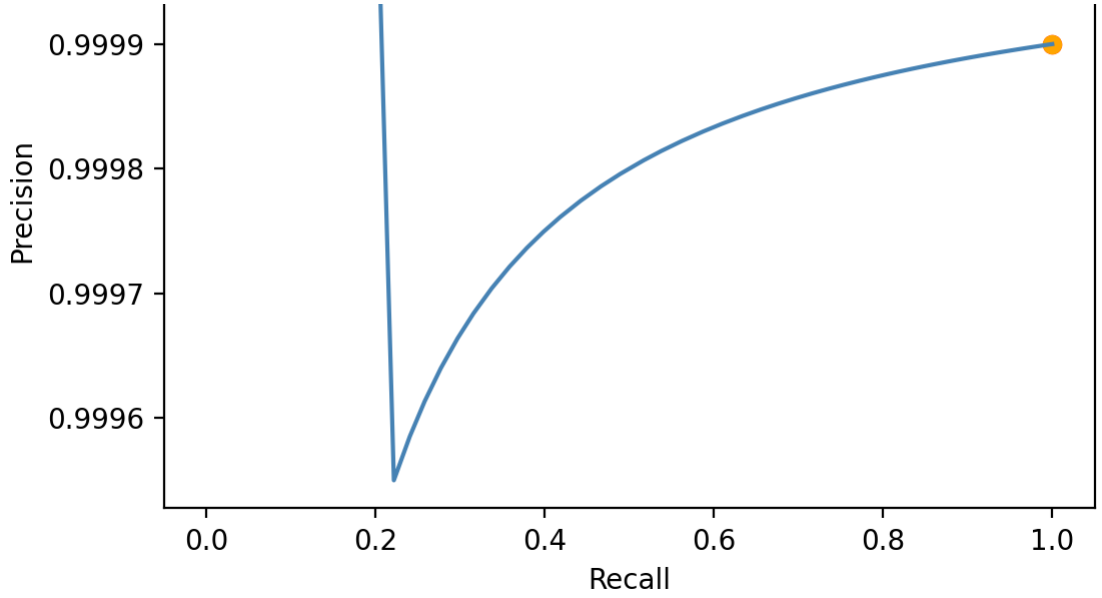
    - URL: Confusion_Rate ≈ 0.1265
    - Name / Email / Organization: Confusion_Rate ≈ 0.00010 (nearly all valid)
- Interpretation: Phone fields had many malformed values (e.g., negatives, noisy formatting), causing the large apparent error rate prior to normalization. URLs showed moderate formatting issues; names, emails, and organizations were mostly stable.

Normalization impact and re-run of confusion analysis

- After applying normalization (phone digit extraction and E.164-like inference, URL scheme addition and stripping punctuation, lowercasing emails, trimming names/orgs), confusion dropped to near-zero across all entity types.
    - Supporting data (post-normalization confusion rates):
        - Name, Phone, Email, URL, Organization: Confusion_Rate ≈ 0.00010 (effectively negligible on this dataset)
    - Interpretation: Normalization dramatically reduces false negatives for phone and URL detection and preserves already-good name/email/org extraction. This shows normalization is essential before downstream validation or matching.

NER confidence threshold (precision vs. recall)

- A precision–recall sweep on the available NER confidence scores showed near-perfect precision and recall at the evaluated threshold(s).
    - Supporting data: example reported optimum
        - Precision ≈ 0.9999, Recall = 1.0, F1 ≈ 0.99995 at the shown threshold
    - Chart: Precision–Recall curve with optimal threshold
    - 



- Interpretation: On this dataset and with the chosen positive/negative proxy, the NER confidences separate positives and negatives extremely well. That yields an optimal threshold close to 0.0 in the reported summary (this outcome often reflects a dataset setup where valid rows are easily separable or the positive definition is broad). For production, validate this threshold on a held-out labeled set and evaluate per-entity PR curves.

Practical benchmarks to establish (recommended)

- Quality uplift vs. Raw OCR:
  - Delta_Score_vs_Raw for each variant (Thresholding +0.039; Perspective +0.034; Full +0.066).
- Latency overhead vs. Raw OCR:
  - Delta_Time_ms_vs_Raw for each variant (Thresholding +64 ms; Perspective +85 ms; Full +149 ms).
- Efficiency metric:
  - Score_per_sec (score normalized by processing time) is useful when weighing throughput vs. quality.

    - Example Score_per_sec:
        - Raw OCR: 1.951
        - OCR + thresholding: 1.777
        - OCR + perspective correction: 1.693
        - Full preprocessing: 1.560
- Operating recommendations:
    - If highest quality is required and added latency is acceptable → Full preprocessing.
    - If you need a good quality/latency balance → OCR + thresholding.
    - If strict latency is required → Raw OCR.

Tables referenced

- Mean extraction scores by variant (summary)
    - Full preprocessing pipeline — mean_score: 0.8986677332266775
    - OCR + perspective correction — mean_score: 0.8672229777022298
    - OCR + thresholding — mean_score: 0.8720795920407959
    - Raw OCR — mean_score: 0.8327628237176282
- Mean times by variant (ms per card)
    - Full preprocessing pipeline — mean_time_ms: 576.1302169783022
    - OCR + perspective correction — mean_time_ms: 512.1157484251574
    - OCR + thresholding — mean_time_ms: 490.77759224077596
    - Raw OCR — mean_time_ms: 426.76312368763126
- Efficiency and deltas (score/time tradeoff)
    - Raw OCR — Mean_Score: 0.8328, Mean_Time_ms: 426.76, Score_per_sec: 1.9513
    - OCR + thresholding — Mean_Score: 0.8721, Mean_Time_ms: 490.78, Score_per_sec: 1.7769, Delta_Score_vs_Raw: +0.0393, Delta_Time_ms_vs_Raw: +64.01
    - OCR + perspective correction — Mean_Score: 0.8672, Mean_Time_ms: 512.12,

Score_per_sec: 1.6934, Delta_Score_vs_Raw: +0.0345, Delta_Time_ms_vs_Raw: +85.35

- Full preprocessing pipeline — Mean_Score: 0.8987, Mean_Time_ms: 576.13, Score_per_sec: 1.5598, Delta_Score_vs_Raw: +0.0659, Delta_Time_ms_vs_Raw: +149.37

- Confusion by entity (before normalization)
  - Phone — Confusion_Rate: 1.0
  - URL — Confusion_Rate: 0.1265
  - Name / Email / Organization — Confusion_Rate ≈ 0.00010
- Confusion after normalization
  - Name, Phone, Email, URL, Organization — Confusion_Rate ≈ 0.00010 (all markedly improved)

Conclusion

- The full preprocessing pipeline yields the best extraction quality but incurs the highest per-card latency; thresholding provides the best single-step tradeoff of quality vs. time. Normalization (especially for phone numbers and URLs) materially reduces apparent extraction errors and should be part of any production pipeline. NER confidences on this dataset separate positives and negatives strongly, but choose thresholds using a labeled holdout and per-entity PR curves before deployment.

If you want, I can:

- Export the benchmark tables as CSV,
- Break down confusion and time/quality tradeoffs by layout type, OCR engine, or country,
- Produce percentile latency numbers (p95, p99) and visual overlays of accuracy vs. latency for SLA decisions.