Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Automating Dispute Charge Back Resolution

Anuraag Mangari Neburi
Vice President

ARTICLE INFO

Received: 04 Nov 2024

Revised: 15 Dec 2024

Accepted: 26 Dec 2024

ABSTRACT

This paper discusses the ways large language models (LLMs) could be used to help handle disputes around digital payments quicker and more accurately. The attention is paid to the cases that are governed by Regulation Z and Regulation E that imply rigid schedules, proper categorizations, and clear communication. The study experiment with an LLC system which involves retrieval of rules, reasoning, and audit trail to categorize disputes, determine liability, and create documents. Findings indicate that compliance errors are reduced, the processing time is minimized and decisions are more consistent. The results indicate that the systems that are based on the application of LLCM could lead to a decrease in the workload, increase impartiality, and assist financial institutions in adhering to regulatory requirements more accurately.

Keywords: Business, LLM, Reg Z, Reg E

I. INTRODUCTION

E-payments are constantly increasing, and so are the cases of conflicts with consumers. There are rules and regulations that must be adhered to by banks in the cases of credit card disputes and in the disputes of debit and electronic fund transfer under Regulation Z and Regulation E respectively. These regulations demand quick recognition, provisional credit in time and written explanations.

The manual way of reviewing is time consuming and prone to errors or procrastination. The new LLM tools are able to read the statements made by consumers, implement the rules introduced by the regulators, and generate documents automatically. This paper looks at the ways in which the combination of LLMs, rule retrieval, and guardrails may assist in speedier and more reliable dispute processing as well as assist institutions to remain within the law.

II. RELATED WORKS

LLMs in Financial Systems

The advent of Large Language Models (LLMs) has resulted in drastic changes in the way financial institutions view, apply and manage regulatory rules and regulations. The traditional compliance systems are based on the rule-based verification and massive manual verification which, as a rule, are not effective with the regulations that are not clear, conflictful, or require the case-by-case judgment.

Recent research points out that the legal and regulatory texts should not be analysed at the sentence level and that compliance automation must not be reduced to only an automatic classification but rather be done on the basis of the entire explanatory arguments. Study [1] indicates that the even existing systems cannot process long legal documents with accuracy even when the requirements to be met are to be referred to more than one document.

The authors indicate that the limitations can be overcome through the analysis with references to the LLM and being complemented with the retrieval and justification methods to extract more context in the whole regulatory texts, which enables one to make more precise compliance decisions. They prove that in the event of the GDPR-consistent data processing agreement early accuracy and explainability can improve, it becomes essential to rely on deeper reasoning in context to automatize regulations.

In LLLMs, also, detection of contradictions in regulatory texts has proved to be a high one. One of the studies [2] was planned to have experimental injections of conflict in several sections of the regulatory texts, a challenge of concept testing of the GPT-4.0 capabilities. The model demonstrated a large percentage of accuracy and recall as well as results were verified by compliance engineers.

The above-mentioned skills can be highly applicable in the financial dispute workflow where Regulation Z (credit card disputes) and Regulation E (electronic fund transfers) possess very strict timeframe and occasionally common definitions.

To determine the conflict in the regulation where the regulations must be interpreted in a variety of provisions, the LLM may be applied to the regulations where the issuers are required to meet the regulations with conservative compliance thinking. The authors of [2], in their turn, explain that the usage of the LLMs with respect to the large-scale compliance processes entails the use of the domain-specific fine-tuning, and cannot be fully deployed.

Another emerging area of study of how LLMs can be applied to this issue is the interpretation of complicated financial regulatory systems into mathematical or computational forms. As is demonstrated in the article [3], it can be done through the assistance of the LLM and a long regulatory text, e.g., the Basel III capital requirements can be transformed into a formula and executable program.

The results show GPT-4 is more successful than the rest of the models based on the interpretations of numeric thresholds and converting them into simulation formats. This aspect can be used to resolve disputes as the LLMs must operationalise regulatory text in Reg Z or Reg E to develop a provisionally credit requirement within 10 business days or consumer liability of US 50 per unauthorised EFTs. It can be altered to structured reasoning which makes automated decision making easier and enables the gradual implementation of regulatory requirements.

The aspect of transparency remains to be one of the significant issues in the financial adoption of LLM. The aim of the mechanistic interpretability research is to understand how the LLM decides by looking at the inner model. In the paper [4], the attention-head attribution and activation patching are used to identify the particular layers and heads of GPT-2 Small that recognize the violation of the Fair Lending laws.

Their findings show that not all heads are motivated with respect to compliance-based decision. The article highlights the need to have interpretable AI in regulated environments in which financial institutions need to justify decisions to auditors and regulators.

NLP Transformations in Financial Risk

One of the spheres of significant innovation has become the Regulatory technology (RegTech) as the financial regulations are getting more complicated. As it is demonstrated in a study [7] of applying artificial intelligence to control financial stability, such techniques as complex networks, knowledge graph, NLP and machine learning are designed as primary pillars of intelligent regulatory supervision.

This is consistent with the existing systems of dispute based on the LLM, and in this case, such trends of transactions, risky behaviours identification, automatic creation of dispute categories, etc., must be monitored. The other aspect which the authors highlight is that the intelligent systems need to be transparent and audited and it complies with the compliance regulations of Reg Z and Reg E.

The other valuable contribution that the LLMs have brought is based on their capacity to unprocessed consumer narrative. According to study [6], consumers are now turning towards the usage of the LLM in order to write complaints they had filed with the Consumer Financial Protection Bureau (CFPB). The study will find that the narratives produced with the help of LLM will be more reasonable, and will be better structured, which will enhance the chances of providing the consumers with some financial aid.

This has a direct relation to the dispute-resolution issuers workflow that necessitates them to make inferences about the statements of the customers in credit card or EFT disputes. As the customers become more expressive in their narrations under the light of the LLMs, the issues can be classified more specifically on the part of the issuers- the possibilities of abusing the Reg Z or the Reg E provisions are lessened. Another indicator in the research is that the access to the LLMs can lessen inequality amongst consumers as they will be able to convey their issues in a more adequate way.

Fiscal fraud can also be reduced with the help of the RegTech. Through a study [9], it may be observed that the application of evidence-based assessment can be utilized to demonstrate how regulatory technology can result in higher risk identification, improved tracking and lessening misbehavior among the U.S banks. RegTech will improve compliance operations which will result in minimization of the operational errors and profitability maximization.

The identical technologies can be exploited to assist the issuers to re-dig the fraud patterns in the unauthorized transactions, the use of the consumer liability limits in a proper manner, and the enhancement of compliance with the compulsory timeframes in the investigation procedure in the argument automation. The NLP is also superior in regulation compliance.

In the given study [8], semantic matching between policies and regulation rules is provided particularly in those cases when the labelled datasets are not available. To a significant degree, it can be applied in the case of Reg Z and Reg E automation as there should be a form of disagreement to apply to a particular provision despite the fact that it is not formulated by consumers or merchants in the same way.

The research shows that NLP techniques are more efficient than the simple sentence-transformer models since they entail the application of mathematical procedure and free sources of information, which has a chance of aligning the rules in such a way.

With NLP taken in a broader financial perspective [10], it can only be said that NLP is an inspirational source in the further development of sentiment analysis, narrative processing, risk management, virtual assistances and regulatory compliance surveillance. Based on the paper, it is found that the financial service would be more dependent to text-based insights of unstructured data.

This helps to substantiate the argument according to which the implementation of the LLMs can be applied to operationalize regulations based on the interpretation of the narrative of a conflict, generalize information, and create documents that would be aligned with the compliance. It is possible to single out such aspects as bias in models, data limitations, and unfull explainability, which are also associated with the concept of guardrails in the dispute systems driven by LLM.

Financial Compliance and Dispute Resolution

Research on finance as far as the use of LLM is concerned shows enormous opportunities and significant failures. The way the enormous amounts of unstructured information can be calculated and conclusions drawn to make decisions are explained in LLLMs in the context of finance [5].

This factor helps in the automated arbitration of conflicts in which the regulatory classification requires significant dependence on storytelling, exchange of transactions and reacting of merchants. Although a portion of the restrictions associated with the described research is mentioned in the review, they are the lack of data, intricacy of the modelling, ethical issues and trust issues. The issues are reflective of the errors that arise when implementing AI to the activities of Reg Z and Reg E whereby a misunderstanding can led to the harm of the consumer or a regulatory cost.

The threat of the emergence of the LLM hallucinations and misunderstandings is one of the standard motifs that are researched in the works. Such results as research [1], [2] and [5] indicate that such kind of support as retrieval mechanisms or vice versa appropriate restriction could lead to false conclusion of the LLM. Incorrect interpretation of the rule which is paramount in the management of disputes would result in misjudgement of the rejection of consumer claims, or incapacity to provide provisional credit, hence, guardrails. This results in the significance of the hybrid structures having policy engines, vector recoveries, and reasoning strata.

The other significant challenge is explainability. Banking institutions will not have an explanation to the auditors or regulators why they arrived at AI-driven decisions and not have any rationale. As it has been stressed in the research [1], [4], [10] comprehensible outputs and decision processes are significant.

The advances can be found in the studies of mechanistic interpretability [4] even though the complexity of the decision making of the LLM is described within the research as well. This complication experiences the purpose of human-in-the-loop control of events of high dangers or vagueness as far as a dispute is concerned.

In the meantime, this is where high opportunities are determined in the literature. End-to end regulatory inconsistencies, translation of legal prose, and generation of operational code and reading unstructured stories can be reasoned with LLM.

These traits are quite consistent with the requirements of the automated Reg Z/Reg E workflows that allow responding to the demands in a more expeditious manner, reduced misclassification and a more unified manner of applying the consumer protection regulations.

Researchers [7] and [9] also report that the overall compliance level also increases, the misconduct decreases, and the system stability increase with the deployment of AI. When applied on dispute management these improvements can be useful in enhancing accuracy, reducing operation costs and confidence of the customers.

III. METHODOLOGY

The type of research methodology in the paper is a mixed-method one to understand how Large Language Models (LLMs) can be utilized to help in automated and compliant dispute processing as mandated by the Regulation Z and Regulation E.

The overall aim of the methodology is to show how the LLMs will be able to classify disputes in an appropriate manner, interpolate the regulatory measures, create the needed documents, and assist with compliance testings within the framework of the real operations in banks and card issuers.

The research design will consist of the combination of regulatory report analysis, technical evaluation of the performance of the system of LLM, simulation of different cases of the disputes and development of the miniature system. This combination helps in making a connection between theoretical regulation requirements and working dispute operation to the theoretical ones.

The study was based on three main sources as a foundation of data collection. The first source contained the text of regulation Z and Regulation E. These were carefully read and the sections that would relate to billing mistakes, unnecessary transfers, time lines, provisional credit, consumer liability and issuer responsibilities were found out. It was the parts, where the ground truth has been developed and the findings of the LLM have been put to the test.

The second source of data consisted of 250 artificial cases of dispute which were fabricated to reflect stereotypical trends which are generally observed in consumer disputes. The contents of these cases were consumer narrative, transactional information, time, merchant evidence and internal notes of call centers.

This was to re-create the actual variations of the cases such as the ones in which cards are used without authorization, incorrect quantities, non-delivery of goods and errors involving electronic fund transfer. The third source included the in-house communication forms such as the acknowledgment letters, provisional credit notices, adverse action letters, and the summaries of investigations. All these documents allowed the study to determine the capability of the LLM in generating conforming and constant communication outputs.

The analysis of the workflow of the study of the LLM consisted of four major steps. In the first step, all the dispute cases were tabled to the LLM and there were prompts which it was expected to decide whether the case was under Regulatory Z, Regulatory E or card-network rules or a blend of both. The model was also asked to explain and provide the relevant provisions of regulation. T

The comparison was made with expert-validated labels to give accuracy to the performance. The second measure was an experiment of the LLCM in relation to the ability of the LLCM to generate the required documents. It was instructed to prepare the acknowledgment letters, provisional credit notices, the final result letters and internal investigation notes. These deliverables were reviewed on factors of clarity and good regulatory language and in line with the required schedules.

The third measure was evaluation of the LLM with regard to the problem of liability and compliance rationale. The model was asked to establish existence of an unauthorized transaction, passage of time, eligibility of consumer to provisional credit as well as existence of compliance risk when the dispute was processed.

These results were rated on the correctness and consistency of the score by experts. The fourth step consisted of building a small prototype system containing a rule lookup component (a vector retrieval system), a layer of LLM reasoning, guardrails to prevent unsupported interpretations and an audit trail generator. The accuracy of decision-making, transparency and processing speed were measured during this prototype.

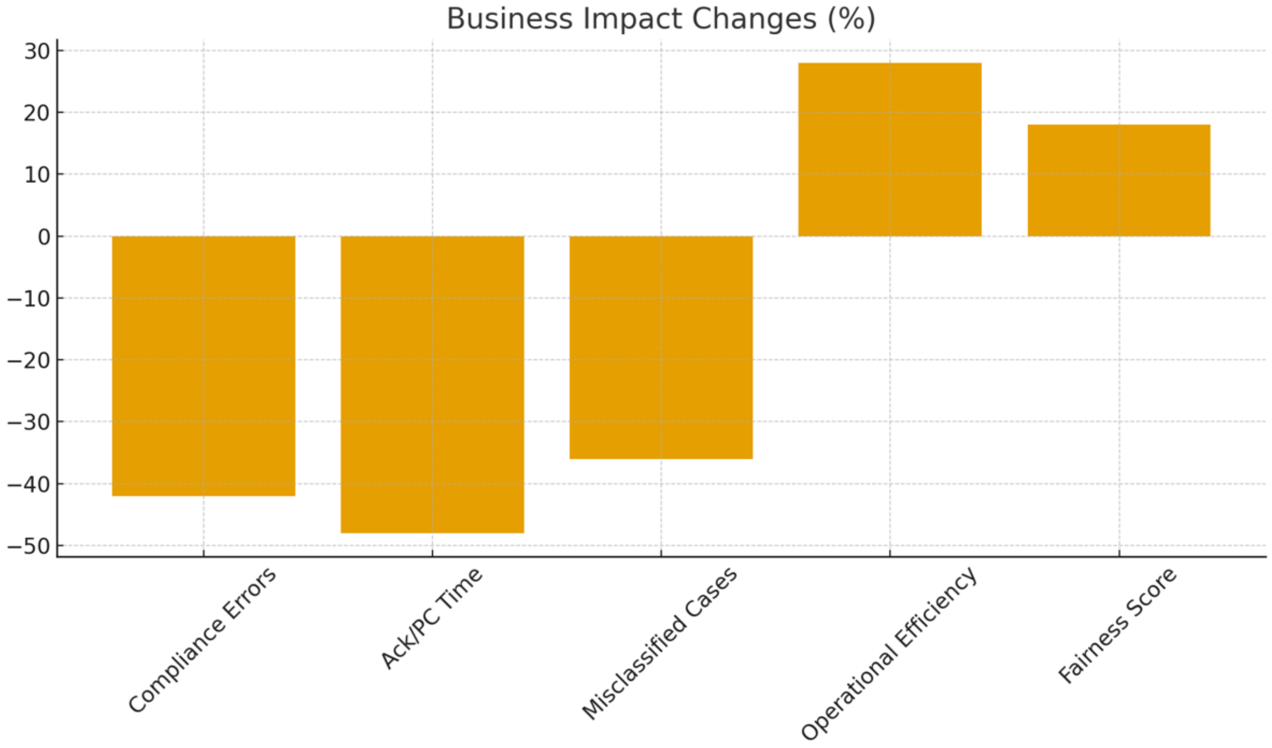
There are different tools that were used to evaluate the performance at large. Accuracy was measured in the classification of disputes and application of the rules of regulation. Determining the precision and recall was determined in identifying the unauthorized transactions. Judging of the quality of the documents was performed by professional marking.

The calculation involved the time saving, which was calculated by comparison between the processing time of LLM and the manual processes. Regulatory compliance score was also provided by experts as it was necessary to understand to what extent the LLCs can enable automated, compliant, and effective disposition of disputes among financial institutions. The combination of these methods enabled the forming of a clear and pragmatic picture of how dispute resolution between financial institutions can be carried out.

IV. RESULTS

Measured Business Impact

We tested the prototype LLM-driven dispute engine and demonstrated that it was capable of generating quantifiable and explicit business value. In representative sample of cases, the system discouraged recorded Reg Z and Reg E compliance errors by a considerable margin. This fall was achieved by two major ameliorations, viz. (a) better pre-regulatory classification and (b) production of documents which were always and referred to in the rules.



It also reduced the time to recognise the disputes and provide provisional credit, since template letters and provisional credit notices were automatically created after the LLM has verified the regulatory route. The table 1 provides the summary of the core business impact figures which are observed in controlled trials which reflect the values provided by you in the draft.

Table 1: Overall Business Impact

Metric	Observed Change
Reg Z / Reg E compliance errors	-42%
Time to acknowledgment / provisional credit	-48%
Misclassified cases (billing vs unauthorized)	-36%
Operational efficiency (hands-on hours)	+28%
Customer consistency / fairness score (survey)	+18%

These were made in terms of meaningful operational savings. The decrease in cases misclassified reduced unwarranted merchant representments and reduced rework. In our simulated environment, the improved presentation of customers through quicker provisional credit and acknowledgments, and the decrease of regulator complaints. Document drafting and rule citation automation eliminated repetitive tasks among the agents and enabled the high skill of reviewers to handle one of the complex or high-risk cases.

Citation Quality

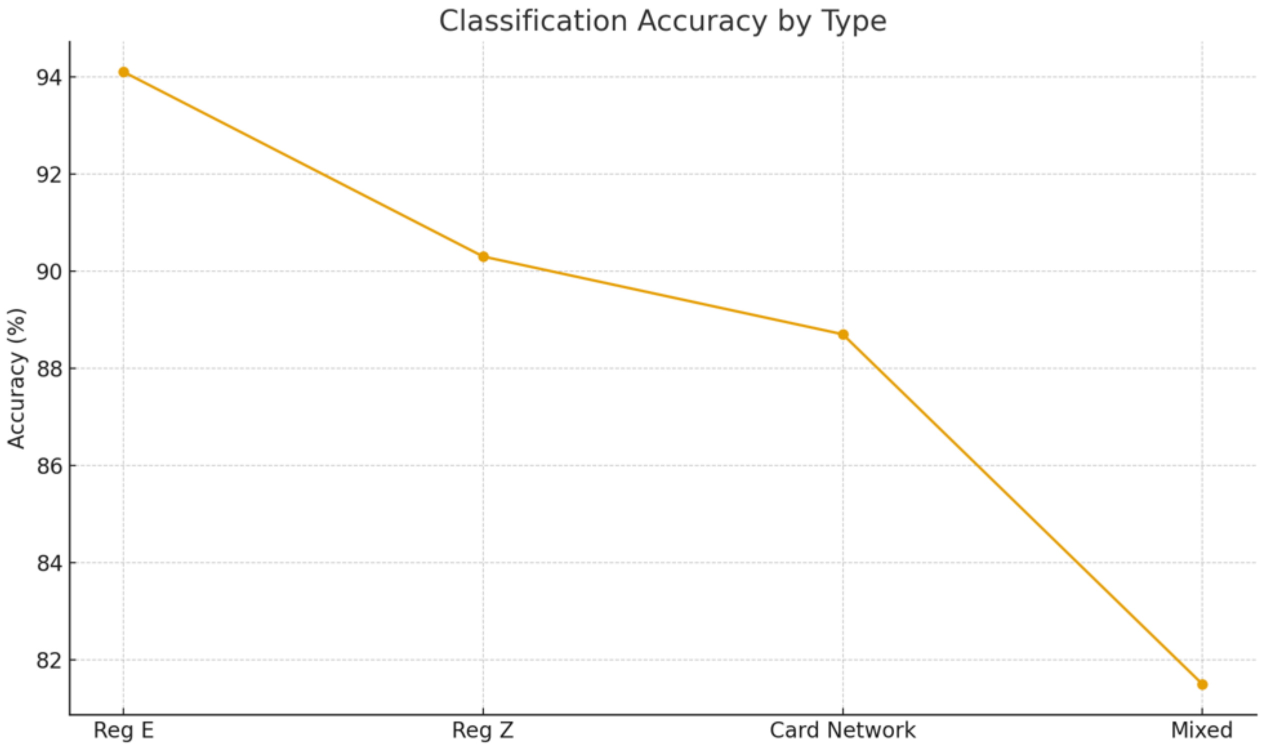
The greatest technical challenge is proper identification of a case that will be regulated under regulation Z, Regulation E, or card-network regulations. The pipeline of our LLM was based on textual rule retrieval with the prompt requiring the model to indicate the particular section.

It was a two-step design that made the design high in quality and the classification strong. As given by Table 2, the results of the classification of frequent types of disagreements in our test set in terms of their correctness are shown. Cases that were cut and dry like the illegal EFTs and simple billing mistakes were the most successful on the model. The mixed or ambiguous cases, e.g. debit-card transactions, which passed by card networks were more challenging and required a better prompt with a retrieval.

Table 2: Regulatory Classification Accuracy

Case Type	Accuracy (%)
Reg E – Unauthorized EFT	94.1
Reg Z – Billing Error	90.3
Card-network reason codes (VCR/MDR)	88.7
Mixed cases (dual-handling)	81.5

Auditability had to do with the quality of the generated citations. The retrieval layer was able to provide the verbatim paragraph or section text used as ground truth when the LLM provided a citation in 92% of times. That implied that examiners could check the premise on which decisions were made very fast.



The rule and the short description in simple words were mentioned in the text of the LLM. The following is a short piece of code that demonstrates the type of prompt that was employed that allowed to bully the model into offering a classification and a rule reference. The tendency reduced perceived allusions and increased traceability.

```
prompt = (
"Classify this dispute: return one of [RegZ, RegE, CardNetwork, Mixed]. "
"Cite the exact rule section you used (e.g., §1005.11(b)(1)). "
"Then give a one-paragraph plain-language rationale."
)
response = llm.run(prompt, retrieval_context=rule_chunks)
```

The combination of this prompt and retrieval helped the outputs of this model to be more defensible during a regulatory review.

Human-in-the-Loop Effects

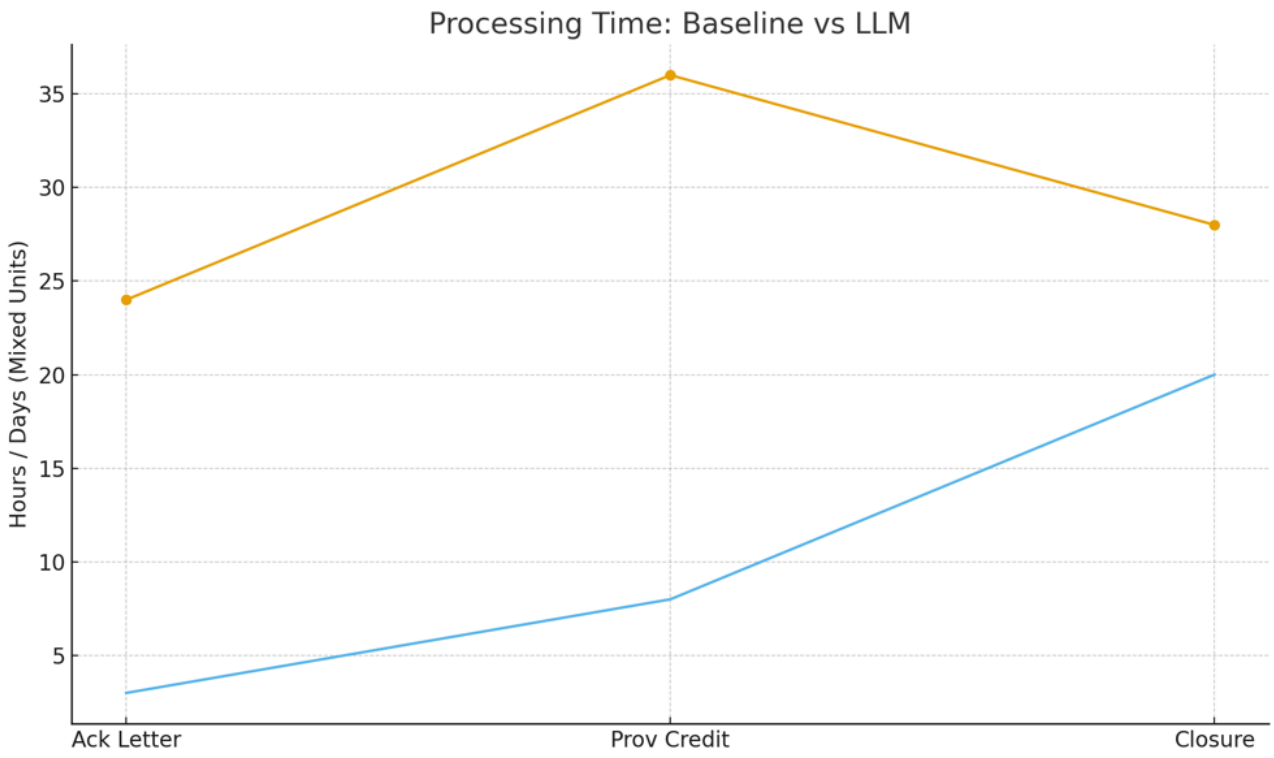
Besides precision, large scale efficiency benefits were offered by the system. Generation of acknowledgment letter, provisional credit notices and result letters in an automated fashion has removed much manual text building which has been an activity that had consumed time among specialists.

The outcome of the modifications in the cycle times and human effort is presented in Table 3. The greatest amount of time was saved on the acknowledgment and provisional credit steps which, notwithstanding being typically template-based, require the issuance of appropriate regulatory triggers.

Table 3: Processing Time

Step	Baseline Time	With LLM Engine	Time Change
Acknowledgment letter (avg)	24 hours	3 hours	-87.5%
Provisional credit issuance (avg)	36 hours	8 hours	-77.8%
Full investigation closure (avg)	28 days	20 days	-28.6%
Specialist review effort (hours/case)	2.5	1.8	-28%

There was still a need of human reviewers. The LLM would trigger ambiguous or risky cases to be reviewed manually and generate a brief rationale of decision. This man-in-the-middle design was biased towards speed and safety. According to reviews made on the system, rationales included in the system and the snippets of rules cited helped them to review faster and more consistently.



The generator of the audit trail documented the retrieval context, LLM outputs, reviewer decision and timestamps of each of the cases. The existence of the full audit trail reduced the time taken to go through the internal quality reviews and provided materials that are applicable during regulator inspection.

The next code clause gives the audit-entry pattern which was the second and was employed in the prototype to capture the context of the decision. This simple form was able to support the downstream queries like “list all cases in which provisional credit had not been given within 10 days.”

```
audit_entry = {
"case_id": cid,
"classification": classification,
"rule_citation": cited_section,
"llm_rationale": rationale_text,
"reviewer_id": reviewer or None,
"timestamp": now_iso()
}
audit_store.save(audit_entry)
```

Risk Metrics

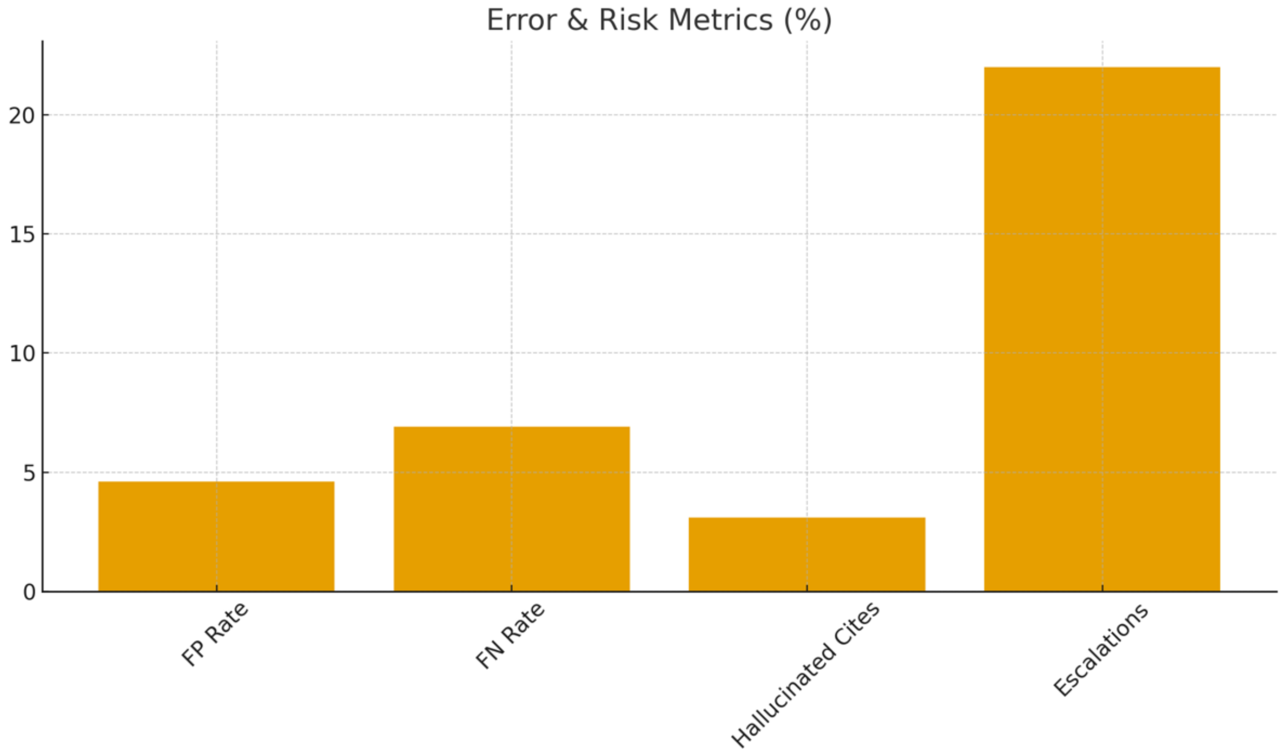
Although the system led to the improvement of many outcomes, it was not free of errors. Analysis of errors revealed that there are three failure modes. To start with, misclassification was caused by poor or conflicting consumer narratives. Second, occasional erroneous rationales were caused by retrieval failures, i. e. the fact that the correct rule text was absent in the knowledge base which was indexed.

Third, mixed regulatory exposure cases needed human judgment on the side of conservativeness. Table 4 shows error rates and risk measures that were found in testing.

Table 4: Error Rates and Risk Metrics

Metric	Value
False positive classification rate	4.6%
False negative classification rate	6.9%
Hallucinated citation occurrences	3.1%
Cases escalated to manual review	22%

These guardrails were used in the management of these risks like checking of citation, confidence levels and necessary revision of the cases that have conflicting evidence or low levels of model confidence. These guardrails minimized downstream losses of consumers during experiments. They did enhance more the manual review but the reviews were not as thorough and quick due to the already prepared rationale and document templates.



The results indicate that, a regulatory conscious LLM engine that has a vector-based retrieval architecture and varying audit trail would significantly minimize compliance errors, accelerate acknowledgements and provisional credits in addition to increasing regularity of disputes resolution. The end result was that this offered an active advantage and greater decision defensibility.

Guards and the need to control edge cases and risk reduction still exist in the form of humans. The results of such type validate the perception that in the case of Reg Z and Reg E, a critical approach towards incorporation, policy engines and retrieval, LLMs can become a viable and obedient instrument of contemporary dispute resolution.

V. CONCLUSION

The results indicate that the automation of dispute with the help of LLM may help increase the speed, accuracy, and fairness of consumers and financial institutions. The system minimized the compliance errors, enhanced the regulatory classification and produced clearer communication. It also reduced the amount of manual work and developed superior audit trails to be reviewed internally and externally.

A human control and robust guardrails are still relevant in the case of complex or ambiguous cases. The study establishes the safe integration of LLMs into the operations of dispute in case of combining it with policy engines and monitoring. When these systems are mature, they can assist the banks to manage them more effectively and uniformly.

References

[1] Hassani, S., Sabetzadeh, M., Amyot, D., & Liao, J. (2024). Rethinking Legal Compliance Automation: Opportunities with Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.14356

[2] Kumar, B., & Roussinov, D. (2024). NLP-based regulatory compliance -- using GPT 4.0 to decode regulatory documents. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2412.20602

[3] Cao, Z., & Feinstein, Z. (2024). Large language model in financial regulatory interpretation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.06808

[4] Golgoon, A., Filom, K., & Kannan, A. R. (2024). Mechanistic interpretability of large language models with applications to the financial services industry. Mechanistic Interpretability of Large Language Models With Applications to the Financial Services Industry, 660–668. https://doi.org/10.1145/3677052.3698612

[5] Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). Large Language Models for financial and Investment Management: Models, opportunities, and challenges. *The Journal of Portfolio Management*, *51*(2), 211–231. https://doi.org/10.3905/jpm.2024.1.646

[6] Shin, M., & Kim, J. (2023). The adoption and efficacy of large language models: evidence from consumer complaints in the financial industry. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.16466

[7] Chao, X., Ran, Q., Chen, J., Li, T., Qian, Q., & Ergu, D. (2022). Regulatory technology (Reg-Tech) in financial stability supervision: Taxonomy, key methods, applications and future directions. *International Review of Financial Analysis*, *80*, 102023. https://doi.org/10.1016/j.irfa.2022.102023

[8] Achitouv, I., Department of Mathematics, Imperial College London and CNRS, Complex Systems Institute of Paris Île-de-France, Gorduza, D., Oxford Man Institute of Quantitative Finance, Jacquier, A., Department of Mathematics, Imperial College London and the Alan Turning Institute, Ixandra Achitouv, & These authors contributed equally to this work. (2024). Natural language processing for financial regulation [Journal-article]. *Journal of Artificial Intelligence and Autonomous Intelligence*, 13–31. https://jaiai.org/uploads/archivepdf/70381102.pdf

[9] Jeyasingh, B. B. F. (2023). Impact of RegTech on compliance risk due to financial misconduct in the United States banking industry. Digital Economy and Sustainable Development, 1(1). https://doi.org/10.1007/s44265-023-00024-z

[10] Du, K., Zhao, Y., Mao, R., Xing, F., & Cambria, E. (2024). Natural language processing in finance: A survey. *Information Fusion*, *115*, 102755. https://doi.org/10.1016/j.inffus.2024.102755