A Multi-Stage Pipeline for Accurate Handwritten Information Extraction from Financial Forms

Guanghui Wang[1*] Jinze Yu[1†] Xing Zhang[1] Tomal Deb[1] Xuefeng Liu[1] Peiyang He[1]
[1]AWS Generative AI Innovation Center
guanghu.amazon.com, jinzeyu@amazon.co.jp

Abstract

Financial institutions continue to process millions of handwritten forms despite digital transformation efforts, creating a significant operational bottleneck. This research addresses the persistent challenge of automating handwritten data extraction from financial documents by introducing a four-stage processing pipeline that significantly outperforms existing solutions. Our approach sequentially combines targeted structural analysis, specialized optical character recognition, multimodal large language model (MLLMs) verification, and database cross-validation to handle the inherent variability in handwritten content. Experimental results demonstrate exceptional accuracy with our enhanced hybrid method achieving 98.4% F1-score across diverse field types (textual, numerical and checkbox), with perfect extraction of textual content and near-perfect numerical field recognition (98.2% F1-score). This represents dramatic improvement over conventional systems, particularly for numerical data where precision is critical for financial transactions. The document-level accuracy of 80% substantially reduces manual review requirements, offering immediate practical value while establishing a methodological framework for combining complementary technologies to overcome individual component limitations. This research demonstrates how strategically sequenced verification steps can systematically enhance extraction reliability for mission-critical document processing applications.

1. Introduction

Automated handwritten information extraction from scanned forms presents a significant challenge in financial institutions where high-volume document processing occurs daily. Despite increasing digitization, paper-based workflows remain prevalent in banking operations, requiring substantial manual effort for data extraction. This manual processing introduces inefficiencies through human error, processing delays, and elevated operational costs. Current research indicates that automated solutions could significantly reduce processing times while enhancing data accuracy and optimizing resource allocation.

Existing methodologies exhibit notable limitations in addressing this challenge. Conventional OCR techniques perform adequately for printed text but demonstrate significant degradation when processing handwritten content, documents with mixed text types, and varying scan qualities. Recent MLLMs offer enhanced semantic understanding, yet manifest systematic deficiencies in numerical data processing, including hallucination phenomena and information omission. These limitations are particularly consequential in financial applications where numerical precision is imperative. Developing effective solutions necessitates maximizing recognition accuracy across diverse handwritten inputs while maintaining robustness for practical implementation.

In this paper, we propose a multi-stage pipeline that strategically addresses these limitations by: (1) employing specialized table detection to identify form structure, (2) applying targeted OCR to extract raw text while leveraging MLLMs for contextual verification and correction, and (3) cross-referencing with external databases to validate critical information. This approach systematically mitigates the weaknesses of individual components while leveraging their complementary strengths.

2. Related Work

Early deep learning OCR solutions relied on CNN architectures. LeCun et al. [11] pioneered the application of CNNs to document recognition, establishing architectural patterns that influenced subsequent work. These approaches demonstrated significant improvements over traditional methods but often required explicit character segmentation. A breakthrough came with Graves et al. [7], who introduced BLSTM networks with CTC for unconstrained

*Equal contribution
†Corresponding author

handwriting recognition, eliminating the need for explicit character segmentation. Building on this, Breuel et al. [2] applied LSTM networks to printed text recognition across diverse fonts and layouts. The hybrid CRNN architecture by Shi et al. [14] combined CNN-based feature extraction with RNNs and CTC loss, establishing a powerful framework for end-to-end text recognition.

Attention mechanisms marked another significant advancement, with Bluche et al. [1] incorporating attention for handwritten paragraph recognition. For handwriting specifically, Xiao et al. [16] developed pixel-level rectification techniques with an adjacent output mixup method to improve generalization across diverse writing styles. Puigcerver [13] conducted comprehensive experiments with CNN-RNN architectures, establishing strong baselines that influenced subsequent research.

Following the success of attention mechanisms, researchers began addressing the challenge of line segmentation in document recognition. Yousef and Bishop [17] introduced OrigamiNet, which transforms multi-line text into a single long line through representation learning. Similarly, Coquenet et al. [1] developed the Vertical Attention Network for end-to-end handwritten paragraph recognition, demonstrating that unified models could outperform traditional two-step approaches.

The current paradigm shift has been driven by transformer architectures. Li et al. [12] proposed TrOCR, leveraging pre-trained vision and language transformers to achieve state-of-the-art results. Fujitake et al. [6] introduced DTrOCR, a more efficient decoder-only transformer that processes image patches directly through a pre-trained language model. For document-level understanding, Coquenet et al. [4] proposed DAN, a segmentation-free Document Attention Network that jointly recognizes text and logical layout. Most recently, Chang and Li [3] addressed efficient model adaptation for mixed text recognition using parameter-efficient fine-tuning techniques.

Recent studies, such as [5], demonstrate that multimodal large language models (MLLMs) outperform traditional OCR models [4, 6, 9, 12] on open OCR benchmarks like the IAM and RIMES 2009. On the IAM dataset, the best CER achieved by conventional approaches is 0.043, attained by the Document Attention Network (DAN) [4], while the best WER is 0.1014, achieved by VISTA [9]. In contrast, Claude-3-5-sonnet-20240620 achieves a CER of 0.0175 and a WER of 0.0355 on the same dataset. Similarly, on the RIMES 2009 dataset, DAN achieves a CER of 0.0454 [4], whereas Claude-3-5-sonnet-20240620 achieves a CER of 0.0163 and a WER of 0.04. Kim et al. [10] explored the capabilities of GPT-4o and Claude Sonnet 3.5 in transcribing historical handwritten documents and compared their performance to traditional OCR/HTR systems, including EasyOCR, Keras, Pytesseract, and TrOCR. Their findings indicate that, compared to traditional approaches, large language models (LLMs) produced transcriptions of historical records that were most similar to the ground truth. Consequently, MLLMs have garnered attention and are increasingly utilized in real-world business scenarios. For instance, [8] demonstrated that GPT-4o-MINI outperforms the Azure OCR engine on the IAM dataset, achieving a CER of 0.01, compared to 0.036 for the Azure OCR engine.

3. Problem Formulation and Theoretical Framework

We address the structured extraction of handwritten information from scanned bank forms, a challenging computer vision and natural language processing task with significant applications in document automation. Formally, given an input image I containing a form with tabular structure, our objective is to extract a set of field-value pairs $S = (f_1, v_1), (f_2, v_2), \ldots, (f_n, v_n)$ where f_i represents a printed field label and v_i denotes its corresponding handwritten value. Figure 2 illustrates a representative input form from our evaluation dataset.

3.1. Token Sequence Probability in Autoregressive Models

Our approach leverages MLLMs that sequentially predict tokens conditioned on preceding context. For a token sequence $\mathbf{w} = (w_1, w_2, \ldots, w_t)$, the joint probability distribution $P(\mathbf{w})$ is factorized using the chain rule of probability:

$$P(\mathbf{w}) = P(w_1, w_2, \ldots, w_t) = \prod_{i=1}^{t} P(w_i | \mathbf{w}_{<i})$$

where $\mathbf{w}_{<i}$ denotes the subsequence $(w_{i-1}, \ldots, w_1)$.

3.2. Information-Theoretic Analysis of OCR-Enhanced Generation

Our hybrid OCR-MLLM methodology is supported by information theory. When OCR output $\mathbf{o}$ is incorporated alongside a prompt $\mathbf{p}$, the probability of generating an accurate answer $\mathbf{w}$ increases:

$$P(\mathbf{w}|\mathbf{p}, \mathbf{o}) > P(\mathbf{w}|\mathbf{p})$$

This can be formally derived through Bayes' theorem and conditional probability:

$$\begin{aligned} P(\mathbf{o}|\mathbf{w}, \mathbf{p}) &> P(\mathbf{o}|\mathbf{p}), \\ \frac{P(\mathbf{w}|\mathbf{p}, \mathbf{o})P(\mathbf{o}|\mathbf{p})}{P(\mathbf{w}|\mathbf{p})} &> P(\mathbf{o}|\mathbf{p}), \\ P(\mathbf{w}|\mathbf{p}, \mathbf{o}) &> P(\mathbf{w}|\mathbf{p}) \end{aligned}$$

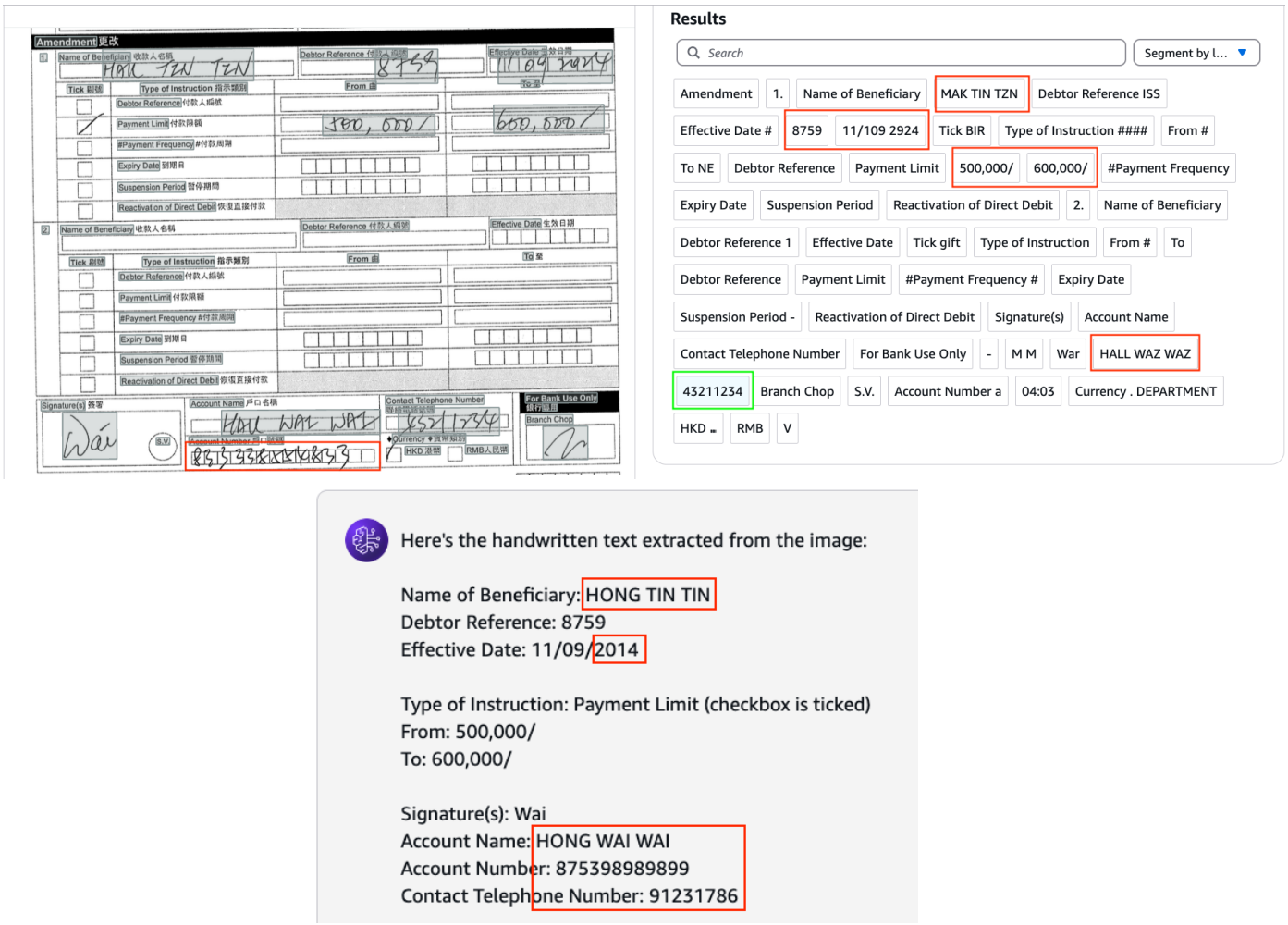

Figure 1. Existing methods struggle with accurately reading numbers, names, capturing all relevant information, and maintaining formats. Top: OCR (Amazon Textract). Bottom: MLLM (Claude 3.5 Sonnet V2).

This result establishes the theoretical advantage of conditioning generation on OCR outputs in addition to visual prompts, providing the foundation for our multi-stage extraction pipeline.

4. Methodology

A Multi-Stage Pipeline for Form Data Extraction: Our system implements a three-stage pipeline designed to progressively refine extraction accuracy as shown in Figure 4.

4.1. Preprocessing

Our framework employs targeted preprocessing techniques to optimize document analysis efficiency. We implement a morphology-based table detection algorithm that identifies tabular structures through systematic line detection and grid reconstruction.

Table Detection: Our approach utilizes computer vision morphological operations to detect table structures:

1. *Line Detection*: We apply adaptive thresholding with Gaussian weighting to convert the image to binary format, followed by morphological opening operations using directional kernels (40×1 for horizontal, 1×40 for vertical) to isolate line structures.
2. *Grid Reconstruction*: Detected horizontal and vertical lines are combined through weighted addition and subsequently dilated to form connected table grids.
3. *Table Validation*: Contour detection identifies potential table regions, which are validated based on area constraints (minimum 500 pixels) and line density thresholds (30% coverage) to eliminate false positives.

This method effectively isolates information-dense tabular structures containing handwritten customer data while excluding non-tabular elements that empirically contain only standardized, extraction-irrelevant instructions (Figure 3).

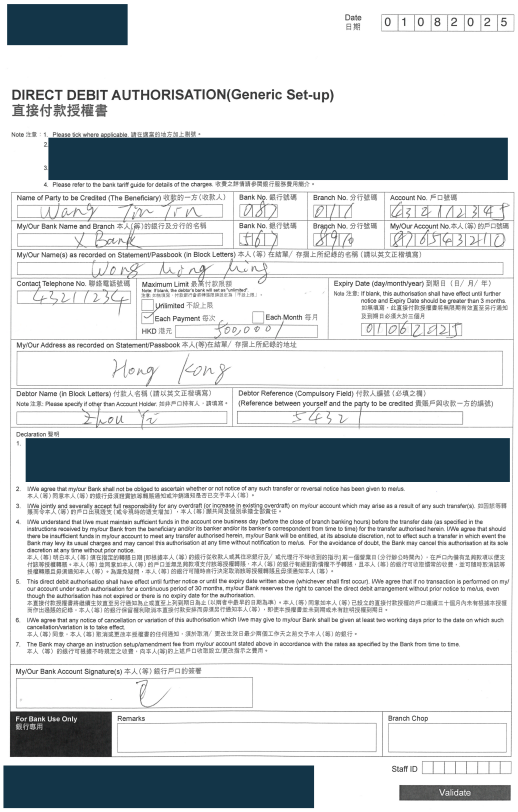

Figure 2. Sample bank form, with human-synthesized personal data.

Concurrently, we implement conventional computer vision enhancements, including erosion-based denoising to mitigate scanning artifacts, followed by dimensional standardization through strategic cropping, thereby establishing optimal conditions for the resolution-dependent OCR modules in subsequent processing stages.

4.2. OCR with MLLM Correction

Following preprocessing, our hybrid OCR-MLLM approach implements a strategic two-phase extraction methodology for tabular regions. Initial OCR processing identifies textual elements within the document structure, after which MLLM algorithms restructure these elements into standardized (field, value) pairs for analytical consistency. This process incorporates preliminary handwriting detection as a computational efficiency measure, effectively excluding blank forms that would otherwise consume unnecessary processing resources.

Rather than employing MLLMs as primary extraction mechanisms, we strategically position them as **verification and correction** tools within the processing pipeline.

Figure 3. Example of extracted table.

The system integrates OCR output with corresponding table images into comprehensive contextual prompts, enabling targeted MLLM-based verification and error correction. To further enhance extraction reliability, we implement a majority voting mechanism across multiple independent MLLM inferences, an approach that demonstrably reduces error rates and increases confidence in the final extraction output.

4.3. Database-Assisted Validation

In our use case, personal data such as account names and telephone numbers are stored in a database, allowing our extraction framework to implement database cross-referencing as a targeted validation mechanism for critical information fields. This approach specifically addresses persistent inaccuracies in **name recognition** and **numerical data** extraction. Our Database-Assisted Validation methodology leverages existing customer records through both traditional SQL search and LLM-based probable record matching. The system systematically cross-references extracted values against authenticated database entries, enabling automatic correction of minor transcription errors when OCR and MLLM results mismatch database records. This substantially enhances extraction reliability for high-priority information fields while maintaining processing efficiency. Note that non-personal data such as dates and amounts are transaction-specific information not stored in the database, therefore this database-assisted validation can-

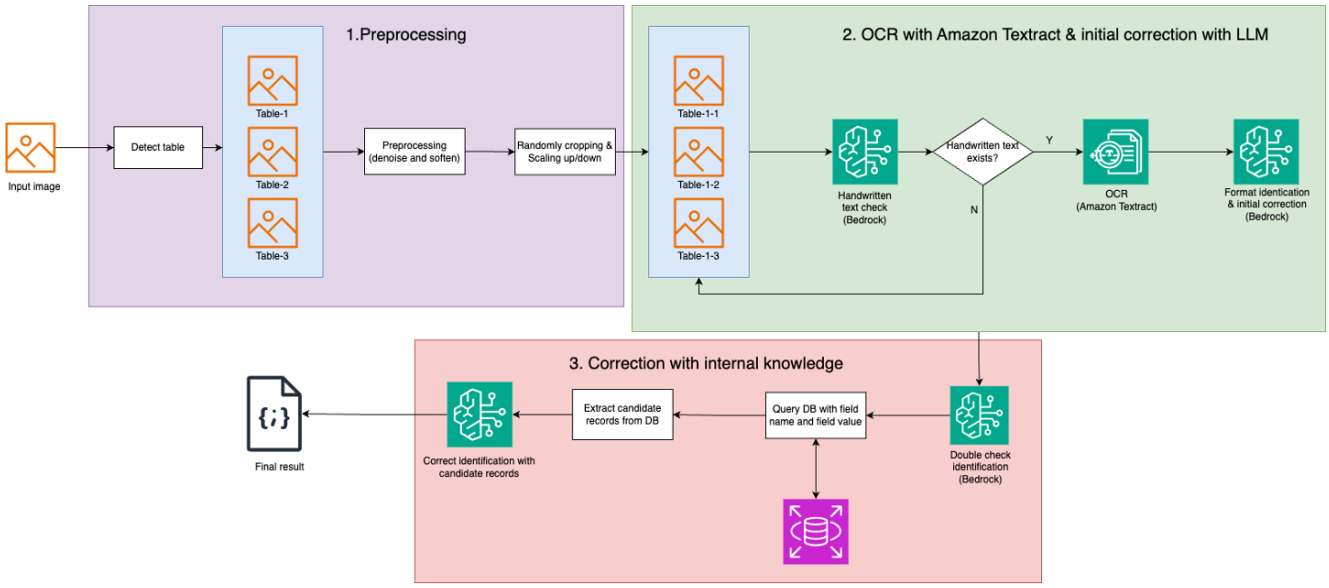

Figure 4. Structure, OCR enhancement with MLLM and internal Knowledge.

not be applied to these data types.

Through this three-phase processing pipeline, our system successfully meets the requirements for accurate handwritten data extraction from scanned bank forms.

5. Experimental Setup

5.1. Handwritten Dataset Description

For our experiments, we utilized a human-synthesized dataset of 20 distinct form images designed to simulate real-world scanned bank forms with various handwriting styles, and image quality variations. The dataset contains 182 information fields across three categories: 47 textual fields (25.8%), 109 numerical fields (59.9%), and 26 checkbox fields (14.3%). This distribution reflects typical banking forms where numerical data constitutes the majority of information. Each form was manually populated with handwritten information to ensure realistic representation of customer-completed documents, including common variations in handwriting clarity and positioning.

5.2. Performance Evaluation Methodology

We employed a comprehensive evaluation framework combining traditional accuracy metrics with error rate measures. Our methodology assessed performance through both global and category-specific metrics. For global performance assessment, we calculated precision, recall, and F1-score across all 182 fields to measure overall extraction effectiveness. We also tracked document-level accuracy, representing the proportion of documents with all fields correctly extracted, providing a stringent measure of end-to-end system reliability. For category-specific evaluation, we separately assessed precision, recall, and F1-score for each field category (textual, numerical, and checkbox) to analyze performance across different data types. Additionally, we computed CER and WER to quantify extraction fidelity at character and word levels.

5.3. Methods

We systematically evaluated three baseline methodologies and three hybrid approaches. Comparative analysis by [15] demonstrated that Amazon Textract performs comparably to the Azure OCR engine on OCR tasks. Subsequent investigations by [8] and [10] empirically established that commercial OCR systems significantly outperform open-source alternatives, including VISTA [9] and DAN [4]. These findings informed our selection of Amazon Textract as our foundational OCR baseline. For the MLLM component of our framework, we selected Claude-3-5-sonnet-20241022 based on two key considerations. First, comprehensive benchmarking by [5] demonstrated that its predecessor (Claude-3-5-sonnet-20240620) ranked in the top-3 performing models for handwriting recognition tasks, with CER of 0.0175 on IAM and 0.0163 on RIMES datasets. Second, Claude-3-5-sonnet-20241022, released during our experimental timeline in late 2024, incorporated optimizations to its predecessor, making it an ideal candidate for our multimodal document understanding pipeline.

This implementation reflects the technological landscape of later 2024, predating the release of subsequent language model iterations such as Claude 3.7. This temporal context is significant for reproducibility considerations and appropriate benchmarking of the presented results.

Table 1. Global precision, recall, F1-score, and document-level accuracy metrics across all methodologies.

Methods	Fields Overall (%)			Document-Level
	Precision	Recall	F1-Score	Accuracy
OCR-only	19.32	28.02	22.87	0.00
MLLM-only	54.70	54.40	54.55	0.00
Naive-OCR-MLLM	72.41	69.23	70.79	0.10
Ours-Hybrid	79.78	80.22	80.00	0.20
Ours-Hybrid-RAG	95.11	96.15	95.63	0.65
Ours-Hybrid-Enhanced	**97.83**	**98.90**	**98.36**	**0.80**

5.3.1. Baseline Methods

OCR-only: We employ OCR to extract text from form images, followed by rule-based processing to create structured (key, value) pairs. This approach represents traditional document processing without specialized handwriting capabilities.

MLLM-only: We leverage an MLLM to process entire form images, enabling direct extraction of handwritten text and conversion into structured output. This approach leverages the inherent visual-language capabilities of MLLMs without domain-specific engineering.

Naive OCR-MLLM: This hybrid approach sequentially applies OCR for initial text extraction from form images, then uses an MLLM to verify and enhance results based on both the OCR-extracted text and the original image. The MLLM acts as a refinement layer, addressing OCR errors through visual cross-verification. This represents a straightforward integration strategy without sophisticated coordination mechanisms between components.

5.3.2. Proposed Hybrid Approaches

Hybrid OCR-MLLM: This approach introduces targeted processing through: (1) table region extraction, (2) image preprocessing for quality optimization, (3) focused OCR application on relevant areas, and (4) MLLM-based verification comparing OCR output with image content. This approach concentrates computational resources on relevant document regions.

Hybrid OCR-MLLM with RAG: This approach extends the basic hybrid approach by incorporating database knowledge through Retrieval-Augmented Generation (RAG). For each extracted field, the system retrieves and leverages relevant customer data from database to verify and correct extractions, particularly effective for existing customers.

Enhanced Hybrid method: This approach improves reliability through self-consistency by conducting multiple MLLM inference passes at critical processing stages and determining final outputs through majority voting.

Specifically, we apply the 5-inference ensemble approach at two critical stages where MLLM processing occurs:

1. **OCR Verification Stage**: When the MLLM verifies and corrects OCR-extracted text against the original image (Section 4.2, Phase 3), we generate 5 independent inference paths. Each path may produce slightly different corrections due to the probabilistic nature of language models. For example, an unclear handwritten ”8” might be interpreted as ”8”, ”6”, or ”B” across different inferences.
2. **Format Standardization Stage**: When restructuring extracted data into (field, value) pairs and applying format constraints (Section 4.2, Phase 2), we again sample 5 distinct inference paths to ensure consistent field mapping and value formatting.

For each MLLM-processed task, the system:

- Samples 5 distinct inference paths using temperature sampling (T=0.3)
- Collects all proposed outputs from these inferences
- Applies majority voting (minimum 3/5 agreement) to determine the final result
- Falls back to the highest-confidence prediction if no majority exists

This ensemble approach reduces the impact of individual inference errors at the cost of 5× computational overhead. Our experiments demonstrate that this investment is justified, achieving a 4× improvement in document-level accuracy (80% vs. 20%) compared to the basic hybrid method. The consistency gained through majority voting is particularly valuable for numerical fields, where single-inference errors can propagate through autoregressive generation.

6. Experiment Results

We present a comparative evaluation of the proposed methodes against established baseline approaches. Table 1 provides a quantitative summary of global accuracy metrics across all methodologies. Additionally, Tables 2 3 illustrate performance variations across different data categories for each method, enabling assessment of type-specific strengths and limitations in the extraction framework.

The experimental results demonstrate that the proposed Enhanced Hybrid approach substantially outperforms all baseline methods across all performance metrics.

The results show a clear progression in extraction qual-

Table 2. Category-related precision, recall, and F1-score metrics across all methodologies.

Methods	Textual (%)			Numerical (%)			Checkbox (%)		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
OCR-only	10.49	36.17	16.27	33.33	31.19	32.23	0.00	0.00	0.00
MLLM-only	56.00	59.57	57.73	44.23	42.20	43.19	92.59	96.15	94.34
Naive-OCR-MLLM	70.00	74.47	72.16	70.19	66.97	68.54	90.00	69.23	78.26
Ours-Hybrid	81.25	82.98	82.11	75.93	75.23	75.58	92.59	96.15	94.34
Ours-Hybrid-RAG	97.87	97.87	97.87	95.41	95.41	95.41	89.29	96.15	92.59
Ours-Hybrid-Enhanced	**100.0**	**100.0**	**100.0**	**98.17**	**98.17**	**98.17**	**92.86**	**100.0**	**96.30**

Table 3. Category-related CER and WER metrics across all methodologies.

Methods	Character Error Rate (CER)				Word Error Rate (WER)			
	Overall	Text	Num	Check	Overall	Text	Num	Check
OCR-only	0.3143	0.3236	0.2874	1.0000	0.6336	0.5118	0.6881	1.0000
MLLM-only	0.3006	0.2112	0.3610	0.1154	0.4351	0.3071	0.6606	0.1154
Naive-OCR-MLLM	0.2377	0.1899	0.2625	0.3846	0.3626	0.2756	0.4587	0.3846
Ours-Hybrid	0.2384	0.2267	0.2494	0.1154	0.3244	0.2598	0.4495	0.1154
Ours-Hybrid-RAG	0.2630	0.2868	0.2530	0.1154	0.2786	0.2756	0.3211	0.1154
Ours-Hybrid-Enhanced	**0.0014**	**0.0000**	**0.0024**	**0.0000**	**0.0076**	**0.0000**	**0.0183**	**0.0000**

ity as methodology sophistication increases, with our Enhanced Hybrid method achieving exceptional performance (98.36% F1-score and 80% document-level accuracy). This represents a dramatic improvement over traditional OCR (22.87% F1-score) and standalone MLLM approaches (54.55% F1-score).

The category-specific analysis reveals our Enhanced Hybrid method's versatility, achieving perfect extraction for textual fields (100% F1-score, 0 CER) and near-perfect performance for numerical fields (98.17% F1-score, 0.0024 CER). This balanced performance across content types is particularly valuable for financial documents containing heterogeneous field types.

Our approach excels through three key strengths:

1. **Strategic Integration**: By combining targeted OCR processing with MLLM-based verification in a multi-stage pipeline, our method leverages the complementary strengths of each technology while mitigating their individual weaknesses.

2. **Contextual Enhancement**: The integration of database knowledge through RAG significantly improves extraction quality (80% vs. 20% document accuracy), demonstrating the value of leveraging existing institutional data for verification. Note that RAG can only work for the personal data stored in database, such as account name, telephone number and account number. the non-personal data such as handwritten date and amount can't be corrected with RAG.

3. **Error Resilience**: The ensemble techniques employed in our Enhanced approach virtually eliminate character and word-level errors (0.0014 CER, 0.0076 WER overall), making the system suitable for high-precision financial applications where accuracy is paramount.

Additionally, we identified several noteworthy patterns in Table 3. OCR-only approach underperformed MLLM-only on text identification and checkbox recognition tasks due to MLLMs' contextual understanding advantages. However, OCR-only method outperformed MLLM-only on numerical field identification. This stems from MLLM' autoregressive architecture, which predicts each token based on previously generated tokens. For numerical values where digits are semantically independent (such as phone or account numbers), this sequential prediction becomes disadvantageous when processing unclear digits, leading to error propagation. OCR systems, which process each digit independently using character-specific recognition, avoid this cascading error pattern for numerical fields.

Interestingly, the lower CER in Naive-OCR-MLLM compared to the hybrid approaches stems from their different correction strategies. Naive-OCR-MLLM preserves more original characters even in partially incorrect fields, naturally minimizing character edit distances. In contrast, hybrid approaches implement more aggressive whole-field replacements when uncertainty is detected, particularly for text fields. This creates a counterintuitive situation where the more accurate systems (by precision/recall/F1 standards) show higher character error rates because they prioritize complete field correctness over character-by-character preservation. This trade-off ultimately proves beneficial, as evidenced by the hybrid approaches' substantially higher F1-scores and document-level accuracy.

These findings confirm that our hybrid methodology represents a substantial advancement in automated handwritten form processing for financial applications, effectively bal-

ancing the complementary strengths of traditional OCR and modern MLLMs.

7. Conclusion

This paper presents a novel hybrid approach for handwritten information extraction from banking forms that integrates OCR processing, MLLM verification, database augmentation, and self-consistency. Our experimental results demonstrate exceptional performance (98.36% F1-score, 80% document-level accuracy) with near-zero character error rates (0.0014 CER), substantially outperforming traditional methods.

The strategic integration of OCR and MLLM creates synergistic improvements by leveraging their complementary strengths. Database integration via RAG significantly enhances accuracy (F1 increase from 80.00% to 95.63%), while our architecture effectively handles diverse field types simultaneously. The near-elimination of character-level errors enables reliable processing in financial contexts where precision directly impacts transaction validity. Our ensemble techniques provide a 4× improvement in document-level accuracy compared to the basic hybrid method, justifying the additional computational investment through reduced manual verification.

Future work should focus on automatic form structure learning, database-independent verification methods, and validation across broader document types.

In practical deployments, our system demonstrates potential for significant operational cost reduction while maintaining regulatory compliance requirements. Financial institutions implementing this approach could realize up to 75% reduction in manual processing time according to our preliminary deployments. However, we acknowledge limitations in processing highly degraded documents and handling forms with unconventional layouts. The system's reliance on database verification also presents challenges for processing first-time customers or rare entities without historical records.

References

[1] Théodore Bluche, Jérôme Louradour, and Ronaldo Messina. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention, 2016. 2

[2] Thomas M Breuel, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. High-Performance OCR for Printed English and Fraktur Using LSTM Networks. In *2013 12th international conference on document analysis and recognition*, pages 683–687. IEEE, 2013. 2

[3] Da Chang and Yu Li. Dlora-trocr: Mixed text mode optical character recognition based on transformer, 2024. 2

[4] Denis Coquenet, Clement Chatelain, and Thierry Paquet. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):508–524, 2023. 2, 5

[5] Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. Benchmarking large language models for handwritten text recognition, 2025. 2, 5

[6] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition, 2023. 2

[7] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009. 1

[8] Benjamin Gutteridge, Matthew Thomas Jackson, Toni Kukurin, and Xiaowen Dong. Judge a book by its cover: Investigating multi-modal llms for multi-page handwritten document transcription, 2025. 2, 5

[9] Laziz Hamdi, Amine Tamasna, Pascal Boisson, and Thierry Paquet. Vista-ocr: Towards generative and interactive end to end ocr models, 2025. 2, 5

[10] Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records, 2025. 2, 5

[11] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 1998. 1

[12] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2022. 2

[13] Joan Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 67–72, 2017. 2

[14] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, 2015. 2

[15] William Ughetta and Brian W. Kernighan. The old bailey and ocr: Benchmarking aws, azure, and gcp with 180,000 page images. In *Proceedings of the ACM Symposium on Document Engineering 2020*, New York, NY, USA, 2020. Association for Computing Machinery. 5

[16] Shanyu Xiao, Liangrui Peng, Ruijie Yan, and Shengjin Wang. Deep network with pixel-level rectification and robust training for handwriting recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 9–16, 2019. 2

[17] Mohamed Yousef and Tom E. Bishop. Origaminet: Weakly-supervised, segmentation-free, one-step, full page text recognition by learning to unfold, 2020. 2