Advances in Medical Sciences and Research
Vol. 1 (2026), No. 3, pp. 171-178
DOI: https://doi.org/10.55578/amsr.2607.011

ARTICLE

Cascaded Intelligence for High-Speed Medical Document Processing

Ato Kasymov[1], Fedor Krasnov[1,*]

[1] *Zentist, Inc., San Francisco, California, United States*

*Corresponding author. Email: fedor.krasnov@zentist.io

Received: 22 May 2026, Revised: 11 June 2026, Accepted: 15 June 2026, Published: 1 July 2026

Abstract

This paper presents a robust, high-throughput architectural framework for automated information extraction from Explanation of Benefits (EOB) documents. In the medical billing industry, the extreme variability of document formats poses a significant challenge for traditional automation. Our research demonstrates that conventional approaches, such as heuristic-based systems and prompt-based Large Language Model (LLM) extraction, fail to achieve the precision required for automated financial posting due to spatial misalignments and neural hallucinations. To address these limitations, we propose a hybrid cascaded pipeline that integrates specialized models for page splitting, layout classification, and field-specific normalization. Furthermore, we implement a multi-layer financial verification engine to ensure arithmetic integrity. The system is deployed in a production environment, processing over 300,000 documents per month, achieving a 95.8% first-pass resolution rate (FPRR) and reducing charge lag by 81%.

Keywords: Information Extraction; Document AI; Cascaded Intelligence; Revenue Cycle Management; Financial Automation; Hybrid Neural Models

1. INTRODUCTION

The modern healthcare system operates under increasing financial pressure, driven by rising operational costs and declining reimbursement rates. Revenue Cycle Management (RCM), defined as the set of administrative and financial processes required to capture, manage, and collect patient service revenue, has evolved from a routine billing function into a complex data-intensive workflow [1]. In the United States alone, claim denials are estimated to cost hospitals and clinical practices approximately $262 billion annually, highlighting the scale of inefficiencies in the healthcare reimbursement infrastructure [2].

One of the primary sources of this inefficiency lies in the structural fragmentation of the insurance ecosystem. Although several major carriers dominate the market, a substantial portion of payers belong to a long tail of smaller insurers, many of which hold less than 1% market share [3]. This fragmentation leads to significant heterogeneity in the format and structure of Explanation of Benefits (EOB) documents, which serve as the primary source of remittance information for healthcare providers. As a consequence, clinical practices must process a wide variety of semi-structured and unstructured documents generated by different payers, each following proprietary layouts and reporting conventions [4].

Operational benchmarks in RCM impose strict performance requirements on this process. Healthcare providers typically aim to maintain a first-pass clean claim rate above 95% and a charge lag of less than seven days [5]. However, manual processing of incoming EOB documents remains a major operational bottleneck. Data entry errors during remittance posting not only increase denial rates but also contribute to revenue leakage, as administrative staff frequently lack the capacity to investigate underpayments across large document volumes [1].

https://doi.org/10.55578/amsr.2607.011

Recent advances in Large Language Models (LLMs) have stimulated significant interest in automated document understanding. Several recent approaches rely on prompt engineering or sequence-to-sequence architectures to convert document text into structured outputs. While these methods show promising results on relatively homogeneous document collections, their performance degrades substantially in real-world medical billing environments. EOB documents frequently contain multi-page tabular structures, irregular layouts, and nested financial relationships that cannot be reliably reconstructed using purely text-based representations.

Empirical observations from production environments suggest that pure text-based or zero-shot extraction strategies exhibit unstable behavior under such conditions. Linear sequence models often fail to preserve spatial relationships between tokens, while prompt-based LLM pipelines remain vulnerable to hallucinated numerical values and non-deterministic outputs when confronted with the long-tail variability of payer formats.

These limitations motivate the need for alternative system architectures capable of integrating spatial, visual, and textual information while maintaining strict numerical consistency. In this work, we investigate a cascaded document intelligence pipeline designed specifically for high-volume EOB processing. The proposed architecture combines visual document segmentation, layout-aware classification, specialized entity extraction modules, and a deterministic financial verification layer that enforces arithmetic consistency across extracted financial fields.

The main contributions of this study are summarized as follows:

1. We propose a cascaded document intelligence architecture for high-throughput processing of heterogeneous Explanation of Benefits (EOB) documents, integrating visual document segmentation, layout-aware classification, and structured entity extraction.
2. We introduce a layout-aware multi-modal classification stage based on LayoutLMv3 representations, enabling robust handling of the long-tail variability of insurance payer document formats.
3. We design specialized extraction modules for both key-value fields and multi-page tabular claim structures, allowing reliable reconstruction of financial relationships between service lines and claim-level summaries.
4. We develop a deterministic financial verification engine that enforces arithmetic constraints across extracted monetary fields, significantly reducing the impact of numerical hallucinations in neural extraction models.
5. We present a large-scale production evaluation on a real-world medical billing workload processing over 300,000 EOB documents per month, demonstrating substantial improvements in extraction accuracy, first-pass resolution rate, and operational charge lag.
6. The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed cascaded architecture. Section 4 presents experimental results and error analysis. Section 5 concludes the paper and outlines directions for future work.

2. RELATED WORK AND LITERATURE REVIEW

The efficiency of Revenue Cycle Management (RCM) in healthcare is directly dependent on the speed and accuracy of processing Explanation of Benefits (EOB) documents. This section reviews current advancements in document understanding, medical entity extraction, and financial data integrity.

2.1. Information Extraction from Form-like Documents

Information Extraction (IE) from EOBs presents unique challenges due to complex tabular structures, heterogeneous layouts, and domain-specific terminology. Recent advances in Document AI have increasingly relied on multi-modal architectures that jointly model textual, visual, and spatial information. The LayoutLM family of models [6] and its subsequent evolution, LayoutLMv3 [7], established strong baselines for form understanding by incorporating document layout directly into Transformer representations.

Further developments have explored tighter fusion of visual and textual modalities. DocFormer [8] introduced an end-to-end multimodal Transformer architecture that jointly encodes textual content, spatial coordinates, and image features through shared self-attention mechanisms. Unlike earlier approaches that relied on separate visual and textual encoders, DocFormer demonstrated that unified multimodal representations improve robustness across diverse document understanding tasks, including key-value extraction and form analysis.

More recently, OCR-independent approaches have emerged as an alternative paradigm. Donut (Document Understanding Transformer) [9] proposed an OCR-free architecture that directly converts document images into structured outputs using a vision encoder-decoder framework. By eliminating dependence on external OCR systems, Donut reduces error propagation between text recognition and information extraction stages and achieves competitive performance on several benchmark datasets.

Despite these advances, practical deployment in high-volume healthcare revenue cycle environments remains challenging. EOB documents frequently contain payer-specific templates, dense financial tables, handwritten annotations, and arithmetic relationships that require strict consistency checks. Consequently, many production systems continue to combine Transformer-based document understanding with task-specific extraction modules and deterministic validation components to achieve both scalability and reliability.

2.2. Processing Financial and Insurance Data

In the domain of financial document processing, accuracy is paramount. Research on the Kleister-NDA and Kleister-Invoice datasets [10] highlights the difficulty of extracting long-tail entities from semi-structured layouts. For medical insurance claims, systems must not only extract text but also preserve the semantic relationships between service codes and billed amounts [11].

2.3. Verification and Data Integrity

While deep learning models achieve high performance, they are prone to hallucinations in numerical extraction. Studies in automated accounting suggest that incorporating deterministic validation layers is essential for industrial applications [12]. The use of specialized normalization units to handle dates and names has been shown to improve the reliability of end-to-end processing pipelines in high-volume environments [13].

2.4. Architectural Considerations for Scalability

The choice between end-to-end Transformer models [14] and cascaded architectures remains a critical engineering decision. For high-throughput systems processing hundreds of thousands of documents, literature suggests that modular pipelines allow for better error isolation and targeted model optimization compared to monolithic "black-box" approaches [15]. Recent research also reflects two complementary trends in document intelligence system design. The first trend focuses on increasingly capable end-to-end architectures such as DocFormer [8] and Donut [9], which aim to unify document perception and information extraction within a single neural framework. The second trend emphasizes hybrid industrial architectures that combine learned document understanding modules with deterministic business logic and validation layers. While end-to-end models often achieve strong benchmark performance, healthcare revenue cycle applications impose strict requirements on financial correctness, auditability, and predictable processing latency, making cascaded architectures particularly attractive for production deployment.

3. METHODOLOGY

To validate our hypothesis, we developed a multi-stage cascaded pipeline designed to process heterogeneous EOB documents at scale. The architecture transitions from coarse-grained document handling to fine-grained entity extraction and financial validation. This modular approach allows for targeted optimization of each component, ensuring that the limitations of one stage (e.g., OCR noise) are mitigated by subsequent layers.

3.1. Document Splitting and Page Boundary Detection

A primary challenge in production-grade EOB processing is the handling of bulk PDF uploads, where a single file may contain multiple independent EOBs, checks, and correspondence letters. Standard linear processing fails here, as financial integrity depends on correctly identifying document boundaries.

We implemented a visual-feature-based splitting module using a lightweight Convolutional Neural Network (CNN) combined with heuristic layout analysis. The model identifies "anchor" pages (e.g., pages containing check images or unique header structures) to determine split points. This ensures that the subsequent extraction models operate on logically coherent document units, preventing cross-document data leakage.

3.2. Multi-modal Layout Classification

Given the "long tail" of insurance payers identified in [3], a universal extraction model often suffers from low precision due to conflicting spatial cues across different templates. Our system employs a layout-based classifier that categorizes each document into one of more than 50 structural payer templates.

We utilize a Multi-modal Transformer architecture (LayoutLMv3-based) that encodes three distinct signals:

- Textual Content: Semantic information extracted via Optical Character Recognition (OCR).
- Spatial Geometry: 2D coordinates (bounding boxes) of each token.
- Visual Features: A linear embedding of the document image patches.

3.3. Cascaded Information Extraction

Following classification, the extraction process is divided into two parallel paths:

1. Key-Value Extraction: Identifying global attributes such as Payer Name, Check Number, and Total Amount Paid.
2. Table Structure Recognition: EOBs are inherently tabular. We employ a specialized Relation Extraction (RE) head that links "Service Dates," "Procedure Codes," "Allowed Amounts," and "Patient Responsibility" within nested rows.

Unlike standard prompt engineering, which often struggles with multi-page tables, our cascaded approach treats table extraction as a graph-connectivity problem, ensuring that line items spanning across page breaks are correctly associated with the parent claim.

3.4. Field-Specific Normalization Units

Raw extraction results are rarely sufficient for automated financial posting. We implemented a layer of deterministic and probabilistic normalization units:

- Name Parsing: A Hidden Markov Model (HMM) designed to split concatenated strings into First Name, Last Name, and Middle Initial, handling the high variance in provider and patient name formats.
- Date and Currency Standardization: Rule-based engines that convert disparate formats into a unified ISO-8601 standard and normalize currency strings into decimal values.

3.5. Financial Integrity and Verification Engine

The final stage of our methodology is a non-neural validation layer. As neural models are prone to "hallucinations" of numerical values, we enforce a strict financial balancing logic:

$$\sum_{i} NetPayment_i = TotalCheckAmount$$

$$BilledAmount - Adjustments = AllowedAmount$$

If a document fails these arithmetic constraints, it is automatically flagged for manual review.

4. RESULTS AND DISCUSSION

In this section, we evaluate the performance of the proposed cascaded intelligence pipeline against baseline methodologies using real-world production data. The results demonstrate significant improvements in both technical extraction accuracy and operational throughput.

4.1. Dataset Characteristics

The evaluation was conducted on a large-scale heterogeneous dataset derived from active medical billing operations, processing approximately 300,000 EOB documents per month. To ensure statistical significance, we curated a gold-standard test set of 10,000 pages, manually annotated by certified medical coders. The dataset exhibits high structural entropy, covering over 50 distinct payer layouts, including semi-structured scans and complex multi-page tables.

4.2. Accuracy Metrics and Comparative Analysis

We compared the proposed cascaded system against two baseline approaches: heuristic-based systems using rigid templates and Large Language Model (LLM) zero-shot. The performance was measured using three industry-standard KPIs: F1-score (entity-level), First-pass Resolution Rate (FPRR), and Straight-through Processing (STP).

Table 1. Comparative performance of extraction methodologies on a multi-payer EOB dataset.

Metric	Heuristic Baseline	Zero-shot LLM	Proposed Pipeline
F1-score (Entity Level)	0.72	0.88	0.96
FPRR (%)	64.2	79.5	95.8
STP Rate (%)	41.0	62.3	89.2

4.3. Component-wise Contribution Analysis

To quantify the contribution of individual architectural components, an ablation study was conducted on a representative subset of 25,000 EOB documents. Starting from the full production pipeline, components were progressively removed and the resulting First-pass Resolution Rate (FPRR) and Straight-through Processing (STP) rates were measured.

Table 2. Ablation study of the proposed cascaded architecture

Configuration	FPRR (%)	STP (%)
Full pipeline	95.8	89.2
Without verification engine	91.7	74.8
Without payer-specific classifier	88.9	81.6
Without table extraction module	84.1	69.7
LayoutLMv3 extraction only	80.3	61.5

The results demonstrate that no individual component is solely responsible for the observed performance gains. The deterministic financial verification engine contributes primarily to STP performance by preventing propagation of financially inconsistent records, while the payer-specific LayoutLMv3 classifier improves extraction accuracy by reducing downstream schema mismatch

errors. The largest degradation is observed when removing the specialized table extraction module, confirming the importance of explicit modeling of tabular financial relationships in EOB documents.

The results indicate that while LLMs perform reasonably well on text-heavy tasks, they struggle with spatial table alignment, leading to the hallucination of financial values. In contrast, the cascaded pipeline achieves an FPRR of 95.8%, meeting the critical Revenue Cycle Management benchmarks. The lower STP rate for LLMs is primarily due to their inability to consistently pass the financial integrity check, which the deterministic verification layer in our pipeline handles with high precision.

4.4. Error Analysis and Model Limitations

A granular analysis of system failures provides insight into the limitations of neural extraction. As illustrated in Figure 1, the majority of errors are concentrated in two categories.

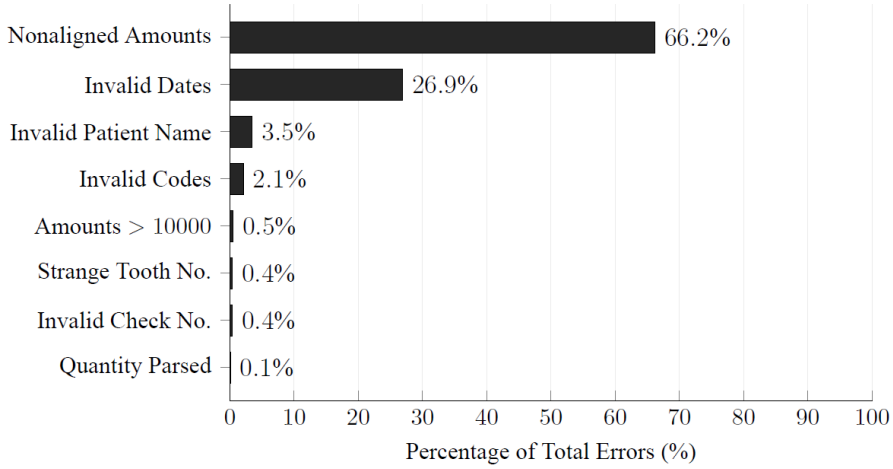

Figure 1. Relative distribution of error categories in the cascaded pipeline

The dominance of nonaligned amounts (66.2%) and invalid dates (26.9%) justifies the implementation of specialized verification and normalization layers.

Nonaligned amounts represent 66.2% of failures, occurring when the model misassociates an adjustment reason with the wrong financial column. This underscores the necessity of our multi-modal approach; whereas linear LLMs process documents as sequences of tokens, our system preserves the two-dimensional spatial context required to maintain row-column integrity.

4.5. Operational Performance and Infrastructure Efficiency

In addition to extraction accuracy, production deployment requires predictable processing latency and infrastructure efficiency. Therefore, operational characteristics of the proposed pipeline were evaluated against the zero-shot LLM baseline.

Table 3. Operational comparison of document processing approaches

Metric	LLM (Zero-shot)	Proposed Pipeline
Average latency per document (s)	7.8	0.94
95th percentile latency (s)	12.4	1.7
GPU memory utilization (GB)	24.0	8.0
CPU utilization (cores)	4	8

Metric	LLM (Zero-shot)	Proposed Pipeline
Estimated processing cost per 1000 EOBs (USD)	38.5	4.7

Measurements were obtained on production-equivalent infrastructure consisting of NVIDIA A10-class GPUs and 32-core CPU servers. Cost estimates include OCR, model inference, and post-processing operations.

The cascaded architecture provides substantially lower inference latency and operating cost than the LLM baseline. This advantage arises from decomposition of the extraction task into specialized modules, allowing computationally expensive transformer inference to be applied only where semantic understanding is required, while arithmetic validation and normalization are executed through lightweight deterministic rules. The resulting architecture supports production-scale workloads exceeding 300,000 documents per month while maintaining predictable processing characteristics.

4.6. Operational and Financial Impact

The implementation of the cascaded pipeline has yielded a significant operational impact. Prior to automation, the average charge lag was 6.4 days, largely due to manual reprocessing and high denial rates. Following deployment, the charge lag was reduced to less than 1.2 days, representing an 81% improvement. By achieving an 89.2% STP rate, the system enables administrative staff to pivot from manual data entry to high-value denial management, directly addressing the revenue leakage challenges inherent in modern healthcare finance.

5. CONCLUSION

The complexity and fragmentation of the medical insurance market demand a departure from traditional, linear document processing methodologies. This study demonstrates that while Large Language Models and heuristic-based systems offer partial solutions, they lack the spatial awareness and deterministic rigor required for production-grade Revenue Cycle Management. By implementing a hybrid cascaded pipeline, we have successfully bridged the gap between neural-network-based extraction and financial integrity, achieving a First-pass Resolution Rate of 95.8% and reducing the average charge lag by 81%.

The conducted ablation study demonstrates that neither LayoutLMv3 classification nor deterministic verification alone is sufficient to achieve production-grade performance; rather, the observed gains emerge from their interaction within the cascaded architecture. In particular, payer-aware layout classification improves extraction robustness, specialized table reconstruction preserves financial relationships, and deterministic verification substantially increases straight-through processing by enforcing arithmetic consistency.

Our error analysis reveals that the primary challenges in EOB processing—spatial non-alignment and date formatting—cannot be solved by increased model scale alone. Instead, the integration of a multi-modal architecture with a secondary financial verification engine ensures that only mathematically consistent data enters the practice management system.

The most sustainable path for high-volume document processing lies in the human-in-the-loop (HITL) ecosystem. By flagging documents that fail the arithmetic integrity checks, the system identifies the most informative samples for manual review. These corrections provide a high-quality data stream for continuous fine-tuning of the underlying LayoutLM and normalization modules. This feedback mechanism ensures that the pipeline remains adaptive to the emerging "long tail" of insurance formats, effectively transforming manual interventions into a strategic asset for model evolution.

Ultimately, the transition to cascaded intelligence transforms the medical billing workflow from a labor-intensive administrative task into a scalable, high-speed information science. Future work will focus on the deployment of active learning strategies to further automate the identification of

novel document structures, moving closer toward a fully autonomous revenue reconciliation framework.

Supplementary Materials: Not applicable.

Funding Statement: This study was not funded by any external sources.

Contribution: The authors contributed to the research and writing of this article and have read/agreed to the published version of the manuscript.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflict of Interest Statement: The authors declare no conflicts of interest.

REFERENCES

[1] Chandawarkar, R., et al. (2024). Revenue cycle management: The art and the science. *PRS Global Open*, *12*(7), e5756.
[2] Guan, S. (2024, December). Predicting medical claim denial using logistic regression and decision tree algorithm. In *2024 3rd International Conference on Health Big Data and Intelligent Healthcare (ICHIH)* (pp. 7-10). IEEE.
[3] Vujicic, M., et al. (2018). Why we need more data on the dental insurance market. *JADA*, *149*(1), 75-77.
[4] Derricks, J. (2021). Overview of the claims submission, medical billing, and revenue cycle management. In *Springer* (pp. 251-276). Springer.
[5] Manley, R., & Satiani, B. (2009). Revenue cycle management. *JVS*, *50*(5), 1232-1238.
[6] Xu, Y., et al. (2020). LayoutLM: Pre-training of text and layout for document image understanding. In *KDD* (pp. 1192-1200). ACM.
[7] Huang, Y., et al. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. In *ACM MM* (pp. 4083-4091). ACM.
[8] Appalaraju, S., et al. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 993-1003). IEEE/CVF.
[9] Kim, G., et al. (2022, October). OCR-free document understanding transformer. In *European Conference on Computer Vision* (pp. 498-517). Springer Nature Switzerland.
[10] Gralinski, F., et al. (2021). The Kleister-NDA dataset and tasks for document understanding. In *EACL* (pp. 38-48). Association for Computational Linguistics.
[11] Majumder, B. P., et al. (2020). Representation learning for information extraction from form-like documents. In *ACL* (pp. 6495-6504). Association for Computational Linguistics.
[12] Palm, R. B., et al. (2019). Attend, copy, parse: End-to-end information extraction. In *ICDAR* (pp. 812-819). IEEE.
[13] Willmen, T., et al. (2021). Health economic benefits through the use of diagnostic support systems and expert knowledge. *BMC Health Services Research*, *21*(1), 947.
[14] Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 5998-6008). Curran Associates.
[15] Borchmann, L., et al. (2021). DUE: End-to-end document understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates.