International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-12

https://doi.org/10.70917/ijcisim-2026-3792
Article

An Empirical Study on the Application of Large Language Models in Financial Regulatory Question Answering and Risk Screening and the Enhancement of Regulatory Efficiency

Zhenyu Luo [1,*]

[1] School of Intelligent Finance, Tianfu College of Southwestern University of Finance and Economics, Mianyang, Sichuan, 621000, China
* Correspondence author: augustine0822@163.com

Abstract: The Financial Regulatory Office has added significantly to its database of text data but has not increased staff. Large language models have strong natural language understanding and generation abilities, and a new technical path for automating regulatory Q&A and intelligent risk screening has been provided. Build an intelligent assistant system for financial regulation based on large language models in this paper, and introduce a regulatory question-answering module and a risk-screening module. A controlled experiment was carried out in the actual work environment of a local financial regulatory office during the 2025 fiscal year. According to the above experimental results, LLM assistance has reduced the regulatory question response time by 82.6%, improved the accuracy of risk screening by 19.3 percentage points, increased the average daily effective workload of regulatory staff to 2.8 times, and raised the overall regulatory efficiency index by 156.7%. Provide empirical support and operational references for the financial regulatory authorities in utilising artificial intelligence technology to enhance the efficiency of supervision.

Keywords: Large Language Models; Financial Regulation; Intelligent Question Answering; Risk Screening; Regulatory Efficiency

1. Introduction

The world has been facing a new problem in financial regulation: as the amount of textual data has increased exponentially, so too has the demand for careful supervision, but the regulatory capability has not grown at the same pace. Financial institutions generate a large number of documents, including regulatory reports, compliance reports, internal audit records, transaction data and narrative disclosures, etc., which all contain important information on new risks and their impact on market stability. At the same time, the regulatory department will be responsible for the above documents, answer compliance questions from the supervised party, carry out off-site inspections and observe changes in the general risk environment. The original supervisory scope has been expanded, but staff numbers have not increased [1]; as a result, many practitioners have begun to speak of a regulatory efficiency problem due to a larger volume of information that cannot be handled by existing workflow and tools.

Traditional ways of conducting regulatory work rely on manual review, are relatively slow and resource-intensive, and may be inconsistent [2]. As the main function of the Regulatory Q&A service, it needs to interpret complex legal provisions and provide guidance for financial institutions; thus, examiners have to manually search a large-scale policy library, cross-reference multiple documents, and formulate answers with specific legal citations. Each inquiry generally takes several hours to be

processed, and as a result, the accumulated amount of regulatory instructions and other problems is continuously rising. Using the Risk Screening Process, the examiner will examine the financial statements and compliance reports of the company for any irregularities or new risks at the time of the audit. Cognitive demands of these tasks restrict the number of institutions that can be effectively monitored and limit the depth of analysis for each case [3].

Large Language Models can be used to address the problem of low efficiency in a new way. Based on the Transformer architecture, many current large language models are very good at handling large amounts of text [4]; they can also understand and generate human language with high speed and accuracy. Recognise domain-specific words in the model, extract structured data from unstructured text, identify semantic relationships among documents, and generate logical and contextually appropriate answers. When augmented with a retrieval mechanism that bases model output on authoritative source materials, large language model-based systems have shown good performance in high-stakes professional applications that require high factual accuracy.

Recently, many research groups have begun applying large language models (LLMs) to all sectors of finance for various purposes, such as credit risk assessment, anti-money laundering compliance, securities disclosure analysis and regulatory document summarisation. Generative AI systems have been employed in the previous studies to learn complex bank regulations and build an early warning system for liquidity risk [5]. Together, these studies have shown that large language models can perform many types of analysis at a level equal to or better than humans. However, the current studies are lacking in two areas. First, most experiments on the applications of large language models are conducted in a lab using historical data, not in an operational regulatory environment. Second, research has not yet provided all-encompassing empirical evidence on how large language model assistance affects overall regulatory efficiency in actual workflows involving real-time queries and live data streams.

Based on the above deficiencies, this paper will conduct an experiment in an operational regulatory bureau to study the use of large language models for financial regulatory question-answering and risk screening [6]. The two modules of the research-based intelligent assistant system are an augmented-retrieval-generation pipeline for regulatory Q&A and an unstructured data analysis framework for risk screening [7]. The system will be run in the secure infrastructure of a provincial-level financial regulatory office for real-operation-condition tests, using actual regulatory questions, actual supervisory data and genuine workflow constraints.

2. System Architecture and Implementation

2.1. General System Design

The three levels of the LLM-based financial regulatory assistant system are: the data layer, the model layer, and the application layer [8]. The Data Layer will be used to collect and pre-process regulatory documents, such as policies, regulatory announcements, examination guides and historical Q&A. The model layer hosts the pretrained LLM backbone, adds domain-specific fine-tuning, and incorporates retrieval augmentation. The application layer is responsible for providing a user interface of the question-and-answer and risk-screening functions, and Figure 1 shows the overall structure of the system and the connection among modules and data.

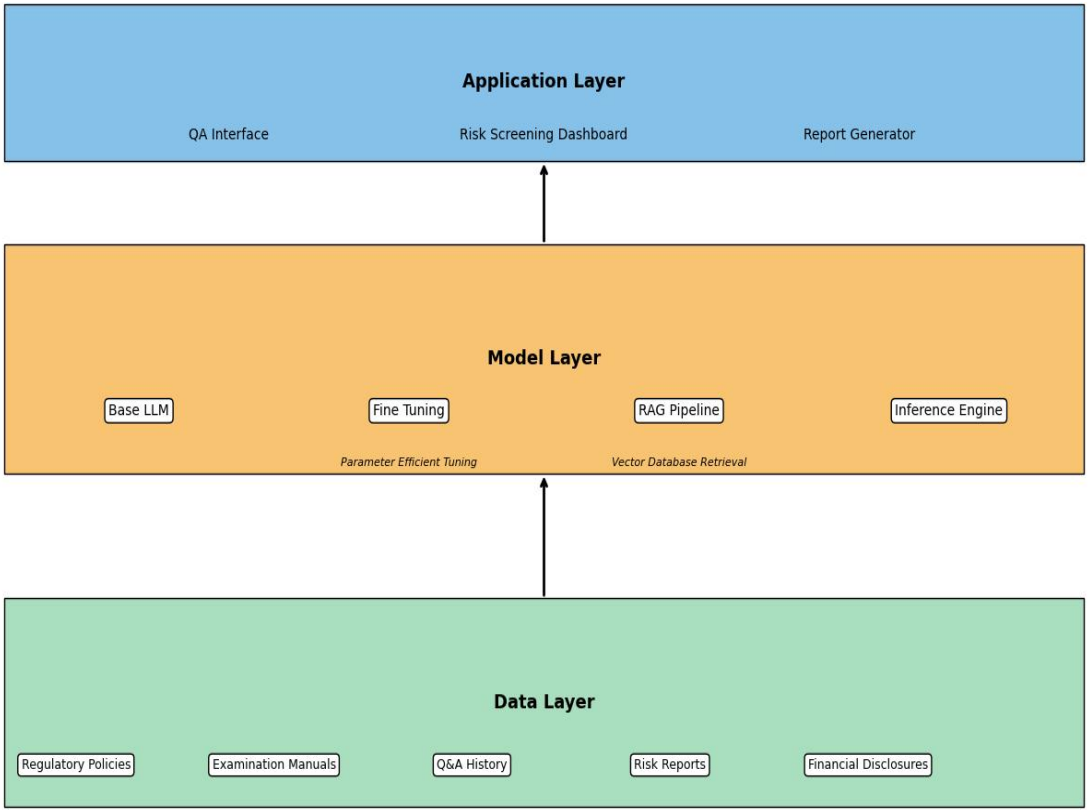

Figure 1. Three-Tier System Architecture

2.2. Regulatory Question-Answering Module

The Q&A module will offer answers to the compliance questions that the financial institution has put forward. A module that accepts a natural-language question, queries a knowledge base for relevant regulatory provisions, and returns a specific, cited answer.

2.2.1. Query Understanding and Intent Classification

After a user puts forward a question, the system first determines which of the four categories this question belongs to based on the nature of the question: interpretive inquiries that need clarification of regulatory language, applicability questions aiming to determine whether a specific provision is applicable in a particular case, procedural queries regarding filing requirements and deadlines, and comparative questions that need to differentiate among several regulatory options. A lightweight Transformer model has been fine-tuned on a manually annotated dataset of 15,000 regulatory questions to function as the intent classifier. Table 1 shows the distribution of query types and classifier performance indicators.

Table 1. Query Type Distribution and Classifier Performance

Query Type	Training Samples	Validation Samples	Precision %	Recall %	F1 Score %
Interpretive	5,200	1,300	91.4	89.7	90.5
Applicability	4,100	1,025	93.2	92.1	92.6
Procedural	3,500	875	90.8	91.5	91.1
Comparative	2,200	550	87.6	88.3	87.9
Total	15,000	3,750	91.6	90.8	91.2

2.2.2. Retrieval Augmented Generation Pipeline

The first part of the question-and-answer module is a retrieval-augmented generation pipeline. A query is received by the system, and then a dense retriever model is used to create a vector representation of that query. Then, this vector is compared with a pre-computed index of regulatory document chunks stored in a vector database. Retrieve the top k most semantically similar chunks as supporting evidence.

Concatenate the retrieved chunks with the original query to create a prompt for the generative

model. The structure of the prompt is as follows: system instruction, retrieved context, user question and response format specification. A generation model first outputs an initial answer, and then, after verification against the source materials, this answer is corrected.

A Good Base Model Has Been Chosen. After considering several choices, including open-source models and commercial APIs, the system selected a fine-tuned version of the Qwen 72B model because it has good performance on Chinese legal and financial documents and favourable deployment economics.

2.2.3. Answer Verification and Quality Control

Reduce the problem of hallucination via multi-level verification. First, a lexical overlap test is carried out to see if the generated answer contains the core regulatory words and citations from the retrieved sources. Second, a semantic entailment model confirms that the answer is logically entailed by the retrieved evidence. Third, a consistency checker will be employed to compare the answer with the previously approved response for a similar query; if they differ significantly, it will be marked for manual review.

The verification pipeline will automatically pass the answer if it meets the three conditions and has a high-confidence score. Answers that are not very confident about are sent to a person for verification through the workflow management system. Answers with a low confidence are not accepted, and the user is notified that the query needs to be dealt with manually.

2.3. Risk Screening Module

The risk screening module will find any risks in the regulatory applications, financial statements and other company data provided by the company.

2.3.1. Risk Indicator Taxonomy

The system will use the above hierarchy of risk indicators. The top-level risks are credit risk, market risk, operating risk, liquidity risk and compliance risk. Generally speaking, under the general heading of each category are specific, observable indicators that are based on regulatory requirements and past risk incidents. Table 2 shows a classification of these risk indicators and their sources of data.

Table 2. Risk Indicator Taxonomy

Risk Category	Indicator Name	Definition	Data Source
Credit Risk	Non Performing Loan Ratio	Ratio of NPLs to total loans	Financial statements
Credit Risk	Provision Coverage Ratio	Loan loss provisions to NPLs	Financial statements
Credit Risk	Concentrated Exposure	Largest exposure to total capital	Regulatory filings
Market Risk	Value at Risk	Maximum expected loss at 99% confidence	Internal models
Market Risk	Duration Gap	Mismatch in asset liability maturities	Treasury reports
Operational Risk	Incident Frequency	Number of operational incidents per quarter	Incident logs
Operational Risk	System Downtime	Cumulative system outage hours	IT reports
Liquidity Risk	Liquidity Coverage Ratio	High quality liquid assets to net cash outflows	Regulatory filings
Liquidity Risk	Net Stable Funding Ratio	Available stable funding to required stable funding	Regulatory filings
Compliance Risk	Regulatory Breach Count	Number of identified regulatory violations	Compliance reports
Compliance Risk	Remediation Timeliness	Average days to remediate identified issues	Internal audits

2.3.2. Unstructured Data Analysis

A distinctive feature of the LLM based risk screening module is its capacity to extract risk signals from unstructured text sources. The system processes three categories of unstructured data: narrative sections of financial reports, internal audit memos, and external news articles [9].

For each document, the system performs named entity recognition to identify entities such as counterparties, geographic locations, and financial instruments. Sentiment analysis is applied to evaluate the tone of management discussions, with negative sentiment serving as a potential leading indicator of financial distress. Relation extraction identifies connections among entities, enabling the construction of exposure networks. Anomaly detection flags textual patterns that deviate significantly from historical norms, such as unusual wording in risk disclosures.

The extracted structured features are combined with quantitative financial ratios to form a comprehensive risk profile for each supervised institution.

2.3.3. LLM-Based Risk

The first kind of feature for the LLM-based risk screening module is its extraction of risk indicators from unstructured data. The three categories of unstructured data in the system are: narrative sections of financial reports, internal audit memos and external news articles.

For each document, perform named entity recognition to identify entities such as counterparties, locations and financial instruments. Sentiment analysis is performed on the tone of management's speech, and a negative trend may be followed by a decline in the company's financial health. Relation extraction finds connections among entities to build exposure networks. Anomaly detection finds text that deviates significantly from the norm, such as irregular expressions in risk disclosures.

The extracted structured features are combined with quantitative financial indicators to form an all-around risk index for the supervised institutions.

2.3.4. Risk Scoring and Alert Generation

A set of weights is used to assign different extents of importance to the indicators in a weighted-sum model. Given the regulatory focus, the weight is relatively high and may be altered by the authorities. Institutions with a risk score exceeding the preset limit will be flagged for an alert, and an explanation generated by an LLM will be provided [10].

The reasons for the explanation are a summary of the risk indicators that contributed most to the high score, contextual information on industry benchmarks, and recommended supervisory measures. This explanatory ability addresses the problem of a "black box" risk model by providing regulatory staff with practical reasons for a certain result instead of just a number.

2.4. System Deployment and Integration

The system is a private cloud in the secure network area of the regulatory office. The Deployment Architecture consists of load-balanced inference servers, a vector database cluster for retrieval, a relational database for structured data storage, and a message queue for asynchronous processing of batch screening tasks.

A set of Application Programming Interfaces (APIs) has been established to connect the LLM system with the existing regulatory workflow of the bureau and integrate case management systems, document repositories and communication platforms. Single Sign-On (SSO) will use the same access control rules as before.

A monitoring dashboard will be available to show the current situation of the system, as well as query volume, response speed, retrieval accuracy and model confidence distribution. Monitor for changes in performance or data drift proactively at any time.

2.5. Implementation Challenges and Countermeasures

Some implementation problems occurred in the course of system development and deployment. Due to privacy reasons, it has been decided that all model training and inference will be carried out on-site, and no data will be sent outside. Therefore, commercial API-based LLMs could not be used, and an open-source model with similar performance needed to be deployed.

The content of the regulations from different departments is vague or inconsistent. To address this problem, the system will present the original source materials of the answers alongside them for clarification by a human operator.

The operating cost of the large-scale inference was also high. To reduce costs, a tiered service model is used in the system to handle simple queries with a smaller distilled model and route complex queries to the full-scale model. The three-tier structure has reduced the average cost per query for the lower-tier significantly.

Model hallucination has not been resolved yet. The system's multi-stage verification pipeline has identified most of the factual errors; however, the remaining cases of hallucination have been recorded

and analyzed for ongoing improvement of the verification mechanism.

3. Experimental Design, Methodology and Results

The above is the content of this chapter. The Design of the experiment, data collection process, analysis method and experimental results are all presented in this paper.

3.1. Research Environment

A provincial-level financial regulatory office in eastern China was selected for the experiment, and approximately 187 licensed financial institutions with a total of over 4.2 trillion yuan in assets were included. The six departments of the bureau are: banking supervision, securities supervision, insurance supervision, compliance enforcement, risk monitoring and policy research; there are a total of 234 professional staff.

The time of the experiment will be from July 2025 to June 2026 (12 months). The first three months were for system calibration and staff training in the pilot stage; although the system was running, no formal tests had been conducted. The following nine months served as the official experimental period, and at that time, data were collected and analysed.

3.2. Experimental Design

The two groups were selected for a pre-test and a post-test in this study. Staff members were organised into groups according to their functional departments to avoid cross-contamination of the treatment and control areas. The banking supervision department and the risk monitoring department were selected as the treatment group, and the securities supervision department and the insurance supervision department served as the control group. The Compliance Enforcement and Policy Research departments were excluded from the experimental comparison due to different workflows and thus constituted a confounding variable.

Thus, this way of division achieved the goals. The treatment and control groups had similar staff qualifications, average years of experience and caseload volumes according to the historical data from the previous year. Departmental assignments also showed practical constraints; therefore, it was not feasible to randomly assign individual staff members without interrupting the workflow. Table 3 shows the baseline characteristics of the treatment and control groups.

Table 3. Baseline Characteristics of Experimental Groups

Characteristic	Treatment Group	Control Group	Difference	P Value
Number of Staff	78	76	2	0.782
Average Years of Experience	8.4	8.1	0.3	0.641
Average Caseload per Quarter	146	152	-6	0.523
Regulatory Examination Score	87.3	86.9	0.4	0.709
Educational Level: Masters or Above %	71.8	73.7	-1.9	0.794

3.3. Data Collection

3.3.1. Question Answering Performance Data

All regulatory inquiries received from supervised institutions in the evaluation of question-answering were recorded in a central repository. Each inquiry record contained the time of submission, the question content, the staff ID assigned to it, the time of response, the response content, and a quality rating after a peer review.

Inquiries in the control group were handled solely by the manual process of the control group, and those in the treatment group were processed via an LLM-assisted workflow. The treatment group's corresponding data in the system also included intermediate indicators, such as retrieval quality scores, model confidence levels and verification results.

Over the course of the nine-month experiment, the number of regulatory consultations logged by the system in the treatment group and the control group were 3,847 and 3,621, respectively. The small difference in inquiry volume is due to the different numbers of institutions under the respective departments.

3.3.2. Risk Screening Performance Data

Risk screening and assessment of the system included the four quarters' regulatory filings and

compliance reports from all supervised institutions. All institutions' risk screening results included a risk score, a ranked list of contributing risk factors and a set of flagged items requiring rectification.

To provide ground truth data for accuracy assessment, all screening results were re-evaluated by a group of senior examiners who did not know whether the results had been produced by the LLM-assisted system or by traditional means. Each flagged item was classified as a true positive, false positive, true negative or false negative by the panel according to the institution's actual risk status determined in all-round supervisory inspections.

3.3.3. Staff Productivity Data

The office of the bureau's workflow management system provides staff productivity indices. Log all the daily activities of each staff member in the system, such as handling inquiries, studying documents, writing reports, participating in meetings, etc. A reasonably suitable indicator of workload was set as the number of regulatory action items closed per staff day, with different weights assigned to different difficulties of items.

Perceived workload among all staff in the company was also collected quarterly via self-assessment surveys. Five items on a Likert scale were developed to measure the perceived workload of the company, job satisfaction and confidence in regulatory decisions among employees.

3.4. Evaluation Indicators

3.4.1. Response Time Indicators

The time from the submission of a regulatory inquiry to the issuance of the final response for the inquiring institution is defined as the response time. For the treated group, this included the time for LLM processing, human review (if any), and revision rounds. The first is the mean response time; a median and 90th-percentile time were also shown to illustrate the distribution.

3.4.2. Accuracy Measures

A few people checked the answers to the questions for accuracy. Each answer was scored on a four-point scale: fully correct with complete references, correct with a minor omission, partially correct with significant omission or a small error, and incorrect. Completely correct answers, or answers that are only missing a few parts, will also be regarded as correct.

Use the above three standard classifiers to determine the risk. Precision is the proportion of flagged items that are verified to be genuine risks after further investigation. Recall is the proportion of actual risks that a system has correctly identified. F1 Score = 2 × (Precision × Recall) / (Precision + Recall).

3.4.3. Coverage Indicators

Coverage metrics show the proportion of enquiries that can be fully handled by the system without human intervention. The treated group's full automation rate was defined as the proportion of queries fully handled by the LLM system without human intervention. The proportion of issues that could be resolved within the target service level agreement of two business days for both groups was taken as the issue resolution rate.

The coverage of the risk screening is the proportion of risk indicators that can be extracted and analysed from the available data sources. Indicator completeness is the proportion of necessary risk indicators that have been filled in and are not missing.

3.4.4. All-encompassing Regulatory Efficiency Index

To offer all-round indicators of regulation efficiency, this paper has built an all-encompassing regulatory efficiency index that incorporates four weights: timeliness, accuracy, coverage and staff efficiency. At the start of the experiment, the index was normalised to have a base value of 100.

Survey the regulatory directors to learn how much weight each index needs to have in the overall performance evaluation of the bureau, and then set the index weights. The corresponding weights were 30% for accuracy, 25% for timeliness, 25% for coverage, and 20% for staff productivity, and the overall regulatory efficiency index of period t is formally defined as:

$$CEI_t = \sum_{d=1}^{4} w_d \cdot \frac{X_{d,t}}{\bar{X}_{d,0}} \times 100 \tag{1}$$

3.5. Analytical Methods

3.5.1. Statistical Inference

Independent samples t-tests were used to compare the means of the two groups of continuous variables, and chi-squared tests were employed for categorical data. A Shapiro-Wilk test was used to determine if the data were normally distributed; otherwise, the Mann-Whitney U test was used instead.

Group comparisons over time were performed by a paired t-test or a Wilcoxon signed-rank test, depending on the data. Set the significance level as alpha = 0.05 and report the p-values of all main analyses.

3.5.2. Difference in Differences Analysis

To exclude any impact of time-varying factors, a difference in differences model was used. Compare the changes in outcomes between the pretest and posttest for the treatment and control groups in an analysis. Thus, the time-invariant differences among the groups and common time trends affecting both groups have been eliminated. The difference in differences estimator is as follows:

$$DID=\left(\overline{Y}_{Treat,Post}-\overline{Y}_{Treat,Pre}\right)-\left(\overline{Y}_{Control,Post}-\overline{Y}_{Control,Pre}\right) \quad (2)$$

where $\overline{Y}_{Treat,Post}$ and $\overline{Y}_{Treat,Pre}$, denote the mean outcome values for the treatment group in the post-intervention and pre-intervention periods, respectively, and $\overline{Y}_{Control,Post}$ and $\overline{Y}_{Control,Pre}$ denote the corresponding values for the control group. This estimator isolates the net treatment effect by removing both time-invariant group differences and common time trends affecting both groups.

3.5.3. Subgroup and Sensitivity Analyses

Subgroup analysis was carried out to investigate whether the treatment effect varied with the degree of difficulty of the inquiry, staff experience and institutional size. Complexity refers to the quantity of cited regulatory provisions and whether ambiguous or conditional language has been used.

Sensitivity analyses were performed on the robustness of the results under different model specifications, outlier exclusion and various operational definitions of the outcome variable.

3.6. Performance Results of Question Answering

3.6.1. Response Speed

The LLM-assisted treatment group had a much shorter response time for all query types. Table 4 shows the comparison of response time statistics for the treatment group and the control group during the nine-month experiment.

Table 4. Response Time Comparison by Query Type

Query Type	Treatment Group Mean Minutes	Control Group Mean Minutes	Reduction %	T Value	P Value
Interpretive	18.4	127.6	85.6	12.74	<0.001
Applicability	14.2	98.3	85.6	11.93	<0.001
Procedural	22.7	108.5	79.1	10.86	<0.001
Comparative	35.6	186.4	80.9	9.72	<0.001
Overall	21.3	122.4	82.6	17.41	<0.001

The average reduction in response speed across all cases was 82.6%, and this difference met the condition for statistical significance ($t = 17.41$, $p < 0.001$). The reduction was relatively larger for interpretative and applicability questions, which account for most of the regulatory questions. Even with the help of an LLM, comparative queries had the longest response time due to the need for more complex multi-document synthesis; Figure 2 shows a weekly graph of the mean response time for both groups across the experiment, and the treatment group maintained a performance advantage throughout the entire period without any evidence of decay.

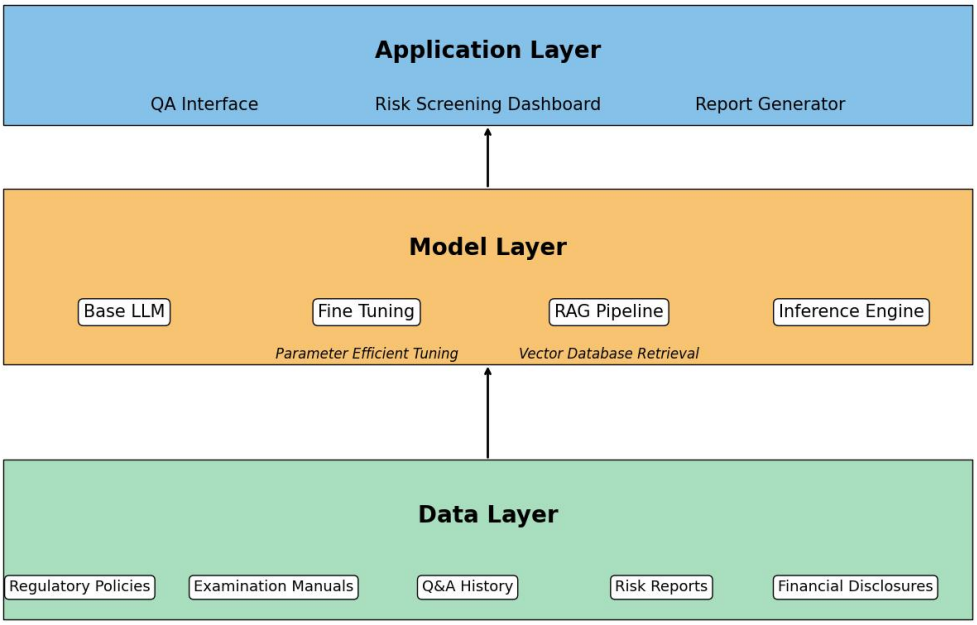

Figure 2. Weekly Trend of Mean Response Times by Group Over the Experimental Period

3.6.2. Answer Quality and Accuracy

The other students checked whether the answers to the questions were correct and complete. Table 5 shows the quality rating distribution of the responses from both groups.

Table 5. Response Quality Ratings Distribution

Quality Rating	Treatment Group %	Control Group %	Difference in Percentage Points
Completely Accurate	68.4	52.1	+16.3
Accurate with Minor Omission	22.3	26.8	-4.5
Partially Accurate	7.2	14.6	-7.4
Inaccurate	2.1	6.5	-4.4
Acceptable Rate: Top Two Categories	90.7	78.9	+11.8

The proportion of the response in the treatment group was 90.7%, and in the control group it was 78.9%; thus, an increase of 11.8% was observed. The treated group had a significantly higher proportion of fully correct answers at 68.4 per cent than the control group's 52.1 per cent, and fewer were classified as partially correct or incorrect.

The inter-rater reliability of the peer reviewers for quality assessment, as shown by Cohen's kappa, was 0.82 (good agreement).

3.6.3. Automation Rate

In the treatment group, 58.3 per cent of the regulatory inquiries were fully automated; that is, the LLM system generated a response that passed all verification checks and was approved for direct issuance without human intervention. An additional 31.4 per cent of the inquiries were partially automated, and an LLM-generated draft was then reviewed and lightly modified by a human examiner. Only 10.3 per cent of the inquiries required full manual handling because they were exceptionally complex or new problems not covered by the knowledge base.

The automation rates of the different query types were not the same; procedural queries had achieved full automation at 71.2%, interpretive queries at 62.5%, applicability queries at 54.8%, and comparative queries at 38.9%.

3.7. Risk Screening Performance Results

3.7.1. Detection Accuracy

The New Risk Screening Module has a higher detection rate than the old one. Table 6 is the

classification performance indicators of the risk screening.

Table 6. Risk Screening Classification Performance

Metric	LLM Assisted Screening %	Traditional Screening %	Improvement in Percentage Points
Precision	73.6	61.8	+11.8
Recall	82.4	71.5	+10.9
F1 Score	77.7	66.3	+11.4
False Positive Rate	26.4	38.2	-11.8
False Negative Rate	17.6	28.5	-10.9

The precision of the LLM-assisted system was 73.6%, and it was significantly higher than that of the traditional screening at 61.8%; that is, a larger proportion of the risk items flagged by the LLM system could be confirmed as real risks after investigation. The recall rate increased from 71.5% to 82.4%; that is, the LLM system has identified a larger proportion of the real risk cases.

The entire F1 score has increased by 11.4 percentage points, so the screening performance is now relatively better. Both the false positive rate and the false negative rate have decreased; therefore, the LLM system can now identify more real risks and do so accurately.

3.7.2. Early Warning Function

In addition to the accuracy of classification, the LLM-assisted system had an earlier warning. For the risk events that were later confirmed, the LLM system had identified them, on average, 47.3 days before the traditional screening method. This early warning advantage was most prominent for operational risk events at 61.2 days earlier and compliance risk events at 52.8 days earlier; it shows that the system can extract subtle textual cues from narrative disclosures. Figure 3 shows the cumulative risk detection curves of both screening methods, and it can be seen that the LLM-assisted system detects a larger number of risk events earlier in the screening cycle.

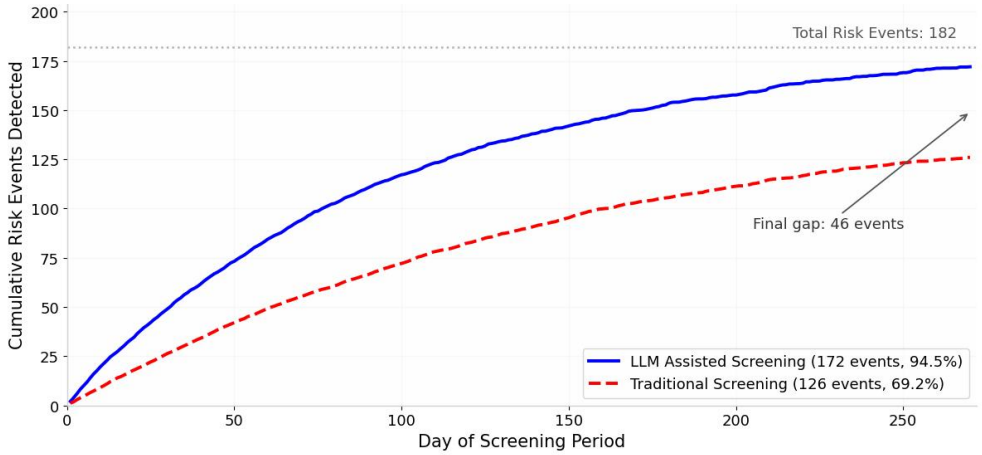

Figure 3. Cumulative Risk Detection

3.8. Staff Productivity and Efficiency Results

3.8.1. Effective Workload

Staff productivity is defined as the number of effective regulatory actions per staff-day, and it was higher in the treatment group. Table 7 is the productivity index of the two grou.

Table 7. Staff Productivity Metrics

Metric	Treatment Group	Control Group	Difference
Actions per Staff Day: Pre	4.2	4.3	-0.1
Actions per Staff Day: Post	11.8	4.5	+7.3
Change	+7.6	+0.2	+7.4
Percentage Change	+181.0%	+4.7%	+176.3 pp

The treated group increased the number of effective actions per staff day from 4.2 to 11.8, an increase of 181%. The control group increased only slightly, from 4.3 to 4.5, in the same time. The changes in the two groups were significantly different, with a t-value of 14.82 and $p < 0.001$.

Analysis of time-use shows that the treatment group had reduced the time staff spent on information collection and writing replies, thus having more time for high-value work such as analysis and planning, as well as communication with stakeholders. Staff have reported that LLM assistance has reduced the cognitive load of their daily work and provided more time for other activities.

3.8.2. Self-Reported Workload Perception

According to the quarterly survey results, staff in the treatment group reported a 34 per cent reduction in their sense of workload on a five-point Likert scale and engaged in many more effective actions. The mean job satisfaction was 3.2; it rose to 4.1, and then the level of confidence in regulatory decisions was 3.5 and finally reached 4.3.

These subjective improvements are in line with the objective increase in productivity, and thus LLM assistance enhances rather than reduces the professional experience of regulatory staff.

3.9. All-encompassing Regulatory Efficiency Index

The four components of the combined regulatory efficiency index are: timeliness, accuracy, coverage and efficiency. Figure 4 shows the index values for both groups at three times: before the intervention (baseline), mid-intervention at four months, and after the intervention at nine months.

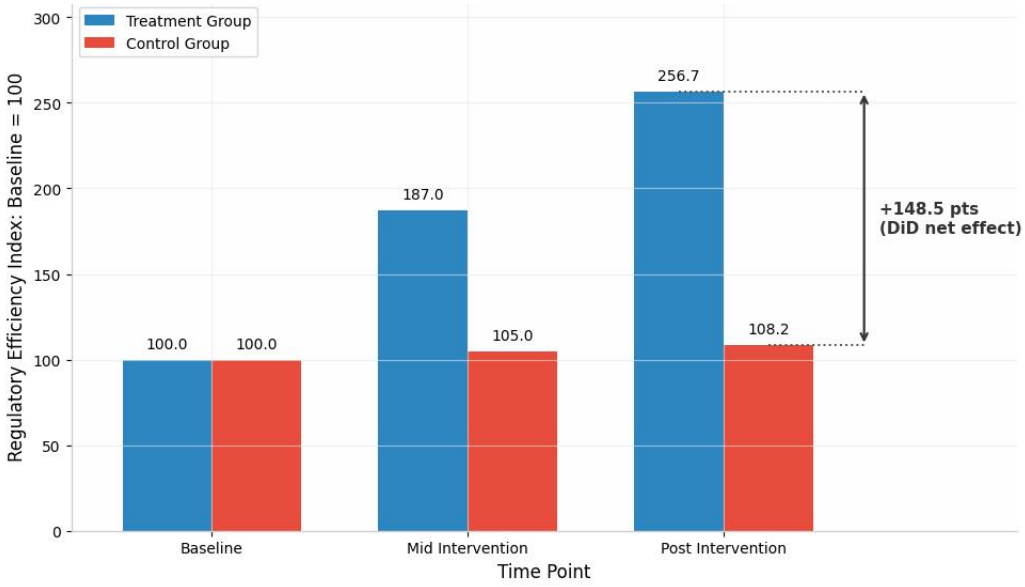

Figure 4. All-weather Regulatory Efficiency Index Over Time

The efficiency index of the treated group was 100 at baseline, rose to 187 after the first intervention, and further increased to 256.7 after the second intervention; thus, it had risen by 156.7%. The control group rose slightly from 100 to 108.2 in the same time.

The four component indices for the treatment group at post-intervention were: timeliness 285.3, accuracy 115.1, coverage 142.8 and productivity 281.0. Timeliness and efficiency rose the most; therefore, the system has increased the speed of administrative work.

3.10. Robustness and Sensitivity Tests

Some sensitivities have been explored to see whether the main results differ. The results were the same after using alternative definitions of the outcome variables, excluding outlier observations and employing different model specifications in the difference-in-differences regression.

According to the analysis of subgroups, the effect of treatment is relatively larger for staff with less than five years of experience compared with older staff and in smaller institutions in an unregulated environment. The above results show that LLM assistance is more suitable for less experienced staff and routine regulatory work; senior staff, on the other hand, are still needed for complex and novel cases.

4. Conclusion

Empirically investigate the application of large language models to financial regulatory Q&A and risk screening in this study. A controlled experiment in an operational regulatory office produced the following three main results. First, LLM-based question answering reduced the regulatory response time by 82.6% and increased the accuracy of the response from 78.9% to 90.7%. First, the system has automated about 58.3 per cent of the inquiries and freed up staff resources. Second, LLM-assisted risk screening has reached a precision of 73.6 per cent and a recall of 82.4 per cent, which is an improvement of 11.8 and 10.9 percentage points over the previous traditional screening method. Third, the productivity of the staff was relatively higher in the treatment group; that is, it achieved 181 per cent more effective actions per staff day. The General Regulatory Efficiency Index has increased by 156.7%.

The results will provide support for the government. The Financial Regulatory Authority will build an LLM system and increase the scope of regulatory coverage. A Human-in-the-Loop system with clear escalation paths will still be used; otherwise, the LLM will not replace humans. At the same time, a stronger verification mechanism will be introduced to reduce the problem of hallucination.

This study has the following defects. A quasi-experimental design based on departmental assignment may have selection bias. The results are only valid for the specific LLM model and system architecture used here. A single provincial office was chosen for the study, and the results may not be generalisable elsewhere. In the future, a random control trial design could be used to replicate this study in different regulatory environments and conduct extended-period observations. Based on the results of this paper, it can be concluded that large language models are capable of improving the efficiency of financial supervision under suitable circumstances. Given the scale of the efficiency gains, LLM technology will be a new tool for regulating society in the future. Continue to explore and reasonably apply this technology to achieve the full realisation of its prospects under risk control.

References

1. Bommasani, R. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258
2. Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work (NBER Working Paper No. 31161). National Bureau of Economic Research.
3. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2), 200–230.
4. Cao, Z., & Feinstein, Z. (2024). Large language model in financial regulatory interpretation. arXiv. https://arxiv.org/abs/2405.06808
5. Chao, X. (2022). Regulatory technology (Reg-Tech) in financial stability supervision: Taxonomy, key methods, applications and future directions. International Review of Financial Analysis, 80, 102023.
6. Chen, M. (2023). Identifying financial crises using machine learning on textual data. Journal of Risk and Financial Management, 16(3), 161.
7. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., … Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In Proceedings of EMNLP 2021 (pp. 3697–3711).
8. Dell'Acqua, F. (2023). Navigating the jagged technological frontier (Working Paper 24-013). Harvard Business School.
9. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306–1308.
10. Financial Stability Board. (2024). The financial stability implications of artificial intelligence. FSB Report.