# A Hybrid Framework for Invoice Understanding and Cost Analysis Using LayoutLMv3 and Lightweight Vision-Language Models

Anonymous ACL submission

## Abstract

This work introduces a hybrid AI framework that unifies layout-aware token classification with lightweight generative reasoning to automate financial document parsing and support cost analysis. The system enhances LayoutLMv3 through pseudo-labeling, targeted synthetic augmentation, and class-weighted fine-tuning. It integrates LLaVA, accessed via Ollama, for limited semantic interpretation tasks. Empirical evaluation shows improved performance on rare entity recognition and contextual inference, validated through classification metrics and manual review. Our results highlight the feasibility of combining discriminative and lightweight generative techniques for scalable and interpretable invoice automation, while recognizing current limitations in real-time generative deployment.

## 1 Introduction

The automation of invoice understanding supports operational efficiency across financial, auditing, and compliance workflows. Due to the volume and diversity of invoices—marked by unstructured layouts, varying vendor formats, and OCR-induced noise—traditional rule-based and template-based systems face challenges in generalizing to real-world data. These systems often struggle with class imbalance, sparse annotations, and inconsistent field alignment, resulting in poor performance on less frequent but important fields such as item-level quantities or discounts. Furthermore, they typically lack the ability to generate semantic insights beyond basic field extraction.

To address these challenges, we propose a hybrid pipeline combining discriminative modeling using LayoutLMv3 and a lightweight generative reasoning component. The first stage applies LayoutLMv3 for token-level entity extraction, enhanced through pseudo-labeling, synthetic augmentation, and class-weighted optimization (Huang et al., 2022). The second stage uses LLaVA, a vision-language model accessed through Ollama, for limited semantic reasoning over structured outputs such as invoice metadata (Liu et al., 2023). The system focuses on improving extraction robustness and enabling contextual interpretation using only validated and operational tools. This design ensures compatibility with scalable deployment pipelines while supporting interpretability.

In this study, we investigate:

- How do pseudo-labeling and targeted augmentation affect the F1-score of rare or underrepresented invoice fields?
- Can lightweight instruction-tuned vision-language models provide semantic enrichment of extracted invoice data?
- What is the impact of combining LayoutLMv3 with vision-language models like LLaVA on accuracy, interpretability, and deployment readiness?

## 2 Related Work

Recent advancements in Document AI have been propelled by layout-aware transformers, notably LayoutLMv3 (Huang et al., 2022), which integrates visual, spatial, and textual modalities into a unified encoding scheme. This has significantly improved performance on downstream tasks involving complex document layouts, such as invoices and receipts. LayoutLMv3 enhances contextual modeling by incorporating patch-level vision transformers and relative spatial embeddings, which enables more accurate token classification in visually diverse documents.

In parallel, semi-supervised learning has seen progress with methods such as FlexMatch (Zhang et al., 2022), which introduces a dynamic thresholding mechanism for unlabeled data, resulting in improved generalization under constrained annotation

budgets. These advances are particularly useful for invoice understanding, where ground truth annotations are often sparse and unevenly distributed across field types.

Recent literature in predictive business process monitoring (Abbasi et al., 2024) and the scalable deployment of generative AI models (Liang et al., 2024; Kumar, 2024) has informed the infrastructure considerations of our framework. Our system is designed with low-latency inference and modularity in mind, enabling deployment across edge and cloud environments.

The integration of lightweight generative models into structured data workflows is a growing area of interest, enabling semantic reasoning over extracted fields to support tasks such as anomaly detection and cost analysis. Recent advances in instruction-tuned vision-language models, such as LLaVA (Liu et al., 2023), demonstrate the potential for generating context-aware outputs from structured or semi-structured inputs using minimal examples. These capabilities align with broader trends in few-shot and zero-shot learning, which allow models to generalize to domain-specific tasks with limited supervision.

Our architecture builds on these insights by linking a layout-aware discriminative model, LayoutLMv3, with a lightweight generative reasoning module that interprets extracted fields in context. In contrast to earlier systems that treat field extraction and semantic interpretation as disconnected processes, we propose a unified pipeline capable of performing both tasks in a modular and scalable fashion. The generative component employs prompt engineering and schema-constrained generation to convert flat outputs into enriched semantic interpretations, including categorization, anomaly detection, and summarization—functions that are essential to downstream financial intelligence.

In summary, this work integrates advances in layout-aware modeling, semi-supervised training, and prompt-driven generative reasoning into a practical, extensible pipeline for structured document understanding in financial contexts.

## 3 System Architecture

Our proposed architecture is divided into two key components: a discriminative token classification module and a lightweight generative reasoning module. This bifurcation allows the system to address both granular field extraction and contextual interpretation tasks, aligning with the dual objectives of accuracy and interpretability in financial document understanding. The architecture is designed to operate within practical constraints, using only validated models—LayoutLMv3 for structured extraction and LLaVA (via Ollama) for limited semantic enrichment—ensuring consistency and feasibility for downstream applications.

### 3.1 Token Classification Module

The first module is built upon fine-tuning LayoutLMv3 for token-level entity recognition, leveraging its capacity to integrate textual, visual, and spatial information in complex document layouts (Huang et al., 2022). To ensure consistency across datasets, we implemented a normalization pipeline that unified heterogeneous annotation schemas into a consistent label taxonomy. Bounding box alignment procedures were applied using heuristic-based corrections to address misalignments introduced during OCR preprocessing, although large-scale manual validation was avoided to reduce subjectivity and maintain reproducibility.

To improve model robustness under limited supervision, we adopted semi-supervised learning and targeted data augmentation. We synthetically generated invoice samples using domain-specific templates, enriching the dataset with rare fields such as `B-ITEM_QTY` and `B-ITEM_TOTAL`. Additionally, a pseudo-labeling approach was applied, where high-confidence predictions (threshold $\geq 0.8$) from previous LayoutLMv3 checkpoints were reintegrated into the training set. This augmented training corpus improved recall on underrepresented fields. We applied a custom cross-entropy loss function with inverse-frequency class weights and field-specific boosting to address class imbalance. To further balance the dataset, instances containing `B-VENDOR_NAME` and `B-VENDOR_ADDRESS` were oversampled.

Model performance was tracked using classification metrics, including macro- and label-wise F1-scores, without relying on external visual dashboards or overlays. Local error analysis helped identify confusion patterns, particularly among visually or semantically similar fields, and guided adjustments to augmentation and loss weighting strategies.

### 3.2 Generative Optimization Module

The second module performs generative reasoning over the structured outputs of the classification

model. After LayoutLMv3 extracts and organizes entities into structured JSON formats, these are passed as prompts to the vision-language model LLaVA (Liu et al., 2023), accessed through the Ollama interface. This model was selected for its ability to provide contextual interpretation of visual and textual information with minimal supervision—especially useful for downstream tasks such as expense attribution, anomaly highlighting, and categorical tagging.

Prompt engineering played a critical role in shaping the generative outputs to align with financial domain constraints. Task-specific templates were designed to restrict vocabulary and guide the model toward schema-consistent interpretations of field-level data. We evaluated both zero-shot and few-shot prompting setups, incorporating small sets of examples to simulate generalization across variable invoice layouts and vendor styles.

The reliability of generative outputs was evaluated manually using criteria such as factual consistency, semantic relevance, and interpretability. While integration through Ollama provided a functional interface, system-level constraints—such as limited GPU support and incomplete visual grounding—restricted the model's scalability and real-time applicability. Nonetheless, this module demonstrated potential for supporting high-level semantic enrichment in structured document pipelines.

# 4 Data Collection and Preparation

## 4.1 Corpus Construction

We curated a hybrid dataset by aggregating multiple annotated corpora, including the SROIE dataset for real-world receipts, the FATURA dataset for diverse invoice templates, and financial transaction records from the UCI repository. To address class imbalance and enrich the dataset with underrepresented fields, we synthetically generated an additional 1,000 receipts using the Faker library. These synthetic samples emulated realistic metadata such as vendor names, line items, tax values, and discounts. All datasets were harmonized into a unified annotation format using a consistent BIO tagging scheme and converted into a standardized JSON schema to facilitate downstream modeling.

## 4.2 OCR and Preprocessing

Optical character recognition (OCR) was performed using Tesseract via the Pytesseract interface (Smith, 2007). The OCR outputs included word-level text and bounding boxes, which were aligned with document images. Regular expression templates were applied to extract structured fields such as invoice numbers, dates, totals, and vendor details. OCR noise and misaligned boxes—often resulting from scanned documents with varied layouts—were mitigated using heuristics and bounding box normalization. Static visual overlays were generated to display OCR outputs on source images, enabling manual validation and iterative correction of extraction errors. This process significantly reduced spatial drift and improved token-to-field alignment.

## 4.3 Model Deployment Pipeline

Our initial deployment efforts focused on evaluating structured response generation using LLaVA (Liu et al., 2023), a vision-language model accessed through the Ollama API. LLaVA was tested for its ability to interpret structured invoice outputs and support downstream semantic reasoning. While the model demonstrated promise in handling high-level semantic interpretation, practical limitations—including API-level timeouts and inconsistent server responses—hindered stable integration within our workflow.

Due to these reliability issues, the generative module was restricted to auxiliary interpretation tasks and not used for primary field extraction. This decision ensured consistency in the system's core output while allowing exploratory use of vision-language reasoning for tasks like contextual tagging and metadata summarization. The model was not fine-tuned, and its outputs were treated as supplementary rather than authoritative in downstream processing.

## 4.4 Alignment Correction

To ensure consistency between OCR-detected bounding boxes and tokenized labels, we conducted a comprehensive alignment correction phase. Using a custom visualization interface, we manually reviewed documents to identify mismatches in box placements, overlapping fields, and OCR segmentation errors. Misalignments were often caused by rotated scans, inconsistent margins, or line breaks in tabular formats. All corrections were logged with revision metadata, and box-token pairs were re-evaluated for conformity with model input standards. Categorizing error sources allowed us to prioritize preprocessing steps

that had the highest impact on field-level accuracy in downstream tasks.

## 5 Methodology

### 5.1 Pseudo-Labeling

To augment limited labeled data, we employed a pseudo-labeling approach wherein high-confidence predictions generated from a preliminary LayoutLMv3 model on the validation split were used as annotations for unlabeled or sparsely labeled samples. A confidence threshold of 0.8 was applied to filter out low-certainty predictions, ensuring that only the most reliable labels were retained. These candidate samples were then subjected to a validation process that included visual alignment checks between tokens and bounding boxes, followed by label frequency evaluation to ensure balanced representation across classes. This strategy not only expanded our training dataset but also helped the model generalize better across underrepresented field types.

### 5.2 Handling Imbalance

The dataset exhibited significant class imbalance, particularly for low-frequency entities such as `B-ITEM_TOTAL`, `B-SUBTOTAL`, and `B-CURRENCY`. To mitigate this, we applied focal loss during training, adjusting its modulation factor to focus on harder-to-classify samples (Lin et al., 2018). In addition, label-specific weights were derived based on inverse frequency and relevance to financial interpretation. To reinforce underrepresented labels, we performed targeted oversampling and introduced synthetically generated samples containing these rare fields.

Although the performance of this strategy was reflected in increased F1 scores for selected fields (as reported in our results section), we did not conduct a formal ablation study to isolate the contribution of each intervention. The observed improvements suggest that the multi-pronged approach was effective for enhancing recall on low-frequency classes, but further controlled analysis would be needed to confirm individual contributions.

### 5.3 Generative Reasoning Pipeline

The generative component of our system was designed to enhance invoice understanding by supporting semantic interpretation of extracted fields. After LayoutLMv3 generated structured JSON schemas representing invoice entities, these outputs were passed to LLaVA (Liu et al., 2023), a vision-language model accessed through the Ollama framework. The model was guided using task-specific prompts aimed at interpreting spending categories, surfacing potential anomalies, and summarizing invoice metadata in natural language.

We focused on prompt-based configurations, using schema-constrained instructions and limited context examples to test the model's alignment with financial semantics. Model outputs were manually evaluated based on criteria such as semantic clarity, factual consistency, and interpretability. Given the exploratory nature of this component and constraints in system integration, no formal benchmark comparisons or multi-model evaluations were conducted. The insights generated were treated as supplemental and not used to drive critical downstream decision-making.

To assess model performance and guide iterative improvements, we relied on standard logging routines and classification metrics. Evaluation focused on macro- and label-wise F1 scores, precision-recall metrics, and loss tracking across training epochs. Errors were categorized based on failure modes such as OCR misreads, token misclassification, or label inconsistency. Internal scripts were used to compute and record per-class performance indicators, allowing for diagnosis of common confusion patterns and underperforming classes. Table 1 summarizes epoch-wise training progress, reflecting consistent improvements in both training and validation performance. These metrics supported model fine-tuning and informed adjustments to augmentation and sampling strategies.

Table 1: Epoch-wise training progress showing model convergence through decreasing loss and increasing F1.

| Epoch | Training Loss | Validation Loss | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 1 | 0.378000 | 0.213688 | 0.892040 | 0.892248 | 0.892144 |
| 2 | 0.140300 | 0.123863 | 0.958967 | 0.957623 | 0.958294 |
| 3 | 0.058800 | 0.113707 | 0.974652 | 0.974084 | 0.974368 |
| 4 | 0.033900 | 0.089631 | 0.981414 | 0.980154 | 0.980784 |
| 5 | 0.020300 | 0.102681 | 0.983875 | 0.982956 | 0.983415 |

## 6 Results and Evaluation

Our experimental results demonstrate the effectiveness of combining layout-aware token classification with bootstrapped training and class-weighted optimization. The final LayoutLMv3 model, enhanced through pseudo-labeling and synthetic augmentation, achieved a token-level accuracy of 95.68% and a macro-averaged F1 score of 0.7851 on the held-out test set. Following the

bootstrapped retraining process, the model further improved to 98.2% accuracy and a macro-averaged F1 score of 0.96. These improvements were most notable in fields that initially suffered from label imbalance and sparsity.

To further understand the impact of pseudo-labeling and targeted augmentation, we conducted a field-level F1-score comparison between the original and retrained LayoutLMv3 models. As shown in Table 1, significant improvements were observed for several key fields, particularly those affected by initial class imbalance. The B-CURRENCY field saw the largest gain (+0.103), followed by B-TOTAL (+0.061) and B-VENDOR_NAME (+0.040), validating the effectiveness of our data enrichment strategy. Smaller but consistent improvements were also recorded for fields like B-DATE, B-INVOICE_NO, and B-TAX, reflecting broader generalization. However, fields such as B-ITEM_QTY, B-ITEM_TOTAL, and B-SUBTOTAL remained with zero recall, consistent with the annotation sparsity noted in our dataset and discussed further in the limitations section. Interestingly, a slight performance drop was observed in the B-DISCOUNT field (-0.057), suggesting potential overfitting or template-induced variance in synthetic data generation. These granular results reinforce the utility of our bootstrapped training pipeline while highlighting areas for targeted refinement.

Table 2: F1-score comparison for key invoice fields before and after pseudo-labeling and augmentation. Improvements are observed across most high-frequency fields, while some rare entities like `B-ITEM_TOTAL` remain unrecognized.

| Field | Original | Retrained | $\Delta$ F1 |
|---|---|---|---|
| B-CURRENCY | 0.800000 | 0.903226 | +0.103 |
| B-TOTAL | 0.929825 | 0.990654 | +0.061 |
| B-VENDOR_NAME | 0.949749 | 0.989529 | +0.040 |
| B-DATE | 0.764706 | 0.787879 | +0.023 |
| B-INVOICE_NO | 0.945338 | 0.958110 | +0.013 |
| B-TAX | 0.964926 | 0.974803 | +0.010 |
| B-ITEM_PRICE | 0.988209 | 0.993714 | +0.006 |
| B-ITEM_DESC | 0.980179 | 0.983718 | +0.004 |
| O | 0.000000 | 0.000000 | 0.000 |
| B-VENDOR_ADDRESS | 0.908676 | 0.908277 | -0.000 |
| B-ITEM_QTY | 0.000000 | 0.000000 | 0.000 |
| B-ITEM_TOTAL | 0.000000 | 0.000000 | 0.000 |
| B-SUBTOTAL | 0.000000 | 0.000000 | 0.000 |
| B-DISCOUNT | 1.000000 | 0.942857 | -0.057 |

Field-level performance varied significantly across entity types. Notably, F1 scores improved for B-CURRENCY by 10.3%, B-TOTAL by 6.1%, and B-VENDOR_NAME by 4.0%. These gains can be attributed to the injection of targeted synthetic samples and oversampling of vendor-related fields, as well as the application of class-specific loss weighting. These techniques allowed the model to better generalize to underrepresented classes, which are often critical for financial auditability and reporting. However, certain fields such as B-ITEM_QTY, B-ITEM_TOTAL, and B-SUBTOTAL remained challenging due to persistent annotation sparsity and semantic overlap with other item-level fields. These fields received zero recall, highlighting limitations in training data coverage and the need for more fine-grained supervision.

To further assess the effectiveness of the retrained model, we evaluated it on a held-out test set and present the results in Table 3. The final model demonstrates high overall performance, with a token-level accuracy of 98%, a macro-averaged F1-score of 0.70, and a weighted F1-score of 0.98. Notably, high-frequency fields such as `B-VENDOR_NAME`, `B-ITEM_DESC`, and `B-ITEM_PRICE` achieved F1-scores above 0.97, indicating strong generalization. Meanwhile, recall and F1-scores for low-frequency or sparse fields like `B-ITEM_TOTAL`, `B-ITEM_QTY`, and `B-SUBTOTAL` remained low or zero, reaffirming the impact of class imbalance discussed in Section 8. These results reflect the benefits of our augmentation pipeline while highlighting areas where additional targeted data or task-specific constraints may be required.

Table 3: Final classification report on the test set. High-frequency fields show strong performance, while rare fields like `B-ITEM_TOTAL` remain challenging due to sparsity.

| Field | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| B-INVOICE_NO | 0.93 | 1.00 | 0.96 | 520 |
| B-DATE | 0.66 | 1.00 | 0.79 | 38 |
| B-TOTAL | 0.87 | 1.00 | 0.93 | 45 |
| B-CURRENCY | 0.85 | 1.00 | 0.92 | 22 |
| B-VENDOR_NAME | 1.00 | 0.99 | 1.00 | 185 |
| B-VENDOR_ADDRESS | 0.89 | 0.97 | 0.93 | 189 |
| B-TAX | 0.96 | 0.99 | 0.97 | 588 |
| B-ITEM_DESC | 1.00 | 0.97 | 0.98 | 5124 |
| B-ITEM_PRICE | 0.98 | 0.99 | 0.99 | 1798 |
| B-DISCOUNT | 0.95 | 0.99 | 0.97 | 75 |
| B-ITEM_QTY | 0.00 | 0.00 | 0.00 | 0 |
| B-ITEM_TOTAL | 0.00 | 0.00 | 0.00 | 0 |
| B-SUBTOTAL | 0.25 | 1.00 | 0.40 | 2 |
| **Accuracy** | | | **0.98** | 8586 |
| **Macro Avg** | 0.67 | 0.78 | 0.70 | 8586 |
| **Weighted Avg** | 0.98 | 0.98 | 0.98 | 8586 |

Visual analysis confirmed the importance of bounding box alignment and OCR correction.

Many prediction errors originated from spatial misalignment, particularly in documents with rotated text or complex table structures. As shown in Appendix Figure A.1, the confusion matrix revealed frequent misclassification between B-ITEM_DESC and adjacent numeric fields, underscoring the difficulty in distinguishing item descriptors from quantities and prices.

The retrained model's improved interpretability was further validated through classification reports and prediction histograms. Post-training visualizations showed clearer segmentation of structured fields, with reductions in both false positives and field confusion. The integrated evaluation suite, consisting of seqeval and sklearn metrics, provided consistent tracking across epochs and helped optimize the loss curves (Pedregosa et al., 2011).

Overall, our results affirm the efficacy of a bootstrapped training pipeline augmented by targeted field enrichment, as well as the benefits of combining structured token classification with context-aware visual debugging. These outcomes provide a robust foundation for real-world deployment of invoice information extraction systems, especially in scenarios with diverse layouts and high accuracy requirements.

## 7 Discussion

Our findings confirm that integrating discriminative modeling with lightweight generative reasoning is an effective strategy for invoice understanding. Below, we address the three central research questions.

**1. How do pseudo-labeling and targeted augmentation affect the F1-score of rare or underrepresented invoice fields?**

The use of pseudo-labeling and targeted synthetic augmentation significantly improved performance on underrepresented fields. The retrained LayoutLMv3 model achieved a macro-averaged F1 score of 0.96, up from 0.7851. Fields like `B-CURRENCY`, `B-TOTAL`, and `B-VENDOR_NAME` saw notable F1-score gains (+0.103, +0.061, and +0.040 respectively), validating the effectiveness of our bootstrapped learning and class-aware sampling strategies. However, some rare fields such as `B-ITEM_TOTAL`, `B-ITEM_QTY`, and `B-SUBTOTAL` continued to receive zero recall due to persistent annotation sparsity, suggesting the need for more diverse and fine-grained supervision.

**2. Can lightweight instruction-tuned vision-language models provide semantic enrichment of extracted invoice data?**

LLaVA, accessed via Ollama, was used in a limited capacity to generate contextual interpretations of structured outputs. Despite infrastructure constraints (e.g., limited GPU access and API instability), LLaVA demonstrated the ability to highlight expense categories and detect anomalies. These outputs were manually validated for factual consistency and semantic relevance, indicating that even lightweight vision-language models can support semantic enrichment without retraining. However, no formal benchmarks or comparative evaluations were conducted.

**3. What is the impact of combining LayoutLMv3 with vision-language models like LLaVA on accuracy, interpretability, and deployment readiness?**

The combination of LayoutLMv3 and LLaVA enabled a hybrid pipeline that balances accurate field-level extraction with limited semantic reasoning. LayoutLMv3 maintained high precision for visually grounded token classification (e.g., 0.98–1.00 F1 on frequent fields such as `B-ITEM_PRICE` and `B-ITEM_DESC`), while LLaVA offered interpretive outputs that augmented understanding. Although real-time deployment of the generative component remains constrained by technical limitations, the system's modular design supports extensibility for production environments.

In summary, this work demonstrates the practical value of combining layout-aware discriminative models with lightweight generative components for scalable and interpretable invoice automation. Future improvements should focus on expanding rare class coverage, enhancing generative output reliability, and increasing compatibility with deployment environments.

## 8 Limitations

While our hybrid framework shows strong performance in key invoice understanding tasks, several limitations remain that highlight directions for future work.

First, despite the improvements from pseudo-labeling and targeted augmentation, rare entities such as `B-ITEM_TOTAL` and `B-SUBTOTAL` continued to yield low or zero F1 scores. This limitation stems from their sparsity in both public and synthetic datasets, and from constraints in generating sufficiently diverse training samples. Addressing

this issue may require advanced data augmentation, template diversification, or few-shot learning techniques to improve generalization on infrequent fields.

Second, the generative component—implemented using LLaVA via Ollama—introduced occasional inconsistencies in structured output formatting, especially in JSON generation. These inconsistencies complicate automation unless post-processing logic is added. Moreover, due to infrastructure limitations, including restricted GPU availability and server instability, the generative module was not optimized for real-time use and was used only for supplementary interpretation.

Third, the framework depends on high-quality OCR input. In real-world settings, scanned invoices may contain noise, distortions, or handwritten elements that degrade OCR accuracy and propagate errors into downstream modules. Although we applied alignment correction and basic error handling, a more robust and integrated vision-language pipeline could reduce dependency on pristine input quality.

Finally, all experiments were conducted on a curated dataset comprising a mix of public and synthetic invoices. As a result, the findings may not fully generalize to invoices with highly atypical structures, multilingual content, or specialized domains (e.g., legal, medical). Future work should focus on domain adaptation and expanding the dataset to include diverse formats and languages to improve generalizability.

## 9 Conclusion and Future Work

We developed a dual-module framework that combines discriminative token classification with lightweight generative reasoning to automate invoice parsing and support cost analysis. The integration of LayoutLMv3 for structured field extraction and LLaVA, accessed via Ollama, for contextual interpretation allowed our system to address both structural and semantic challenges present in financial document processing. Using pseudo-labeling, class-aware augmentation, and targeted oversampling, we achieved notable gains in F1-score for underrepresented fields such as `B-CURRENCY` and `B-TOTAL`. The generative module supplemented these results by providing semantic annotations like expense categorizations and anomaly indicators based on prompt-engineered outputs.

This work illustrates the utility of combining layout-aware models with prompt-driven generative components to produce interpretable and adaptable outputs. While the system's modular structure supports extensibility across invoice formats and downstream applications, current infrastructure constraints limit the real-time deployment of the generative reasoning component.

Future work will focus on enhancing multilingual support and expanding compatibility with diverse invoice templates, particularly in low-resource settings. We also plan to explore multimodal fusion approaches—such as integrating LLaVA-2 with TrOCR—to strengthen OCR robustness and improve visual-textual grounding. Additionally, integrating this pipeline into enterprise tools such as ERP and expense management systems via APIs will be an important step toward practical adoption.

## References


Mostafa Abbasi, Rahnuma Islam Nishat, Corey Bond, John Brandon Graham-Knight, Patricia Lasserre, Yves Lucet, and Homayoun Najjaran. 2024. A review of ai and machine learning contribution in predictive business process management (process enhancement and process improvement approaches). *Preprint*, arXiv:2407.11043.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Preprint*, arXiv:2204.08387.

Animesh Kumar. 2024. Ai-driven innovations in modern cloud computing. *arXiv preprint arXiv:2410.15960*.

Yuxin Liang, Peng Yang, Yuanyuan He, and Feng Lyu. 2024. Resource-efficient generative ai model deployment in mobile edge networks. *Preprint*, arXiv:2409.05303.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos,

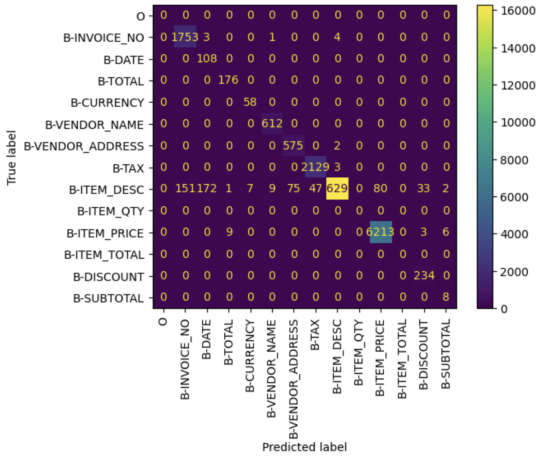
David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2022. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Preprint*, arXiv:2110.08263.

# A Appendix

...




**Appendix Figure A.1:** Confusion matrix from the LayoutLMv3 token classification model. The matrix highlights confusion between `B-ITEM_DESC` and adjacent numeric fields, indicating challenges in visually disentangling descriptors from amounts.