Design Document - Invoice Validation System

1. Overview

This document describes a proof-of-concept invoice validation system that matches OCR-extracted fields against ground truth and database records, classifies discrepancies, and outputs validation status with confidence.

2. Data and Inputs

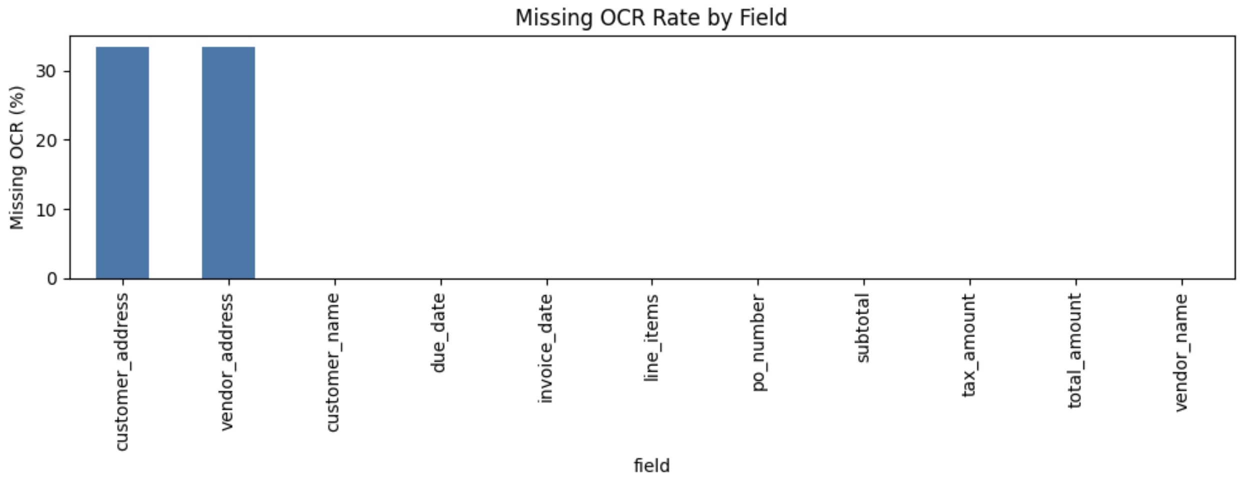
The system operates on three JSON sources:

- ground_truth.json: expected values for each invoice (labels).
- ocr_results.json: OCR-extracted values and confidence for each invoice (detections).
- database.json: reference data (purchase orders, vendor master, customer info).

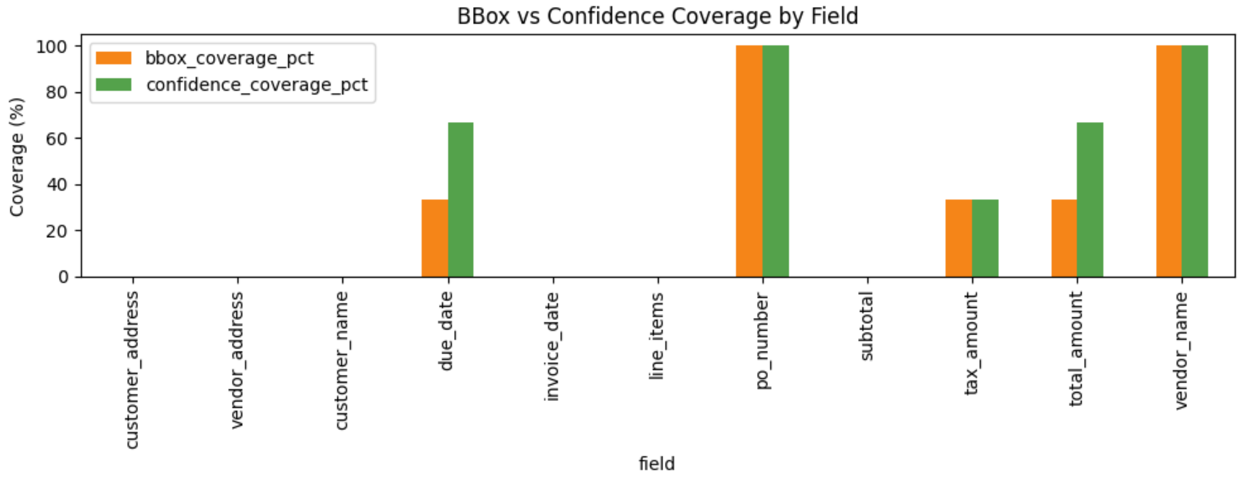
Primary fields include vendor/customer name and address, PO number, invoice and due dates, amounts (subtotal, tax, total), and line items.

3. Data Analysis & Understanding

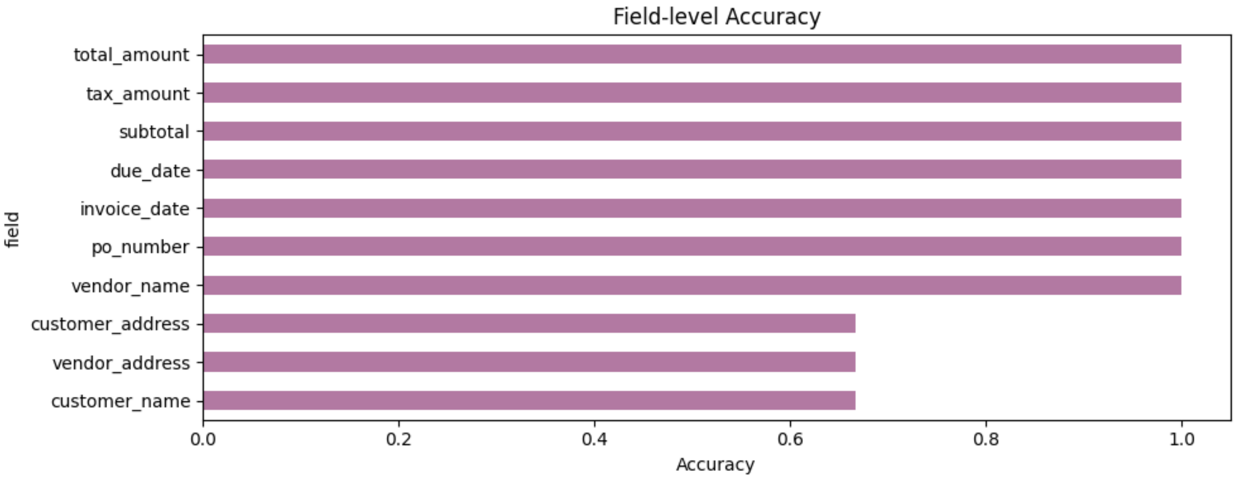
→ Missing OCR rate by field: Missing OCR rate concentrates on addresses, this aligns with OCR outputs that leave some address fields blank.



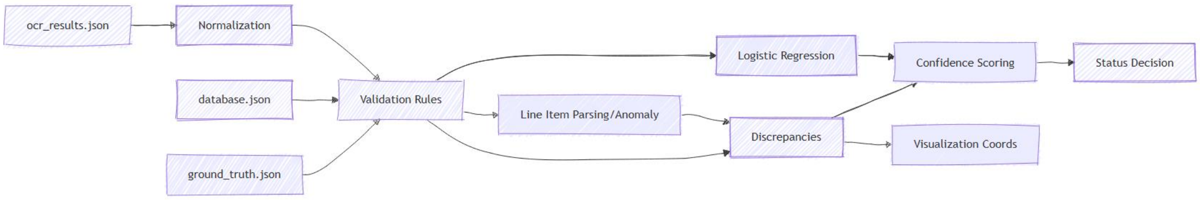
→ BBox vs confidence coverage: BBox vs confidence coverage is high for key fields (po_number, vendor_name) but sparse for others, indicating partial OCR annotations.



→ Field-level accuracy: Field-level accuracy is near 1.0 for dates and amounts, while names/addresses drop due to truncation/abbreviation.



4. Approach



The diagram illustrates the processing pipeline: OCR data is normalized, compared with ground truth and the database to create discrepancies.

These discrepancies are used to calculate confidence and determine status; they also generate highlight coordinates for the final JSON output.

1. *Data ingestion*

- Read ground_truth.json, ocr_results.json, database.json.
- Basic schema validation to ensure required keys exist.

2. *Normalization*

- Text: lowercase, strip punctuation, collapse whitespace.
- Company suffix normalization (inc, llc, ltd, co).
- Address abbreviation normalization (street -> st, drive -> dr).
- OCR confusable mapping (O/0, I/1/l, S/5, B/8).

3. ***Matching and validation***

- ID fields (PO): exact match after OCR-confusion normalization.
- Names/addresses: fuzzy similarity (token set + edit ratio).
- Dates: tolerance window in days.
- Amounts: absolute and relative tolerance.
- Database references:
    - PO vendor used as primary expected vendor name.
    - vendor_master address used for vendor address validation when available.
    - customer_info billing_address used for customer address validation.
    - tax_rate from PO used to cross-check tax_amount.
- Line items: parse text to (description, qty, unit_price, total).
    - Item not in PO list -> critical.
    - Quantity above max -> critical.
    - Line total mismatch -> warning.

4. ***Discrepancy scoring and status***

- Discrepancy severity: critical / warning / informational.
- Overall confidence combines OCR confidence and rule penalties.
- Status logic:
    - critical -> needs_review (POC default).
    - warnings or low confidence -> needs_review.
    - otherwise approved.

5. ***Option A (ML)***

- Logistic regression predicts probability a field is wrong.
- Features: OCR confidence, text length, digit ratio, fuzzy score, amount/date diffs, field identity.
- Model output calibrates confidence per field and overall score.

5. Rule Rationale & Thresholds

- PO/ID fields: strict matching is required because they drive payment linkage
- and vendor approval; OCR-confusion normalization (O/0, I/1/l, S/5, B/8)
- reduces false rejections from common OCR errors.
- Names/addresses: fuzzy matching handles abbreviations and truncation; thresholds
- are softer to avoid penalizing minor formatting differences.
- Dates: tolerance of 1-3 days is treated as warning based on typical OCR/entry
- drift; larger gaps are critical.
- Amounts: combined absolute and relative tolerance captures rounding/tax noise
- across both small and large invoices.
- Tax: tax_amount is cross-checked against PO tax_rate when available to validate
- accounting consistency.
- Line items: items not in PO list or quantity above max are critical because
- they violate approved purchasing rules; line total drift is warning to allow
- minor rounding.
- Status policy: in POC, critical issues route to needs_review to keep a
- human-in-the-loop; production can flip to rejected if desired.

6. Assumptions and Limitations

- OCR output is structured and aligned to field names.
- Ground truth is available for training (self-supervised labels).

- Limited sample size: ML model is illustrative only.
- Address parsing is token-based, not full geocoding.

7. Scalability Considerations

- Batch or streaming validation supported (stateless rules).
- Thresholds and mappings are centralized in config for tuning.
- Add vendor-specific templates to improve field extraction.
- Introduce async processing for large volumes.

8. Deliverables Mapping

- Notebook: data overview and metrics (data_overview.ipynb).
- Core engine: src/* modules.
- Sample outputs: sample_outputs/*.json (via scripts/generate_sample_outputs.py).
- Tests: pytest with 15 cases.

9. Tests

Pytest coverage includes 15 cases: missing fields, OCR confusion (O/0), date out-of-range, tax and amount tolerance, line item anomalies, and address abbreviation handling.

10. Assumptions and Limitations

OCR output is structured and aligned to expected field names.

Sample size is small; ML model is illustrative only.

No real invoice images provided; visualization bounding boxes may be placeholders.

11. How to Run

```
python -m venv .venv
.\.venv\Scripts\Activate.ps1
python -m pip install -r requirements.txt
python .\scripts\run_demo.py
python .\scripts\generate_sample_outputs.py
```

12. Outputs and Results

- INV-2024-001: status=needs_review, confidence=0.541 (2 discrepancies). po_number [warning] expected=PO-78910 detected=P0-78910; customer_name [warning] expected=Suri Technologies detected=Suri Tech

- INV-2024-002: status=needs_review, confidence=0.085 (3 discrepancies). po_number [warning] expected=PO-45678 detected=PO-4567B; customer_address [warning] expected=456 Data Street, Austin, TX 78701 detected=; line_items[1].quantity [critical] expected=3 detected=5

- INV-2024-003: status=needs_review, confidence=0.680 (1 discrepancies). vendor_address [warning] expected=77 Lakeview Ave, Chicago, IL 60601 detected=

13. Notebook Table

Invoice Count Summary

Unnamed: 0	gt_invoices	ocr_invoices	in_both	missing_in_ocr	missing_in_gt
0	3	3	3	0	0

Missing OCR / Coverage by Field

Unnamed: 0	field	missing_ocr	missing_ocr_pct	bbox_coverage_pct	confidence_coverage_pct
0	customer_address	1	33.333	0.000	0.000

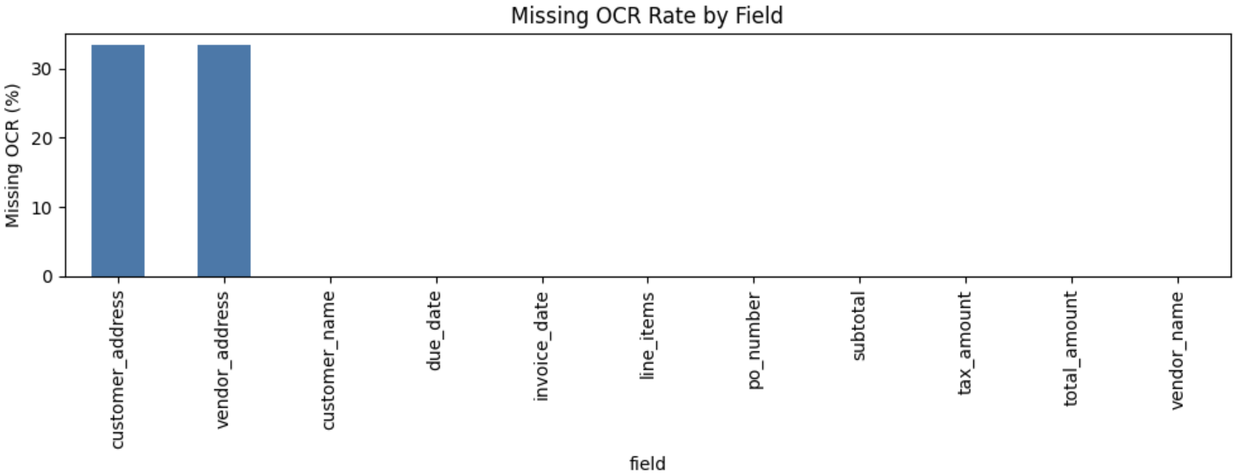
9	vendor_address	1	33.333	0.000	0.000
1	customer_name	0	0.000	0.000	0.000
2	due_date	0	0.000	33.333	66.667
3	invoice_date	0	0.000	0.000	0.000
4	line_items	0	0.000	0.000	0.000
5	po_number	0	0.000	100.000	100.000
6	subtotal	0	0.000	0.000	0.000
7	tax_amount	0	0.000	33.333	33.333
8	total_amount	0	0.000	33.333	66.667
10	vendor_name	0	0.000	100.000	100.000

Field-level Accuracy (Table)

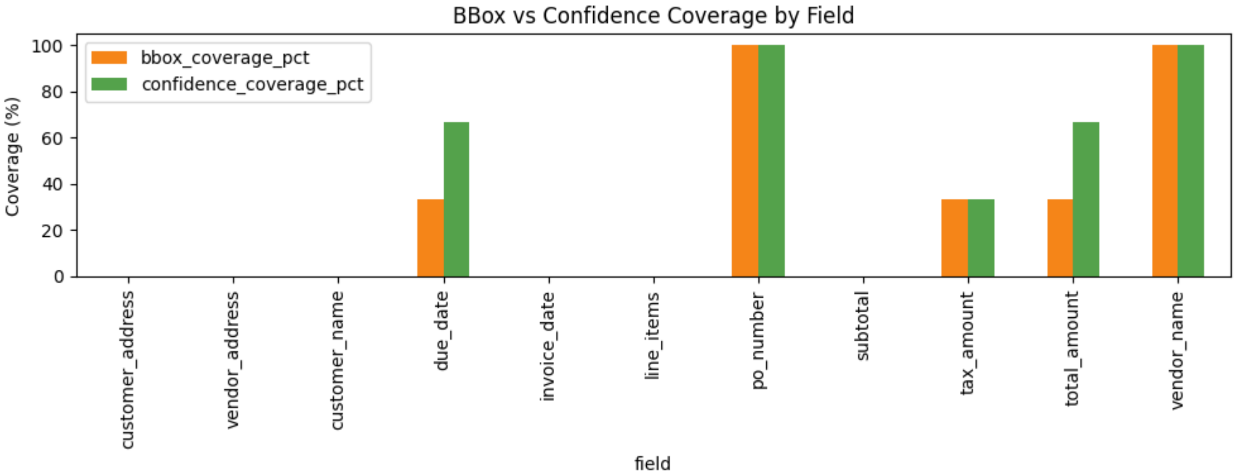
Unnamed: 0	field	accuracy	total
1	customer_name	0.667	3
2	vendor_address	0.667	3
3	customer_address	0.667	3
0	vendor_name	1.000	3
4	po_number	1.000	3
5	invoice_date	1.000	3
6	due_date	1.000	3
7	subtotal	1.000	3
8	tax_amount	1.000	3
9	total_amount	1.000	3

Data Visualizations

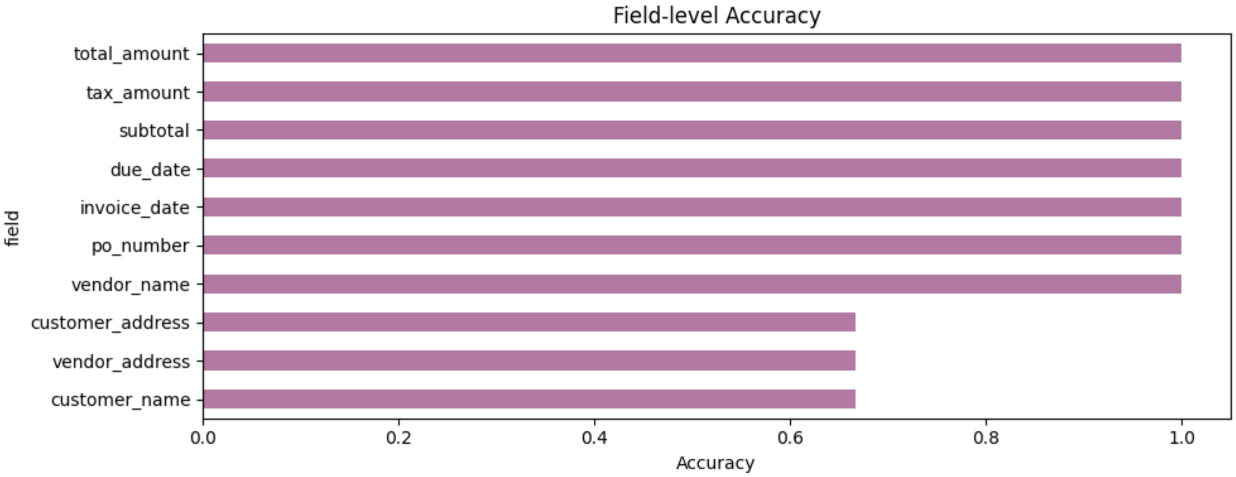
Missing OCR Rate by Field



BBox vs Confidence Coverage by Field



Field-level Accuracy


Field-level Accuracy
total_amount
tax_amount
subtotal
due_date
invoice_date
po_number
vendor_name
customer_address
vendor_address
customer_name
field
Accuracy