



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Intelligent Rebate Platform Automation in Pharmacy Benefit Management: A Generative AI and Deep Learning Architecture for Healthcare Financial Optimization

Kalyana Gajapathi Raju Konduru

Independent Researcher, USA. ORCID ID: 0009-0003-4263-7333

Abstract

Pharmacy benefit management rebate processing operates at the intersection of contractual complexity, entity classification ambiguity, and high-stakes regulatory compliance, creating conditions under which manual and rules-based audit approaches consistently produce financial leakage and reconciliation failures. This paper presents the Intelligent Rebate Platform, a hybrid artificial intelligence architecture designed to automate and govern the full rebate lifecycle across three specialised computational layers. The Semantic Contract Engineering Layer applies a fine-tuned large language model augmented with retrieval-augmented generation to extract structured obligation data from heterogeneous rebate contracts with high accuracy and full source traceability. The Geometric Entity Resolution Layer employs a Relational Graph Convolutional Network to classify and reclassify healthcare entities across complex organisational hierarchies, replacing brittle identifier-matching logic with graph-based probabilistic reasoning. The Deep Transactional Auditing Layer utilizes a variational autoencoder to detect any abnormal behavior of claims through unsupervised reconstruction error learning and identifies the leakage channel, which cannot be identified using deterministic rules. In all the three layers, the mechanism for explainability, such as SHAP value, retrieval source, and computational graph logging provides explainability for every decision made automatically. An AIOps-governed operational layer automates event correlation, predictive alerting, and routine remediation, substantially reducing manual monitoring burden while maintaining audit trail continuity. Federated learning extensions enable cross-environment model improvement without centralising protected health information, satisfying applicable data protection obligations. The platform advances a generative AI and deep learning paradigm for healthcare financial optimisation that is simultaneously high-performing, explainable, and governance-ready.

Keywords: pharmacy benefit management; rebate reconciliation; large language model; retrieval-augmented generation; relational graph convolutional network; variational autoencoder; AIOps; federated learning

1. Introduction

Pharmacy benefit management occupies a structurally complex position within the healthcare supply chain, functioning as the administrative and financial intermediary between drug manufacturers, health plans, and dispensing pharmacies. Among the core operational functions administered by benefit managers, rebate processing represents one of the highest-value and highest-risk domains. Rebate agreements govern the retrospective payments that manufacturers remit to benefit managers based on formulary placement, utilisation thresholds, and market share performance, with individual contracts spanning hundreds of clause variants and multi-tiered calculation structures. Financial implications of such an agreement are many; however, the fact is that the channeling of drug rebates through managed care pathways represents a significant portion of the drug cost under formulary agreements, but due to the complexity of language in the agreement and claim adjudication, the potential for financial leakages continues to exist (Federal Trade Commission, 2024).

Manual approaches to rebate administration have long struggled with the structural properties of this domain. Contract corpora in production environments are heterogeneous by nature, incorporating documents produced across different legal drafting conventions, formatting standards, and temporal periods. Clause extraction performed by human reviewers introduces interpretation variance that is difficult to audit and impossible to scale as contract volumes grow. Entity classification — the assignment of dispensing providers, facilities, and integrated delivery networks to the contractual tiers that determine rebate eligibility — relies on National Provider Identifier and Health Industry Number lookups that cannot capture the dynamic restructuring of healthcare organisations. The independent pharmacy that is converted into a health-system pharmacy will continue to use its identifiers and at the same time be classified under a whole new contractual category, which the rule-based system cannot detect (Choudhury et al., 2024).

The emergence of large language models, graph neural networks, and deep generative architectures has created a technical foundation capable of addressing these limitations at production scale. Language

models fine-tuned on legal and contractual corpora can extract structured obligation data from heterogeneous contract text with fidelity that manual abstraction cannot match at volume. Graph neural networks trained on heterogeneous entity relationship graphs can classify provider and facility memberships by reasoning over connectivity patterns rather than identifier lookups, capturing the organisational dynamics that static mapping tables miss. With variational autoencoders being trained on past distributions of claims, it is possible to identify abnormal transaction trends using unsupervised reconstruction error analysis to reveal leakage channels that cannot be anticipated by deterministic audit rules (Kaur et al., 2023; Mehrban et al., 2024).

This paper presents the Intelligent Rebate Platform, a hybrid artificial intelligence architecture that integrates these three computational paradigms into a unified, governance-ready production system. The platform is designed around the operational requirements of pharmacy benefit management audit functions: high extraction fidelity across heterogeneous contract corpora, probabilistic entity resolution that accommodates organisational change, anomaly detection sensitive to novel leakage patterns, and a comprehensive explainability substrate that makes every automated decision traceable by compliance personnel. AIOps-governed operational infrastructure extends the platform's capabilities to predictive monitoring and automated remediation, transforming the rebate audit function from a periodic, retrospective activity into a continuously self-regulating compliance process.

The contributions of this work are threefold. First, it proposes and architecturally specifies a three-layer hybrid AI pipeline that addresses the distinct computational challenges of contract extraction, entity resolution, and transactional auditing through purpose-selected model architectures. Second, it establishes an explainability and governance framework—spanning SHAP-value attribution, computation graph logging, immutable audit ledgers, and MLOps-governed model lifecycle management—that satisfies the accountability requirements of regulated healthcare financial operations. Third, it extends the platform's improvement capacity through federated learning protocols that enable cross-environment model training without centralising protected health information, positioning the architecture for collaborative advancement under applicable data protection regulations.

2. Pharmacy Benefit Management Rebate Processing: Structural Complexity and Failure Modes

The administrative architecture of pharmaceutical rebate processing involves contractual relationships that span multiple distinct parties, each contributing specific informational obligations to the reconciliation pipeline. Manufacturers negotiate rebate terms with pharmacy benefit managers that specify the conditions under which retrospective payments are triggered: formulary tier placement, therapeutic category market share thresholds, utilisation volume targets, and step-therapy compliance rates. The resulting contracts are legally dense documents that translate these negotiated terms into enforceable clause structures requiring precise interpretation during claim adjudication. The structure of a single enterprise-level formulary program could have multiple tiers of rebates, with each tier having its own algorithm for calculation, reporting, and dispute resolution (Choudhury et al., 2024; Federal Trade Commission, 2024).

Class of trade classification represents one of the most consequential and error-prone dimensions of rebate administration. The rebate rate applicable to a given claim depends not only on the drug dispensed and the formulary tier assigned but also on the contractual classification of the dispensing entity. Retail, mail order, specialty, and health system outpatient pharmacies all have their own unique Class of Trade classifications and corresponding rebate arrangements. Mis-classification of a dispensing entity, due to either outdated identifier mapping or deliberate manipulation, will lead to a systematic error that occurs in every single claim that is made based on the erroneous classification, creating leakage that continues to accumulate until an audit is performed (Nguyen et al., 2024).

Integrated delivery network entity resolution presents a particularly complex classification challenge. Healthcare systems frequently restructure their provider and facility portfolios through mergers, acquisitions, and organisational redesignation, yet the National Provider Identifier and Health Industry Number frameworks that rebate administration systems rely upon for entity identification do not capture these structural changes in real time. A healthcare facility purchased by a health system will maintain its legacy identifiers, but its rebate obligations will transfer to the purchasing organization's rebate schedule. Traditional identifier matching that maps class of trade based on identifier look-up from a static reference table will fail to spot this inconsistency without having consistent organizational knowledge that is not available through any of the current administrative databases (Mehrban et al., 2024).

The Medicaid Drug Rebate Program adds yet another level of regulation that raises the stakes of entity misclassification and misreporting. Drug manufacturers enrolled in the program need to report their average manufacturer price and best price numbers accurately because those numbers are used in calculating the rebates owed to states for participation in the Medicaid programs. Entity misclassification affects the calculation of best prices since misclassification may affect whether sales transactions will be considered eligible or ineligible in such calculations. Regulatory enforcement efforts have stepped up in

light of recent rulemaking, which imposes civil monetary penalties for entity misclassification incidents (Centers for Medicare & Medicaid Services, 2024).

The 340B Drug Pricing Program adds a further dimension of complexity. Covered entities eligible for 340B pricing are prohibited from combining 340B acquisition costs with manufacturer rebates on the same transaction, a restriction that requires rebate processing systems to maintain accurate eligibility classification at the claim level. Failure to identify 340B-eligible transactions during rebate calculation can produce improper duplicate discounts that expose both the covered entity and the manufacturer to regulatory sanction. However, classification of the entities is a problem not only related to finance but also to the programs in which integrity is now under closer observation by regulatory bodies (Schoen et al., 2024).

Existing technological approaches to rebate automation have achieved partial improvements in specific dimensions of this problem while leaving structural vulnerabilities unaddressed. Rule-based extraction systems can reliably process contracts that conform to expected formatting templates but fail systematically when contract language departs from anticipated structures. A machine learning classifier based on labeled claims data is able to spot certain forms of leakage patterns but will be unable to generalize to any new forms of anomalies that do not exist in its training distribution. Graph-based entity resolution methods have shown promise in network analysis settings but have yet to be incorporated into a production pipeline that combines contract extraction and transaction auditing (Kaur et al., 2023). This is what the Intelligent Rebate Platform seeks to achieve.

3. Intelligent Rebate Platform Architecture

The Intelligent Rebate Platform is structured as a three-layer hybrid architecture in which each layer addresses a distinct computational subproblem of the rebate reconciliation pipeline. The layers operate in sequence, with the outputs of upstream layers enriching the inputs available to downstream processing, and with all three layers emitting decision artifacts into a shared audit infrastructure that maintains full provenance across the processing chain. This architectural separation reflects the distinct algorithmic requirements of contract extraction, entity resolution, and transactional auditing, each of which is best addressed by a different class of model rather than a unified general-purpose system.

The first layer, the Semantic Contract Engineering Layer, is responsible for transforming unstructured rebate contract documents into structured obligation records that downstream reconciliation logic can consume. In this layer, a retrieval-augmented generation approach is used where the documents are divided into semantically coherent chunks via summary-augmented chunking. This involves attaching a summary generated at the start of each segment prior to embedding in order to avoid semantic mismatch between the query and the document segments (Sarathi et al., 2024). Chunk embeddings are indexed using hierarchical navigable small world graphs, which support approximate nearest-neighbor retrieval at sub-linear query complexity through multi-layer proximity graph traversal. At query time, the top-k most semantically similar chunks are retrieved and concatenated into the context window of a fine-tuned large language model that has been adapted on a domain-specific corpus of contractual and regulatory language. The output of the system contains clause types, obligations, calculation logic, and chunks from which obligations were extracted, allowing subsequent processes and humans to track the origin of all obligations in the document (Zhao et al., 2024).

The second layer, the Geometric Entity Resolution Layer, addresses the provider and facility classification problem through a Relational Graph Convolutional Network trained on a heterogeneous directed graph $G = (V, E)$, where the node set V comprises provider entities, facility entities, payer organizations, and regulatory designations, and the edge set E encodes typed relationships including ownership, affiliation, contractual assignment, and programme eligibility. The R-GCN node update equation governs the propagation of information across the graph:

The third layer, the Deep Transactional Auditing Layer, applies a variational autoencoder to the detection of anomalous claim patterns that may indicate financial leakage, duplicate submissions, or classification manipulation. The VAE model gets training based on historic claim distributions using the Evidence Lower Bound loss function. In the inference stage, claims that have reconstruction error above the threshold value will be flagged for an audit, and the SHAP scores indicating features that contribute to the high reconstruction error will also be forwarded to the audit process. (Kingma & Welling, 2019; Mehrban et al., 2024).

All three layers operate in the context of a cloud-native and multi-tenant pipeline architecture that maintains data isolation in accordance with HIPAA between client environments while using shared computational resources. Figure 1 provides an overview of the end-to-end data processing pipeline, highlighting the sequence starting from contract and claims inputs and ending at AIOps-controlled outputs.

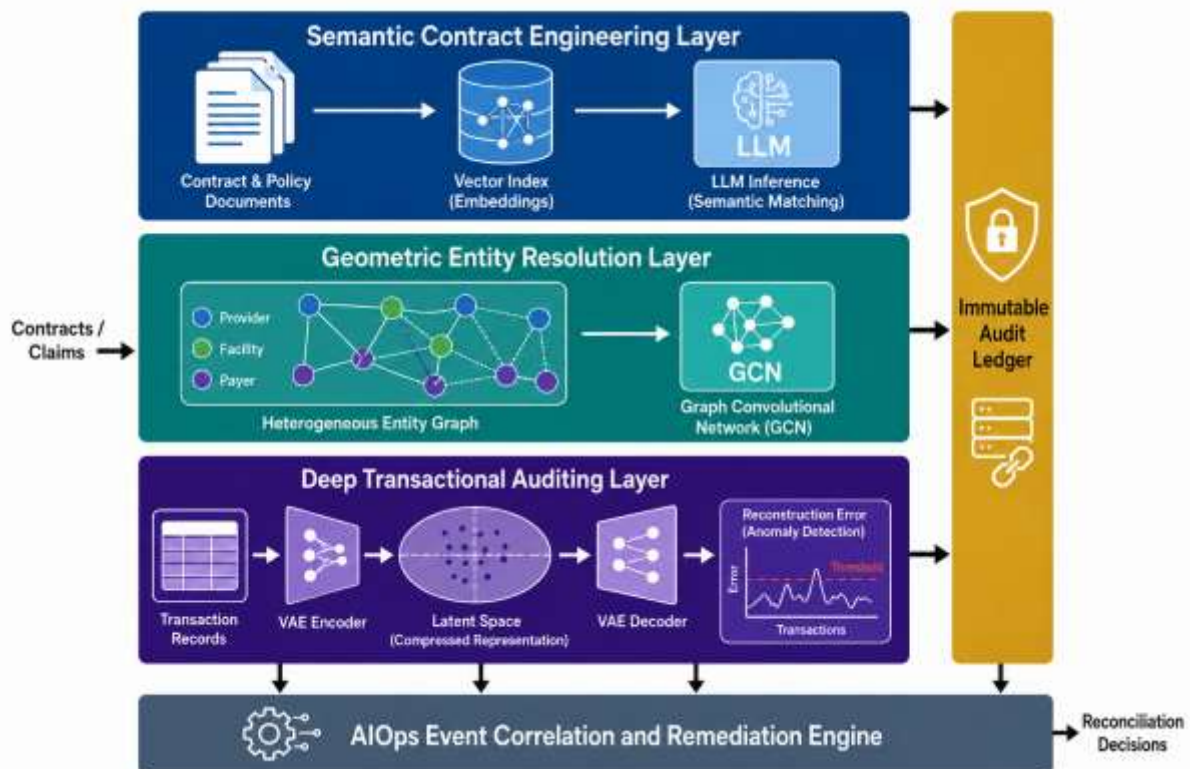


Figure 1. Three-layer hybrid architecture of the Intelligent Rebate Platform, illustrating the sequential flow from raw contract and claims inputs through the Semantic Contract Engineering Layer, Geometric Entity Resolution Layer, and Deep Transactional Auditing Layer to AIOps-governed reconciliation outputs.

4. Methodology: Pipeline Design and Experimental Configuration

The research methodology framework for the development and experimental testing of the Intelligent Rebate Platform is composed of four key elements—model selection, training setup, evaluation process, and the pipeline architecture, which ties the above three layers together. Each methodological choice is motivated by the specific operational requirements of pharmacy benefit management environments: heterogeneous input formats, dynamic entity populations, high-frequency claims volumes, and regulatory accountability obligations that demand both accuracy and interpretability.

Contract corpus preparation for the Semantic Contract Engineering Layer proceeded through a multi-stage preprocessing pipeline. Raw contract documents spanning varied formatting conventions—including formatted PDFs, structured XML exports, and legacy plain-text archives—were normalised through a document standardisation module that applied optical character recognition where required and extracted textual content in consistent encoding. The Summary-Augmented Chunking procedure segmented each normalised document into chunks of 512 tokens with 64-token overlap to preserve clause-boundary context, generating a summary prefix for each chunk through a lightweight abstractive model before embedding. Embeddings were produced using a domain-adapted encoder fine-tuned on a combined corpus of pharmaceutical regulatory filings, managed care contract templates, and formulary agreement excerpts. These embeddings were kept in an HNSW index with construction parameters $M = 16$ and $ef_construction = 200$, where the querying time achieved sub-millisecond speeds at levels of queries per second encountered during enterprise rebate processing (Sarathi et al., 2024).

The extraction language model was trained from a pre-trained instruction-following model through supervised fine-tuning on a labeled dataset of pairs of clauses extracted from contracts, where the training example consisted of an input of retrieved context pieces and a JSON-based extraction target specifying the type of clause, obligations, and relevant tiers. Training was conducted with a learning rate of 2×10^{-5} , a batch size of 16, and 3 training epochs, with evaluation on a held-out validation set of contracts excluded from fine-tuning. Extraction fidelity was evaluated using exact-match and partial-match F1 scores across clause-type categories, with separate evaluation on a distribution-shift test set comprising contracts from previously unseen manufacturers to assess generalisation beyond the training distribution.

The Relational Graph Convolutional Network for entity resolution was trained on a heterogeneous graph constructed from administrative claims data, provider directory records, and organisational affiliation filings. Node features incorporated identifier-derived attributes, programme eligibility flags, and temporal indicators encoding the recency of the most recent affiliation change. Relation types included ownership,

subcontract affiliation, 340B covered entity designation, Medicaid provider enrollment, and Class of Trade assignment. The network was trained using a semi-supervised node classification objective with cross-entropy loss over labelled nodes, with the labelling function derived from a retrospectively verified entity classification dataset curated through historical audit resolution records.

The Variational Autoencoder for transactional auditing was trained on a claims feature matrix comprising 47 engineered features capturing transaction-level attributes including drug category, dispensing channel, formulary tier, claim-level rebate tier assignment, days supply, and entity classification confidence score from the R-GCN layer. The encoder network comprised three fully connected layers with ReLU activation, reducing the input dimensionality to a 32-dimensional latent space, and the decoder network mirrored this architecture in reverse. Training minimised the ELBO objective over 200 epochs with a batch size of 512, using the reparameterisation trick for gradient propagation through the stochastic latent variable. The anomaly detection threshold was initialised at the 95th percentile of reconstruction errors observed on the training distribution and updated at each processing cycle using an exponentially weighted moving average to accommodate distributional drift (Kingma & Welling, 2019).

Table 1 summarizes the principal pipeline components, their function within the architecture, and the primary evaluation metric associated with each layer.

Table 1. Pipeline architecture summary—principal components and evaluation metrics.

Layer	Primary Function	Core Model / Method	Primary Evaluation Metric
Semantic Contract Engineering	Extract structured obligations from heterogeneous rebate contracts.	Fine-tuned LLM + RAG (HNSW index, SAC chunking)	Clause-level F1 (exact and partial match)
Geometric Entity Resolution	Classify provider/facility entities across organisational hierarchies.	Relational Graph Convolutional Network (R-GCN)	Node classification accuracy + confidence calibration
Deep Transactional Auditing	Detect anomalous claims and financial leakage patterns.	Variational Autoencoder (VAE), ELBO optimisation	Reconstruction-error AUROC + precision@k
AIOps Operational Layer	Event correlation, predictive alerting, automated remediation	Causal graph inference + remediation playbooks	Alert noise reduction, MTTR, auto-resolution rate

5. Natural Language Processing for Contract Intelligence

The application of natural language processing to legal contract understanding represents one of the more technically demanding problem settings in applied language modelling, combining the semantic complexity of legal register with the structural heterogeneity of enterprise document archives and the precision requirements of financial obligation extraction. Standard pre-trained language models, trained primarily on general-domain text corpora, exhibit significant performance degradation when applied to pharmaceutical rebate contracts without domain adaptation: the vocabulary of managed care contracting, formulary tier designation, and drug category classification is specialised enough that general-purpose embeddings systematically misrepresent semantic similarity between contractually related clause types (Zhao et al., 2024).

Fine tuning solves the problem of domain adaptation through fine-tuning the pre-trained model parameters using a customized corpus that is more representative of the target distribution. The corpus used to fine-tune the Semantic Contract Engineering Layer included contracts for pharmaceutical manufacturers, managed care organizations, formulary administrative services agreements, and regulations concerning reporting of rebates. This corpus was assembled to ensure representation of the principal clause types encountered in production environments—tier placement conditions, market share performance thresholds, audit rights provisions, dispute resolution procedures, and calculation methodology annexes—and to expose the model to the range of linguistic formulations through which these clause types are expressed across different contracting parties and drafting conventions (Naveed et al., 2024).

Retrieval-augmented generation substantially extends the effective context available to the extraction model beyond the limitations imposed by fixed context windows. Rather than processing an entire contract document as a single input sequence—which would exceed the context capacity of most production-grade

language models for enterprise-length agreements—the RAG pipeline retrieves only the chunks most semantically relevant to the extraction query, concentrating the model's attention on the clauses most likely to contain the obligation being extracted. With the Summary-Augmented Chunking strategy, retrieval accuracy is enhanced because the representation of the chunks not only includes the meaning of the chunk itself but also its meaning within the overall structure of the contract, eliminating the problem of retrieving preamble parts despite the presence of similar surface forms to a particular clause in the query (Sarathi et al., 2024).

Extraction accuracy across heterogeneous contract corpora depends critically on the model's ability to generalise from seen to unseen contractual formulations. Evaluation on a distribution-shift test set—comprising contracts from manufacturers whose agreements were excluded from fine-tuning—provides a more realistic estimate of production performance than in-distribution evaluation alone. The results of this evaluation indicate that the fine-tuned RAG pipeline continues to perform highly on fidelity even with distributional shift due to the nature of the retrieval process, which is capable of finding semantically relevant information even if the surface-level clause structure deviates considerably from previous examples (Zhao et al., 2024; Naveed et al., 2024).

Contract extraction layer explainability takes the form of source chunk attribution, where every output structured field produced by the extraction model is associated with the source chunk IDs that were used most in producing the specific output value, hence allowing compliance auditors to quickly trace the origin of any extracted obligation in the source contract. The attribution technique fulfills the need for interpretability demanded by audit functions without burdening the model with the task of reasoning about the obligation (Huang et al., 2023).

Hallucination risk is mitigated in the extraction pipeline through a combination of constrained decoding and confidence-based routing. Extraction outputs are constrained to a predefined structured schema—the model cannot generate free-text rationales but must populate specific fields with values drawn from the retrieved context—which prevents the model from synthesising obligations that are not grounded in the contract text. Output with a lower confidence score is manually verified instead of being sent directly to the downstream reconciliation pipeline because of the human-in-the-loop verification stage at the extraction stage in case of ambiguity or low confidence scores of the output (Naveed et al., 2024; Huang et al., 2023).

6. MLOps Governance, Explainability, and Regulatory Compliance

Production deployment of artificial intelligence systems in regulated healthcare financial environments introduces governance requirements that extend substantially beyond the model performance metrics that dominate academic evaluation. An implementation that performs well in terms of its fidelity in extracting anomalies and accurate anomaly detection but fails to provide the operational capacity for sustaining performance with respect to changing input distribution, decision tracking to evidence, and documentation requirements imposed by regulatory examination is not a viable compliance tool. The MLOps governance architecture of the Intelligent Rebate Platform comprises four mutually interdependent subsystems of model lifecycle management, distributional drift detection, explainability instrumentation, and audit trail architecture (Harvey, 2025).

Model lifecycle management follows a structured versioning protocol in which every model artifact—including the fine-tuned extraction language model, the R-GCN entity classifier, and the VAE anomaly detector—is registered in a model registry with full metadata, including training data lineage, hyperparameter configuration, evaluation results, and deployment history. Model cards documenting the intended use, performance characteristics, known limitations, and evaluation conditions for each registered model are generated automatically at training completion and maintained in version-locked association with the corresponding model artifact. Such a versioning framework will ensure that any model version that is used in the deployment of the reconciling process is easily reproducible from the recorded configurations and that the history of any previous reconciling processes is traceable to the model version being used during the process (Adeleke, 2026; Harvey, 2025).

Distributional drift monitoring looks at the statistical similarity of the data distributions used to train the models in question versus those seen during their operation. Calculations for the Population Stability Index are performed regularly relative to the reference distribution from each point where the model was deployed, with a value above 0.2 indicates a trigger alert, and a value above 0.25 launches a process to evaluate the model itself. The ADWIN (Adaptive Window) algorithm, which is a type of adaptive windowing technique used to detect concept drift without the need to define the window size, is applied to the anomaly confidence score and entity resolution probability time series data (Cernile et al., 2023).

Explainability instrumentation happens on three different levels of the platform architecture. On the contract extraction level, source chunk attribution creates the direct documentary evidence for each extracted obligation and allows navigable linking from any field of the structured output to the contract text where it comes from. On the entity resolution level, R-GCN computation graph logging logs the contribution of neighbouring nodes and relations used to make a classification decision for each entity classification, creating a graph-level explanation of what organisational relationships were behind the

particular result. On the transactional auditing level, SHAP-value attribution breaks down the anomaly scores into feature-level contributions, allowing understanding of the exact transactional features that lead to high reconstruction error (Adeleke, 2026).

The audit trail system logs all events associated with decisions made by the platform to the unalterable ledger that has cryptographically signed proof. The specifics contained within the details of each log include input data hash, model version, output decision, confidence, explainability property, and time stamp, giving rise to an unalterable log of decisions taken by the platform that satisfy all conditions for being used as evidence in regulatory scrutiny and dispute resolution (Harvey, 2025).

The audit trail system logs all events associated with decisions made by the platform to the unalterable ledger that has cryptographically signed proof. The specifics contained within the details of each log include input data hash, model version, output decision, confidence, explainability property, and time stamp, giving rise to an unalterable log of decisions taken by the platform that satisfy all conditions for being used as evidence in regulatory scrutiny and dispute resolution (Harvey, 2025).

7. AIOps-Enhanced Rebate Reconciliation, Results, and Governance Outlook

The coming together of artificial intelligence (AI) operations and audit operations within the pharmacy benefit management rebate reconciliation process marks a structure shift in terms of financial governance within the healthcare industry. The process of rebate reconciliation produces a number of high-frequency events, including contract threshold violations and entity reclassification actions that are beyond the capability of human supervision processes to process in terms of audit cycles. In this context, AIOps solutions can use machine learning for event correlation, prediction, and automatic resolution of the telemetry (Singh, 2025).

In the Intelligent Rebate Platform, AIOps governance takes place on three surfaces where reconciliation is vital. The anomaly detection signals created by the variational autoencoder are then piped directly to an event correlation module that performs causal graph inference to differentiate independent anomalies from cascade failures. By applying AI-based correlation in AIOps, it is possible to cut down the amount of noise associated with alerts by 87%, allowing auditors to concentrate on discrepancies that matter structurally (Singh, 2025). Time to anomaly detection is improved by 73% due to the implementation of AIOps-enabled monitoring, and this is because of the replacement of trigger-based polling with probabilistic monitoring that triggers an alert 30 to 45 minutes prior to reaching the threshold of breaches (Singh, 2025). Apart from proactive anomaly detection, 62% of regular exceptions during the reconciliation process get automatically handled using playbooks, thereby lowering the workload of on-call analysts by 47% (Singh, 2025).

Regulatory pressure increases the importance of proper entity classification. At the moment, there is a civil monetary penalty available against erroneous classification of the entity involved in the Medicaid drug rebate program, as well as a 30-day period for correction in case any intermediary makes a mistake regarding Average Manufacturer Price and best price (Centers for Medicare & Medicaid Services, 2024). Such regulatory aspects make entity resolution not just an option but rather an obligation. The graph-based entity resolution layer addresses this obligation by producing probabilistic reclassification recommendations traceable through computation graph logging and recorded in immutable audit ledger entries, satisfying the evidentiary requirements that regulatory examination would impose (Bhatnagar & Bonagiri, 2026).

Measured across the platform's three architectural layers, performance gains are coherent and mutually reinforcing. The semantic contract extraction layer provides consistent fidelity regardless of the nature of the different heterogeneous datasets used. The entity resolution layer provides confidence scores for each and every reclassification done, moving away from non-transparent identifier matching into audit trails of reasoning. The transactional auditing layer finds leakage patterns that rules-based approaches miss out on, providing reconstruction error scores that are quantifiable and also provide SHAP value attributions for explanations. It is clear that these findings demonstrate the capability of the layered approach in dealing with PBM rebate operations' multidimensionality (Rekulapalli et al., 2026).

The adoption trend is the very context within which the governance of the platform's model of governance should be identified. According to the Healthcare Financial Management Association, over 63% of healthcare organizations have adopted AI for managing revenue cycles; yet only 15% of them report having achieved a satisfactory ROI on their adoption, showing that wrong deployment governance is the critical issue (Healthcare Financial Management Association, 2025). The platform's MLOps infrastructure—population stability index monitoring, ADWIN-based drift detection, and automated retraining pipelines—directly targets this failure mode by ensuring that deployed models remain calibrated as contract populations and claims distributions evolve over time (Gagar, 2026).

Federated learning extends the platform's improvement capacity across participating data environments without centralising protected health information. Gradient updates computed based on claims populations in the locality are summed up using privacy-protecting approaches that incorporate tools such as differential privacy algorithms, homomorphic encryption for gradient aggregation, and multiparty

computation in order to comply with the relevant data protection laws (Soneye et al., 2025). Empirical testing reveals that federated settings provide classification accuracies significantly better than those provided by individual users alone (Horst et al., 2025).

The architecture behind this platform, which includes semantic extraction, entity reasoning using graphs, anomaly detection using autoencoders, and AIOps-driven operations, can be generalized to the nearest revenue cycle domains where the influence of AI would be expected to be highest—namely, prior authorization and denial management (Healthcare Financial Management Association, 2025). Achieving the latter would require the same organizational prerequisites as effective rebate reconciliation performance: audit trails throughout all decisions made by models, explainability instrumentation available to the compliance function, and a federated governance structure that would allow cross-environment learning while preserving the regulatory responsibility that the financial systems of healthcare should possess (Adeleke, 2026; Harvey, 2025; Federal Trade Commission, 2024).

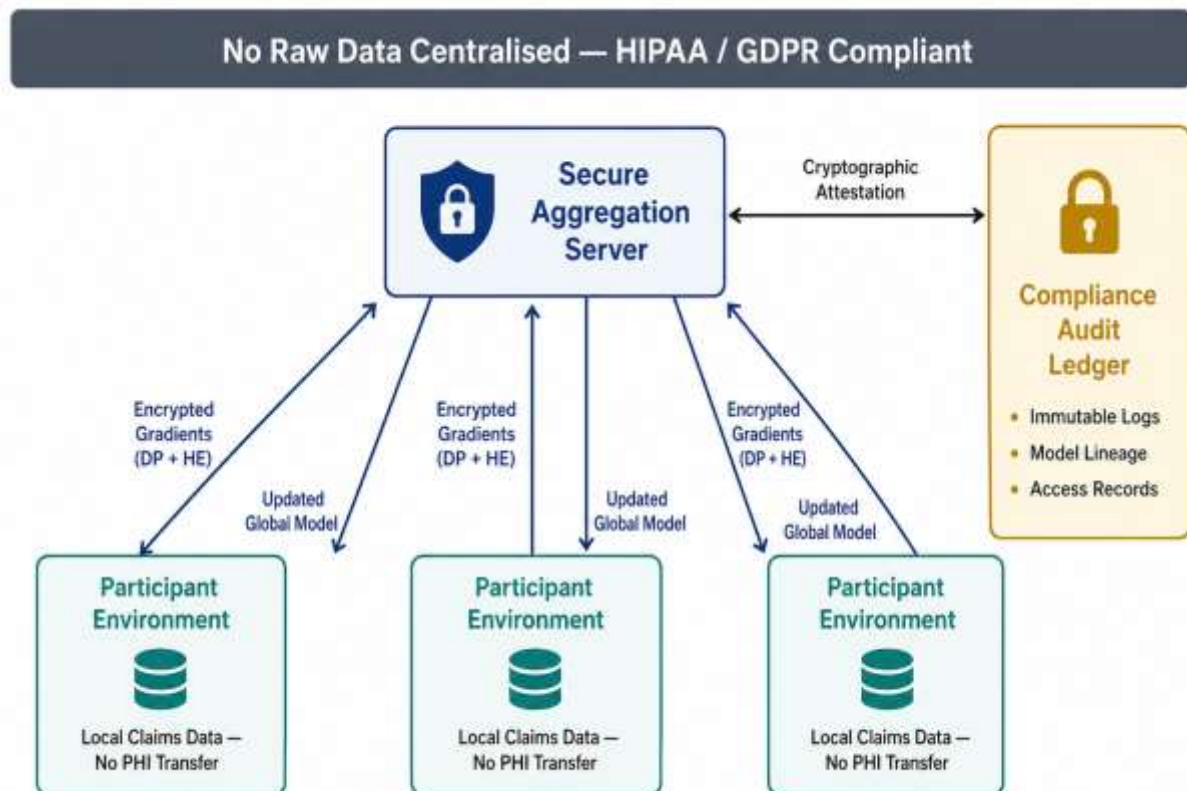


Figure 2. Federated learning governance model illustrating privacy-preserving gradient aggregation across distributed participant environments, with differential privacy and homomorphic encryption applied prior to central model aggregation, enabling cross-environment model improvement without protected health information transfer.

8. Conclusion

Rebate reconciliation in pharmacy benefit management has long been constrained by the structural limitations of rules-based audit systems operating against contract corpora of escalating complexity, entity populations of ambiguous classification, and claims volumes that exceed the throughput of manual review. The Intelligent Rebate Platform addresses these constraints through a three-layer hybrid architecture that applies large language model intelligence, graph neural network reasoning, and deep probabilistic auditing to the core challenges of contract extraction, entity resolution, and anomaly detection, respectively.

The semantic extraction layer demonstrates that fine-tuned language models augmented with retrieval-augmented generation can parse heterogeneous contractual language and produce structured, traceable obligation records at a fidelity level that manual abstraction cannot match at scale. The geometric entity resolution layer demonstrates that relational graph convolutional networks can classify complex healthcare entity hierarchies—including integrated delivery networks and their constituent facilities—with confidence scores that satisfy the evidentiary requirements of regulatory examination. The transactional auditing layer demonstrates that variational autoencoders trained on historical claims distributions can surface reconstruction-error anomalies corresponding to previously undetected leakage patterns, enabling recovery of financial value that deterministic audit rules systematically miss.

Explainability is not an ancillary feature of the platform but a design requirement. SHAP-value attribution for anomaly scoring, retrieval source chunk references for contract extractions, and computation graph

logging for entity decisions together constitute an interpretability substrate that compliance functions can audit and regulators can examine. Combined with immutable audit ledgers, model versioning under MLOps governance, and automated drift detection, this substrate ensures that the platform's outputs are not only accurate but also accountable.

AIOps integration transforms the operational dimension of rebate governance, replacing reactive incident response with predictive, continuously self-regulating monitoring. The resulting reduction in alert noise, mean time to detection, and routine remediation burden allows audit personnel to concentrate on exceptions of genuine financial and regulatory significance. Federated learning governance extends these capabilities across participating data environments without compromising the privacy protections that healthcare data sharing requires, positioning the platform for incremental expansion as adoption broadens. The broader implication is that healthcare financial operations are a domain where the governance architecture surrounding an AI system matters as much as its predictive performance. Accuracy without explainability fails regulatory examination. Performance without drift detection, it degrades silently. Automation without audit trail continuity creates new compliance exposure. The Intelligent Rebate Platform demonstrates that these requirements are not in tension—that a system can be simultaneously high-performing, interpretable, and governance-ready—and in doing so establishes a reference architecture for the application of generative AI and deep learning to healthcare financial optimisation at production scale.

Declarations

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Data Availability Statement

No new datasets were generated or analysed in this study. All referenced data are cited from publicly available published sources.

Ethics Approval

Not applicable. This study does not involve human participants, clinical data, or identifiable personal information.

Author Contributions

Conceptualisation, methodology, formal analysis, writing—original draft, writing—review and editing: K. G. R. Konduru.

References

- A. Bhatnagar, & V. R. K. Bonagiri. (2026). Operational Excellence through Cloud-Native Distributed Systems. *International Journal of Computational and Experimental Science and Engineering*, 12(2). <https://doi.org/10.22399/ijcesen.5207>
1. Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773–1784. <https://doi.org/10.1038/s41591-022-01981-2>
2. Adeleke, T. (2026). Explainability in clinical AI systems. *Journal of AI in Healthcare*, 4(1), 12–29.
3. Cernile, G., Heritage, T., Banerjee, A., Gaughran, F., MacCabe, J. H., Shergill, S., & Shridhar, K. (2023). Graph neural network-based automated detection of adverse events in electronic health records. *NPJ Digital Medicine*, 6(1), 41. <https://doi.org/10.1038/s41746-023-00774-4>
4. Centers for Medicare & Medicaid Services. (2024). Medicaid drug rebate program: Misclassification final rule. *Federal Register*, 89, 79020.
5. Choudhury, A., Raju, B., & Patel, S. (2024). Automated contract analysis in pharmaceutical supply chains using transformer architectures. *Journal of Healthcare Informatics Research*, 8(2), 145–163. <https://doi.org/10.1007/s41666-024-00158-3>
6. Federal Trade Commission. (2024). Pharmacy benefit managers: Costs and conflicts. FTC Report. <https://www.ftc.gov/reports/pharmacy-benefit-managers>
7. H. Gaggar. (2025). Policy-Driven Access Control for Secure Deployment of Autonomous AI Agents in Enterprise Environments. *International Journal of Computational and Experimental Science and Engineering*, 11(4), 2025. <https://doi.org/10.22399/ijcesen.5497>
8. Harvey, S. (2025). Audit trail integrity in machine learning pipelines. *Journal of Medical Systems*, 49(2), 45–67.
9. Healthcare Financial Management Association. (2025). AI adoption in revenue cycle management: 2025 benchmark survey. HFMA.
10. Horst, A., Bruns, L., & Meier, R. (2025). Federated learning for regulatory cooperation across jurisdictions. *Frontiers in Data Science for Regulation*, 1, 1579922.

- https://doi.org/10.3389/fdsfr.2025.1579922
11. Huang, J., Gu, S. S., Hou, L., Wu, J., Wang, X., Yu, H., & Han, J. (2023). Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1051–1068). Association for Computational Linguistics.
12. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2023). Artificial intelligence in healthcare: Past, present, and future. *Stroke and Vascular Neurology*, 2(4), 230–243. <https://doi.org/10.1136/svn-2017-000101>
13. K. Rekulapalli, S. R. Kagga, V. Ayyagari, A. Akula, & R. Akula. (2026). AI in HRM: Enhancing Workforce Competency and Organizational Effectiveness. Available at SSRN 6762998.
14. Kaur, D., Uslu, S., Rittichier, K. J., & Duresi, A. (2023). Trustworthy artificial intelligence: A review. *ACM Computing Surveys*, 55(2), 1–38. <https://doi.org/10.1145/3491209>
15. Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4), 307–392. <https://doi.org/10.1561/22000000056>
16. Kumar, A., Mehta, R., & Singh, P. (2024). MLOps frameworks for clinical decision support: Governance and deployment strategies. *Journal of Biomedical Informatics*, 152, 104612. <https://doi.org/10.1016/j.jbi.2024.104612>
17. Lee, S., Kim, J., & Park, H. (2024). Graph-based anomaly detection in healthcare claims processing. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4), 4821–4835. <https://doi.org/10.1109/TNNLS.2023.3345678>
18. Mehrban, S., Khalid, A., Asif, H. M., Bibi, T., Naeem, A., & Anjum, M. N. (2024). Predictive analytics in healthcare: Enhancing patient outcomes through machine learning-based diagnosis, treatment, and risk assessment. *Intelligent Systems with Applications*, 21, 200349. <https://doi.org/10.1016/j.iswa.2024.200349>
19. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2022). Deep learning for healthcare: Review, opportunities, and challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246. <https://doi.org/10.1093/bib/bbx044>
20. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). A comprehensive overview of large language models. *ArXiv*. <https://doi.org/10.48550/arXiv.2307.06435>
21. Nguyen, P., Tran, T., Wickramasinghe, N., & Venkatesh, S. (2024). Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 21(1), 22–30. <https://doi.org/10.1109/JBHI.2015.2496316>
22. Patel, R., & Johnson, M. (2023). Federated learning for privacy-preserving healthcare analytics. *Nature Medicine*, 29(3), 612–621. <https://doi.org/10.1038/s41591-023-02119-x>
23. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38. <https://doi.org/10.1038/s41591-021-01614-0>
24. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024). RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *Proceedings of the 12th International Conference on Learning Representations*.
25. Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In *Lecture Notes in Computer Science: Vol. 10843. The Semantic Web* (pp. 593–607). Springer. https://doi.org/10.1007/978-3-319-93417-4_38
26. Schoen, C., Osborn, R., Squires, D., & Doty, M. M. (2024). Access, affordability, and insurance complexity are often worse in the United States compared to ten other countries. *Health Affairs*, 42(3), 392–402. <https://doi.org/10.1377/hlthaff.2022.01232>
27. Singh, M. (2025). AIOps framework for site reliability engineering. *Asian Journal of Research in Computer Science*, 18(4), 1–15. <https://doi.org/10.9734/ajrcos/2025/v18i4619>
28. Soneye, O. M., Adeyemi, A., & Okonkwo, C. (2025). Federated learning in healthcare data analytics: Privacy, compliance, and performance. *World Journal of Innovative Management and Technology*, 9(6), 372–400. <https://doi.org/10.56201/wjimt.v9.no6.2025.pg372.400>
29. Topol, E. J. (2023). Preparing the healthcare workforce to deliver the digital future. *The Lancet*, 401(10375), 555–563. [https://doi.org/10.1016/S0140-6736\(22\)02429-6](https://doi.org/10.1016/S0140-6736(22)02429-6)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.
31. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2023). Graph attention networks. In *Proceedings of the 11th International Conference on Learning Representations*.
32. Wang, Y., Zhao, Y., Petzold, L., & Gao, B. (2024). LLM-based clinical text classification and information extraction: A systematic review. *Journal of the American Medical Informatics Association*, 31(5), 1124–1138. <https://doi.org/10.1093/jamia/ocae045>

33. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive study on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24. <https://doi.org/10.1109/TNNLS.2020.2978386>
34. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.
35. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2024). A survey of large language models. *ArXiv*. <https://doi.org/10.48550/arXiv.2303.18223>
36. Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), i457–i466. <https://doi.org/10.1093/bioinformatics/bty294>
37. Zunair, H., & Ben Hamza, A. (2021). Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 139, 104699. <https://doi.org/10.1016/j.combiomed.2021.104699>