Dynamic Evaluation Model for Government Budget Performance Using DeepSeek Multimodal Analysis

Qingquan Huang[1,a,*]

[1]School of Business, Dongguan City University, Dongguan, Guangdong, China

[a]718100657@qq.com

*Corresponding author

Keywords: Budget Performance, Multimodal Analysis, Dynamic Evaluation, Time Series Prediction, Intelligent Government Affairs

Abstract: To address core bottlenecks in traditional government budget performance evaluation—static lag, data silos, and disconnection between results and application—this paper proposes a dynamic evaluation model based on DeepSeek multimodal analysis. The model integrates structured and unstructured financial data (e.g., government accounting, policy texts, and project images) through a finance-specific multimodal fusion engine. Using time series analysis (LSTM) and reinforcement learning, it enables real-time performance tracking and deviation alerts throughout the budget cycle. Empirical analysis of three typical financial projects in Dongguan, China (livelihood, infrastructure, informatization), along with five comparative and ablation experiments, demonstrates that the model reduces the evaluation response cycle from an average of 186 days to within 72 hours, achieves 94.7% accuracy in expenditure deviation warnings (>15%), and improves the F1-Score to 93.4%. The study offers an engineering-ready intelligent solution to chronic issues like “funds waiting for projects” and “last-minute spending,” supporting the shift from compliance control to efficiency governance in public finance.

1. Introduction

Comprehensively advancing budget performance management is an inherent requirement for the modernization of the national governance system and governance capacity. However, the current performance evaluation system heavily relies on static indicators and post-audit financial statements, suffering from inherent defects such as evaluation lag, single data dimension, and disconnection between results and application [1]. Particularly in light of the urgent demand for real-time decision support from Guangdong Province's "Digital Public Finance" platform, traditional methods cannot achieve full-cycle dynamic governance covering "pre-event, in-process, and post-event" phases. In recent years, artificial intelligence technologies, especially large language models (LLMs) and multimodal learning, have provided new pathways to solve these challenges [2]. As an advanced general-purpose large model, DeepSeek shows great potential in semantic understanding, logical reasoning, and multimodal information processing [3]. However, recent studies indicate that directly applying large-scale inference models like DeepSeek to specialized fields such as finance and public finance presents engineering challenges, including domain knowledge adaptation, multi-source heterogeneous data integration, and real-time decision support issues [4].

To address these challenges, this paper proposes three key innovative outcomes tailored to the practical engineering needs of government budget performance management. First, a multimodal dynamic performance governance framework in the fiscal domain is constructed, along with an end-to-end engineering solution covering the entire process from data collection and fusion analysis to decision support. Second, a DeepSeek-based fiscal multimodal semantic alignment engine is developed, effectively bridging the semantic gap between structured financial data and unstructured policy documents and image data. Third, a comprehensive experimental validation system is established, encompassing baseline comparisons (including a rule-based real-time monitoring system for objective evaluation of AI component value), ablation studies, and case studies. It is

  DOI: 10.25236/ehmit.2025.022

particularly noteworthy that to tackle the challenge of limited project instance quantity (n=9), we employ rigorous cross-validation and data augmentation strategies, while conducting detailed model failure error analysis to clearly reveal system limitations.

2. Model Architecture and Engineering Implementation

This paper constructs a specific dynamic evaluation model for government budget performance, with its core workflow following a data flow processing logic and integrating specific algorithms and functional modules to support project implementation.

2.1. Data Acquisition and Perception Layer: Omni-channel Data Access

This layer is the "sensory system" of the model, responsible for collecting data from heterogeneous sources in real-time or near-real-time[5].Structured Data Channel: Through secure APIs and ETL tools, directly connects to the Guangdong "Digital Public Finance" platform, automatically extracting time-series data streams such as budget indicators, fund usage plans, treasury payments, and financial accounting records.

Unstructured Data Channel.Policy Texts: Automatically crawls and parses policy documents, special fund management measures, and project implementation plans published on government portals at all levels.Project Image Data: Obtains project site monitoring videos, drone aerial images, and engineering progress photos through government cloud storage or IoT devices.Public Opinion and Social Sentiment: Collects posts and comments related to projects from news media reports and social media platforms, forming public opinion briefs

2.2. Multimodal Fusion and Intelligent Processing Layer: Core Engine

This layer is the "brain" of the model, responsible for transforming raw, heterogeneous data into a unified semantic feature representation for analysis. This is the most significant innovation of this model [6].

2.2.1. Structured Data Processing Pipeline

Cleansing, imputing missing values, and standardizing financial and business time-series data.Extracting key performance signal features through feature engineering, such as: monthly budget execution rate, year-on-year/month-on-month expenditure growth rate, target milestone achievement rate.

2.2.2. DeepSeek-based Unstructured Data Parsing Engine

Fiscal Policy Semantic Parsing Module.Key Technology: We performed domain adaptation fine-tuning on the general DeepSeek model using hundreds of Guangdong Province fiscal policy documents and professional reports, constructing a "Guangdong Fiscal Terminology Knowledge Base" (e.g., accurately understanding concepts like "special bonds," "integration of agricultural funds," "ex-ante performance assessment").

Engineering Implementation.Input policy texts into the fine-tuned DeepSeek model, triggering instructions (e.g., "Please extract performance targets, fund usage scopes, and key time nodes from the following policy") to output structured triplets of (Performance Indicator, Required Value, Basis Clause) [7].

Project Visual Progress Recognition Module.Key Technology.Utilizes DeepSeek-Vision's multimodal understanding capability combined with traditional computer vision algorithms. Engineering Implementation: The model receives project site images, capable of not only object recognition (e.g., identifying tower cranes, construction vehicles, building materials) but also generating qualitative judgments (e.g., "foundation construction phase," "main structure topping out") and quantitative estimates (e.g., "progress is approximately 40% of the total project") of engineering progress based on the project plan, and automatically comparing them with the planned progress.Utilize DeepSeek's sentiment analysis capability to process the collected text data, outputting satisfaction indices for specific projects or negative public opinion warning levels [8].

2.2.3. Cross-Modal Feature Alignment and Fusion

The main technical method of this article is to design a fiscal semantic alignment algorithm, which maps different data modalities to a unified "fiscal performance semantic space" and aligns all features based on timestamps. For example, integrating fiscal expenditure data at specific time points, current cycle targets parsed from policy documents, actual progress indicators determined through image recognition, and public satisfaction indicators into a global multidimensional feature vector. This method can construct a correlation graph of "capital flow policy objectives actual progress social effects".

2.3. Dynamic Evaluation and Early Warning Layer: Model Decision Core

This layer is the "analysis and judgment system" of the performance evaluation model, mainly based on fusion features for real-time evaluation and prediction.

2.3.1. LSTM-based Performance Trend Prediction Module

This article uses Long Short Term Memory (LSTM) networks to model the inherent complex nonlinear time dependencies in the budget execution cycle. The standard LSTM formula is very suitable for this task because it can capture long-range dependencies, which is crucial for linking the early stages of a project with the final outcome.

In addition to standard financial time series features such as monthly budget execution rate, the feature engineering and model configuration used in this article mainly focuses on constructing cross modal time features. These features include the alignment status between financial expenditures and policy milestones, as well as the difference between capital investment inferred from image data and actual progress. The model configuration used this time adopts a single hidden layer (128 neurons) structure to prevent overfitting, and sets a sequence length of 6 months (N) to predict performance indicators for the next 3 months (M). In addition, Adam optimizer and mean squared error (MSE) loss function are used during the training process.

LSTM effectively captures long-term dependencies in budget execution through its gating mechanism. The core computing process is as follows:

$it = \sigma(Wxi * xt + Whi * ht-1 + bi)$ (Input Gate)

$ft = \sigma(Wxf * xt + Whf * ht-1 + bf)$ (Forget Gate)

$ot = \sigma(Wxo * xt + Who * ht-1 + bo)$ (Output Gate)

$ct = ft \odot ct-1 + it \odot \tanh(Wxc * xt + Whc * ht-1 + bc)$ (Cell State)

$ht = ot \odot \tanh(ct)$(Hidden State)

2.3.2. Reinforcement Learning-based Dynamic Threshold Warning Module

This article mainly focuses on innovation and risk management, abandoning fixed warning thresholds and instead adopting reinforcement learning (RL) agents that can dynamically adjust thresholds. Considering that using only 9 project trajectories may lead to overfitting risks, we have implemented two key strategies. The first is state space simplification, where the state of a reinforcement learning agent is defined by a compact set of high-level features - including project type, project phase, standardized expenditure bias, and completion risk predicted by LSTM - rather than raw high-dimensional data. The second is policy regularization and cross item validation, which means that in the training of Near End Policy Optimization (PPO), we adopt conservative learning rates and policy regularization methods. In addition, by using a leave one method cross validation scheme to evaluate the agent, it ensures that its learning strategy can generalize to unseen items rather than just memorizing specific cases.

This model interacts with a simulation environment constructed based on historical project data through reinforcement learning agents, and designs a reward function: $R(s, a)=\alpha * (\text{budget accuracy}) - \beta * (\text{overspending penalty})+\gamma * (\text{early warning bonus})$, where α, β, and γ are weight coefficients. This reward function can guide intelligent agents to form robust strategies even in limited data situations by incentivizing precise expenditures, punishing violations, and rewarding early and correct warnings.

2.4. Decision Support and Application Layer: Value Output

This layer is equivalent to the "execution center" of the model. It will call business rule engines like Drools, which contain business rules extracted by financial experts, to convert the results of intelligent analysis into instructions that the government can actually operate. The system can also automatically output some results: for example, when a warning is triggered, a "Budget Execution Exception Warning Notice" is automatically generated, which clearly states where the problem lies and provides suggestions such as "meeting the project leader" and "holding a coordination meeting"; If it is predicted that a project will continue to be inefficient, a "Budget Adjustment Plan" will be automatically generated, suggesting "slowing down the speed of fund disbursement" and "redirecting resources to efficient projects". Finally, these results are directly sent to the business process of Guangdong Province's "Digital Public Finance" platform through standardized API interfaces, forming an online closed loop of "monitoring → warning → decision-making → execution".

3. Experiments and Results Analysis

To comprehensively verify the model's effectiveness, we designed five sets of experiments, including baseline comparison, ablation experiments, and case studies.

3.1. Experimental Setup and Data

Dataset and Baselines: Our dataset comprises 9 financial projects from Dongguan City (2023-2024). To ensure a fair and rigorous evaluation, we compare our model against three baselines:

Baseline Models: Selected the traditional static evaluation method based on final account data and the quarterly evaluation model based on logistic regression as baselines:

Baseline 1 (Traditional Static Evaluation): The conventional post-hoc evaluation based on annual final accounts.

Baseline 2 (Logistic Regression Model): A quarterly evaluation model based on structured financial data .

Baseline 3 (Rule-Based Real-Time Monitoring): A newly designed baseline for a balanced comparison on timeliness.

This system executes real-time monitoring using predefined, fixed thresholds (e.g., expenditure progress deviation >15%) on the same data streams as our full model, but it lacks the AI components (DeepSeek parsing, LSTM prediction, RL dynamic thresholds). This comparison directly isolates the value added by the intelligent multimodal analysis.

Based on previous work cases, the following performance evaluation examples are presented in detail in Table 1.

Table 1 Experimental Dataset Overview

Project Category	Project Examples	Data Modalities	Data Volume (items/proj)
Livelihood	Basic Public Health Services, Affordable Housing Construction	Financial data, Policy documents, Satisfaction survey reports	15,000
Infrastructure	Road Upgrading, Park Renovation	Financial data, Contract texts, Engineering progress images, Acceptance reports	25,000
Informatization	Smart City Platform, Government Cloud	Financial data, Tender documents, System logs, User feedback	20,000

3.2. Experimental Setup and Data

3.2.1. Dynamic Warning Effectiveness Analysis

The model's warning effect on the key anomaly event "expenditure progress delay >15%" on the

test set is shown in Table 2. The results show that our full model achieves a significant improvement over all baseline methods. Crucially, it substantially outperforms the Rule-Based Real-Time Monitoring system (Baseline 3), confirming that the performance gains stem from the intelligent multimodal analysis and dynamic thresholding, not merely from real-time data access. Multimodal fusion remains the primary contributor to performance gain.

Table 2 Dynamic Warning Performance Comparison

Model	Accuracy(%)	Recall(%)	F1-Score
Traditional Static Evaluation (Baseline 1)	-	-	-
Logistic Regression Model (Baseline 2)	81.3	75.6	78.3
Rule-Based Real-Time Monitoring (Baseline 3)	84.1	78.9	81.4
Our Model (Structured Data Only)	85.2	80.1	82.6
Our Model (Full Modality)	94.7	92.1	93.4

3.2.2. Ablation Study

To verify the contribution of each module, we conducted ablation experiments, with results shown in Table 3. After removing multimodal data, the F1-Score dropped by 10.8 points; after removing the LSTM prediction module, the response cycle significantly lengthened. This demonstrates the necessity of the multimodal fusion and dynamic prediction mechanism proposed in this paper.

Table 3 Ablation Experiment Results Analysis

Model Configuration	F1-Score	Avg. Response Cycle	Key Findings
Full Model	93.4	<72 hours	-
w/o Multimodal Data	82.6	<72 hours	Increased false alarm rate, insufficient ability to identify complex deviations
w/o LSTM Prediction	89.5	~7 days	Unable to provide early warning, reduced value of in-process intervention
w/o Dynamic Threshold	90.1	<72 hours	Generated numerous invalid warnings for finalization phase projects

3.2.3. Evaluation Response Timeliness Analysis

Response cycle refers to the time difference between the occurrence of an anomaly and the generation of a warning or evaluation result by the system. As shown in Table 4, the model in this paper achieves near real-time evaluation.

Table 4 Evaluation Response Cycle Comparison

Evaluation Method	Average Response Cycle	Evaluation Mode
Traditional Static Evaluation (Baseline 1)	186 days	Post-event Evaluation
Logistic Regression Model (Baseline 2)	30 days	Post-event Evaluation
Rule-Based Real-Time Monitoring (Baseline 3)	<72 hours	In-process Dynamic Evaluation
Our Model	<72 hours	In-process Dynamic Evaluation

3.2.4. Failure Case Discussion

We not only look at successful cases, but also carefully analyze failed cases, especially the 5.3% false negatives (i.e. missed warnings). Through these cases, several recurring issues have been identified, which also expose the shortcomings of current multimodal fusion engines

Firstly, compensatory bias masks the anomaly.For example, there were several instances of

underreporting due to one type of expenditure (such as hardware procurement) being spent slower, while another type (such as software licensing) was spent faster. When the two were offset, the overall expenditure progress remained within the dynamic threshold range, and the model did not trigger an alarm. The reason is that the policy analysis module did not take into account the mutual influence of expenditures between different categories, so the internal structure has become imbalanced, but it did not notice it.

Secondly, visual recognition has "semantic bias".In some complex infrastructure projects, the visual progress module occasionally misjudges the later indoor engineering (such as "electrical system installation") as "nearing completion". This misjudgment overestimated the actual progress, and the signal of financial delay was also washed away.

Thirdly, the policy text itself is written vaguely.There is an unclear statement about the milestone deadline in a policy document, with one side stating "Q3 2024" and the other side stating "end of September 2024". The DeepSeek parser defaults to selecting the latter date, but it was delayed and not recognized because the internal timeline of the model has not yet been broken.

Fourthly, unstructured data channels have latency.A small number of missed reports are due to the suspension of ground projects, but the updated images or social media discussion content on site will not be released for a while, and this period becomes a temporary information blind spot.

Overall, this error analysis indicates that although the model is efficient, its performance is constrained by several factors - the quality and timeliness of sensory input, as well as the depth of semantic understanding. Next, we will focus on three things: first, modeling the internal budget structure; second, improving the visual recognition ability of complex scenes; and third, adding confidence scores to strategy analysis.

3.2.5. Robustness Analysis in a Small-N Setting

Some people are concerned that the limited number of projects may affect the conclusion, so we conducted an additional robustness analysis. We focused on whether the warning strategy of the reinforcement learning agent is stable, and found that for specific project types, stages, and deviation states, its threshold judgments are very consistent in different cross validation rounds, with a coefficient of variation of less than 8%. This indicates that the agent is learning general strategies rather than memorizing specific situations of individual projects. In addition, the LSTM model has similar and stable prediction errors between different projects, which in turn proves that feature engineering and regularization methods are indeed effective.

3.2.6. Engineering Application Value Case Study

Taking a "Smart City Informatization Project" as an example, in the 4th month after the project started, the model detected that its hardware procurement expenditure was far below plan (deviation -40%), while the policy parsing module found that its tender documents had clear requirements for localization rates. The model immediately triggered an orange warning and generated suggestions: "Suspected hardware procurement delay due to supply chain issues. Recommendations: a) Project team to explain the situation; b) Financial department to coordinate supplier resources." This warning was 8 months earlier than the traditional annual audit, winning valuable time for project correction, and it is estimated to have avoided approximately 15% of budget fund idle time.

4. Conclusion

This study successfully constructed a dynamic evaluation model for government budget performance based on DeepSeek multimodal analysis technology. Through verification from multiple dimensions, the model has achieved certain results. This model not only significantly reduces response time, but also surpasses rule-based real-time systems in prediction accuracy, fully demonstrating the core value of this model. Through systematic error analysis, we have thoroughly revealed the limitations of the model in complex compensation anomaly handling and semantic understanding boundaries, providing direction for future research. In the future, we will be committed to transforming models into lightweight SaaS services, expanding grassroots application

scenarios, exploring their practical application boundaries in complex scenarios such as government debt risk warning and tax forecasting, and continuously integrating the reasoning capabilities of advanced models such as DeepSeeker R1 to further enhance the system's causal analysis capabilities and adaptive optimization performance.

Acknowledgements

Funded Programs: Guangdong Province 2025 Annual Accounting Research Project, Program No.kj202503-8;Dongguan City College Teaching Quality Engineering Project, Project Name: Kingdee Digital Economy Industry College, Project Number: 2025zlgc001;Dongguan City College Higher Education Teaching Reform Project, Project Name: Digital and Intelligent Empowerment, Craftsman Spirit Cultivation: Research on High-Quality Development of Applied Talent Training in New Business Disciplines, Project Number: 2025yjjg001.

References

[1] Ni Q. Government budget management, tax incentives, and corporate performance[J]. Finance Research Letters, 2025, 86(Part A): 108768.

[2] Wang Z J, Wang L, Wang B, et al. The Orientation and Path of Key Financial Budget Evaluation in Government Budget Performance Management [J]. Financial Supervision, 2025, (15): 46-50.

[3] Dong L. Research on Path Optimization of Government Budget Performance Management from the Perspective of Financial and Accounting Supervision——Based on the Construction of a "Dual Circulation" Supervision Framework [J]. International Business Accounting, 2025, (10): 64-66+74.

[4] He J N. Research on the Impact of Provincial Government Budget Performance Management on the Efficiency of Financial Science and Technology Expenditure [D]. Shandong University of Finance and Economics, 2024.

[5] Song H L, Meng S Y. Research on Digital Technology Empowering Government Budget Performance Management Based on TOE Framework [J]. Western Finance and Accounting, 2024, (02): 10-12.

[6] Yue H J, Wei Q Q. Research on the Application System of Local Government Budget Performance Management Results——Based on the Analysis of System Texts from 11 Provinces [J]. Times Finance, 2023, (10): 58-63.

[7] Zhang G X. Research on the Impact of Local Government Budget Performance Management Reform on Fiscal Expenditure Efficiency [D]. Shanxi University of Finance and Economics, 2023.

[8] Huang Q. Analysis of ABC model for optimizing financial structure of universities in government accounting comprehensive budget performance management[J]. Applied Mathematics and Nonlinear Sciences, 2025, 10(1): 1-18.