Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 5, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

AI-Powered Financial Insights: Using Large Language Models to Improve Government Decision-Making and Policy Execution

REN, Luqing [1*]

[1] Columbia University, USA

**** REN, Luqing is the corresponding author, E-mail: lr3130@columbia.edu***

Abstract: Given the complexity of fiscal data types and the lengthy policy execution chain, this study constructs an application framework for language models supporting government decision-making. It systematically investigates task modules including decision-making question-answering identification, expenditure forecasting modeling, executive summary extraction, semantic matching, and conflict reasoning. The framework elucidates model architecture design methodologies and semantic fusion mechanisms, while introducing response capability simulation testing and performance evaluation systems. Using heterogeneous fiscal corpora and multi-task experimental data, demonstrates that the model exhibits strong performance in accuracy, generative consistency, and generalization capabilities, supporting intelligent applications across diverse fiscal scenarios.

Keywords: Language Model, Semantic Matching, Execution Analysis.

Disciplines: Artificial Intelligence Technology. **Subjects:** Natural Language Processing.

DOI: https://doi.org/10.70393/6a69656173.333139 **ARK:** https://n2t.net/ark:/40704/JIEAS.v3n5a03

1 INTRODUCTION

As fiscal management evolves toward digitalization, the urgent challenge for governments is how to rapidly extract actionable insights from vast, heterogeneous fiscal big data. These datasets encompass structured budgets, semi-structured revenue and expenditure records, and unstructured policy and performance reports, presenting significant difficulties in semantic understanding, contextual dialogue, and information fusion. Enhancing transparency, real-time visibility, and traceability throughout policy implementation is crucial for strengthening government response mechanisms and accountability chains. Current research suffers from insufficient semantic modeling capabilities, inadequate logical reasoning abilities, and poor cross-task generalization, making it difficult to meet the demands of fiscal operations characterized by dynamic changes, high contextual relevance, and multi-domain coupling. The lack of adaptability makes it difficult for these systems to provide stable support for complex task environments in practical government decision-making systems. Significant portability bottlenecks exist, particularly in areas such as data ambiguity, indicator coupling, and overlapping policy objectives [1].

To address these limitations, this paper proposes an intelligent fiscal decision-support framework built upon large-scale language models. The framework integrates a series of specialized task modules, including fiscal question-answering recognition, budget forecasting, executive summary generation, semantic alignment for performance evaluation, and conflict detection with reasoning capabilities. In addition, a multidimensional performance evaluation system tailored to fiscal textual corpora is constructed to assess model robustness, accuracy, and scalability. Through this architecture, the study aims to establish mechanisms that ensure high semantic consistency, contextual credibility, and task-level interpretability in fiscal policy formulation and execution.

2 SEMANTIC UNDERSTANDING METHODS FOR FISCAL DATA

Semantic understanding of fiscal data serves as the foundational step for large language models to achieve information perception and reasoning within intelligent fiscal decision-making systems [2]. Given the complex nature of fiscal data—encompassing both structured budget details and income/expenditure statements, as well as extensive unstructured policy texts and implementation reports—a layered semantic parsing approach is required. First, structured data undergoes field mapping and label expansion to construct a fiscal ontology lexicon, enabling semantic linkage between indicators. For unstructured text, pre-trained language models perform semantic embedding, leveraging

Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 5, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

span annotation and context dependency modeling to extract fiscal entities and their relationships. To enhance semantic consistency, the system introduces cross-modal alignment mechanisms, uniformly tagging fields such as time, amount, and department across text and tables. By constructing a fiscal knowledge graph, the system further achieves entity fusion and semantic completion, enhancing the contextual understanding capabilities of subsequent question-answering and prediction modules. This semantic understanding approach not only improves the model's accuracy in recognizing fiscal semantic relationships but also lays the semantic foundation for subsequent LLM modeling and executive summary generation.

3 LLM MODELING METHODS FOR DECISION SUPPORT SYSTEMS

3.1 Decision-Making Q&A Identification Mechanism

The fiscal decision Q&A identification mechanism aims to accurately determine whether user input constitutes a fiscal question with decision-making intent, then categorize and route it to different processing modules. The system employs a two-layer classification structure based on language models: the first layer extracts question semantic features via BERT embedding vectors, while the second layer uses softmax functions for multi-label decision intent judgment. Specific classifications include budget evaluation, fund allocation, execution progress, and performance review. For an input query x , whose embedding representation is vector h_x , the probability of a specific intent y is calculated as follows:

$$P(y|x)=\frac{e^{W_y\cdot h_x+b_y}}{\sum_{k=1}^{K}e^{W_k\cdot h_x+b_k}} \tag{1}$$

Where: W_y is the weight vector for category y , and K is the total number of intent categories.

To enhance recognition accuracy, the model incorporates a rule-based enhancement mechanism. Prompt templates containing keywords such as "fiscal budget adjustment" and "execution node delay" are constructed to guide the model toward decision-related intents. During training, Focal Loss mitigates class imbalance, while dynamic weight fine-tuning improves recognition of borderline cases. This mechanism effectively supports task routing and strategy scheduling for fiscal Q&A.

3.2 Fiscal Forecasting Modeling Approach

Fiscal forecasting modeling performs forward analysis of key indicators such as future revenue/expenditure trends and budget execution risks within decision support systems. Given fiscal data's dual characteristics of strong structural elements (e.g., time series, account codes) and semantic non-structural elements (e.g., policy constraints, explanatory notes) [3], this paper constructs a hybrid language-numeric dual-channel forecasting model, whose overall architecture is shown in Figure 1. This model incorporates two input modules: the numerical channel receives historical revenue-expenditure data sequences $\{x_t\}$ and employs a multi-layer Transformer to encode temporal dependencies; the semantic channel processes fiscal policy texts via BERT embedding to obtain contextual semantic representations S . The prediction component integrates these channels through cross-channel attention mechanisms, outputting the forecasted fiscal indicator values for the next period $\hat{y}_{t+1}$. The calculation formula is as follows:

$$\hat{y}_{t+1}=f_\theta(\mathrm{Attn}(T(x_{t-n:t}),S)) \tag{2}$$

Where: f_θ denotes the fused nonlinear regression predictor, $\mathrm{Attn}(\cdot)$ represents the multi-head attention function, and $T(\cdot)$ signifies the numerical Transformer encoder.

The training phase employs a weighted MSE loss function, incorporating phased fiscal weighting coefficients α_t to distinguish critical months and enhance prediction accuracy during budget peak periods. This modeling approach significantly improves the model's ability to capture trends in fiscal revenue and expenditure fluctuations, as well as execution risk inflection points, while demonstrating strong contextual adaptability and generalization capabilities.

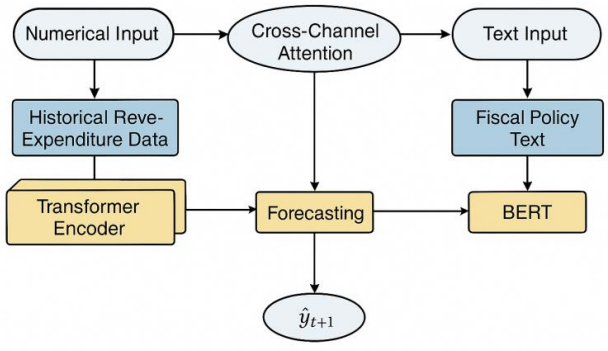

Figure 1: Structure Diagram of the Fiscal Forecasting Model

3.3 Decision Credibility Control

In fiscal decision-making systems, language model outputs directly inform high-sensitivity tasks such as budget recommendations and allocation directives. Rigorous credibility control mechanisms must be established to ensure usability and security. A credibility scoring mechanism is

designed by integrating semantic consistency, data support, and model confidence to dynamically determine the adoptability of model-generated decisions [4]. Given a language model response of ` $A$ `, with corresponding input semantic vector ` $E_x$ ` and output vector ` $E_A$ `, the overall credibility scoring function is defined as follows:

$$\text{Trust}(A) = \lambda_1 \cdot \cos(E_x, E_A) + \lambda_2 \cdot S_K + \lambda_3 \cdot \text{Conf}(A) \quad (3)$$

Where: $\cos(E_x, E_A)$ represents semantic consistency between input and output; S_K denotes matching scores with triplet facts in the financial knowledge graph, serving as a quantitative measure of knowledge support; $\text{Conf}(A)$ is the model's built-in confidence score for generated outputs.

In practical deployment, the system sets a confidence threshold (e.g., 0.75). Outputs below this threshold are automatically routed to manual review channels and tagged with corresponding risk labels (e.g., "contextual inconsistency," "insufficient data support"). Simultaneously, the system enhances the interpretability of model outputs by introducing Attention visualization and sentence-level traceability mechanisms. This provides transparent, traceable reference support for policy formulation, further improving the reliability of the model's practical deployment in fiscal operations.

4 LANGUAGE MODEL OPTIMIZATION MECHANISM FOR POLICY IMPLEMENTATION

4.1 Automated Executive Summary Extraction

To address the vast volume of execution reports and tracking documents generated during policy implementation, we designed an automatic execution summary extraction mechanism combining semantic aggregation and generative control. The system first encodes raw execution text into sentence vectors, extracts sentence-level embeddings ($\{s_1, s_2, \text{K}, s_n\}$) via pre-trained models (e.g., BERT), and then selects summary sentences based on content coverage and positional weighting to enhance content completeness and structural logic. The summary candidate scoring function for this stage is defined as follows:

$$\text{Score}(s_i) = \lambda_1 \cdot \text{Sim}(s_i, D) + \lambda_2 \cdot \text{Pos}(s_i) + \lambda_3 \cdot \text{Redund}(s_i)^{-1} \quad (4)$$

where: $\text{Sim}(s_i, D)$ denotes the semantic similarity between a sentence and the document topic, $Pos(s_i)$ represents the structural position weight of the sentence within the document, and $\text{Redund}(s_i)$ indicates its redundancy with already selected sentences.

During the generative summarization phase, a decoding guidance mechanism based on control vectors is introduced. Key information such as "policy name, target indicators, and implementation timeline" is embedded into the summary template to guide the LLM in generating more focused policy summaries. The training process employs a weighted ROUGE loss function to enhance content coverage and readability. This approach achieves high-quality structured compression of policy implementation corpora and provides standardized inputs for subsequent performance comparisons, ensuring the model maintains strong semantic consistency and content representativeness when interpreting policy execution logic.

4.2 Conflict Detection and Reasoning

During policy implementation, documents generated by different departments, time periods, or data sources may contain semantic conflicts or logical contradictions [5]. Without proper identification, these discrepancies can lead to biases in subsequent performance evaluations. To address this, this paper constructs a conflict detection system integrating knowledge triplet extraction and language reasoning mechanisms. Based on LLM, it automatically extracts fact representations ($\langle h_i, r_i, t_i \rangle$) from implementation summaries and compares them with target logic in the fiscal knowledge graph. Conflict determination is defined by the following consistency function:

$$C = \frac{1}{n} \sum_{i=1}^{n} \delta\left(\langle h_i, r_i, t_i \rangle \text{K}\right) \quad (5)$$

Where: δ denotes the conflict determination function, returning 1 for semantic inconsistency and 0 for consistency; K represents the set of fiscal domain knowledge rules. Conflict types include: amount discrepancies, temporal inversion, broken causal chains, and budget-performance deviations.

Additionally, the system incorporates a natural language reasoning module to construct "premise-hypothesis" pairs, determining whether execution sentences form contradictions with policy objectives to enhance reasoning capabilities. During training, multi-label supervised signals (Entailment/Neutral/Contradiction) improve the model's ability to identify fine-grained contradictions in complex contexts. This mechanism not only ensures internal consistency in policy execution records but also provides a more reliable semantic foundation for performance analysis.

4.3 Performance Semantic Matching Method

Performance semantic matching aims to evaluate the semantic consistency between policy objective texts and implementation outcomes, providing quantitative support at the linguistic level for fiscal assessments. A matching framework based on embedding alignment and multi-scale

semantic aggregation is constructed. Objective statements ($G = \{g_1, g_2, \mathrm{K}, g_m\}$) and implementation statements ($E = \{e_1, e_2, \mathrm{K}, e_n\}$) are encoded separately. Pre-trained models (e.g., RoBERTa) are used to extract sentence vector representations ($\vec{g}, \vec{e}$), and semantic matching scores are calculated via similarity functions:

$$\mathrm{Match}(G, E) = \cos(\vec{g}, \vec{e}) = \frac{\vec{g} \cdot \vec{e}}{\|\vec{g}\| \|\vec{e}\|} \tag{6}$$

where: $\cos(\cdot)$ denotes the cosine similarity function, yielding values between [0,1], with higher values indicating greater semantic consistency.

To enhance adaptability to verbose sentences and abstract expressions, the system further incorporates a Cross-Encoder architecture. This constructs a cross-attention mechanism between target and execution content, enabling end-to-end learning for multi-sentence block alignment structures. Additionally, "Key Performance Point" labels are introduced as auxiliary signals to increase sensitivity to semantic deviations. The final matching score serves as a critical input for the performance monitoring module, aiding in the identification of risk scenarios such as execution deviation, empty-running targets, or false reporting. This supports intelligent fiscal oversight and subsequent scheduling feedback.

5 EXPERIMENTAL VALIDATION AND PERFORMANCE EVALUATION

5.1 Task Set and Evaluation Metrics

To systematically evaluate the performance of the constructed fiscal intelligent decision-making model across multiple tasks, a heterogeneous corpus encompassing five core tasks—decision-making Q&A, expenditure forecasting, executive summaries, conflict detection, and performance matching—was designed. The corpus sources include fiscal policy texts, implementation bulletins, monthly budget reports, and performance review materials. After unified formatting and manual annotation, these materials formed a standardized benchmark dataset. For classification and recognition tasks, accuracy and Macro-F1 served as primary evaluation metrics. Generative tasks were measured using BLEU and ROUGE to assess content coverage and linguistic quality. The semantic matching task incorporated mean cosine similarity and AUROC to reflect the system's sensitivity to semantic deviations. All models operate under custom fiscal scenario prompts. Two adversarial datasets—"high-pressure corpus" and "policy retrospective"—were constructed to test robustness and generalization under extreme inputs. The evaluation framework assesses recognition, generation, alignment, and reasoning capabilities, providing quantitative metrics for subsequent simulation and model selection.

5.2 Response Capability Simulation Testing

To validate the real-time performance and deployability of large language models in fiscal applications, response capability simulations were designed covering request scale, input length, and load fluctuations. Models based on GPT-4 ran on NVIDIA A100 GPUs, with tasks including decision-making Q&A, executive summarization, and performance matching. Table 1 results show that within the 10 to 100 concurrent request range, model response latency increases linearly with concurrency, with maximum delay controlled below 5.2 seconds, indicating stable scheduling. As input tokens increased from 256 to 2048, the average response time per request rose from 0.43 seconds to 2.91 seconds, exhibiting exponential growth, confirming input length as the key performance bottleneck. Under high concurrency (>80), the model maintained stable output quality with a response success rate exceeding 94%, without any abnormal interruptions. These results demonstrate the system's high concurrency support capability and scalability in fiscal business scenarios, making it suitable for multi-department online collaborative invocations.

Table 1. Simulation Results of Model Response Capability

Concurrent Requests (rps)	Input Token Count (tokens)	Average Response Time (s)	Maximum Response Time (s)	Response Success Rate (%)
10	256	0.43	0.68	100.0
20	512	0.79	1.12	100.0
50	1024	1.56	2.41	98.7
80	1536	2.33	4.02	96.3
100	2048	2.91	5.20	94.1

5.3 Model Comparison and Generalization Analysis

To evaluate the applicability and generalization capabilities of different language models in financial tasks, GPT-3.5, GPT-4, and BLOOM-Z were tested on subtasks including question-answering recognition, expenditure prediction, executive summarization, and semantic matching. Table 2 shows that GPT-4 achieved the best overall performance in accuracy, generation consistency, and semantic alignment, with an average accuracy of 91.2%, significantly outperforming GPT-3.5 (85.9%) and BLOOM-Z (82.4%). In the executive summary task, GPT-4's ROUGE-L score improved by nearly 7%, demonstrating its superior text compression and key point extraction capabilities. Figure 2 further illustrates the distribution of the three models across five core capability dimensions. GPT-4 exhibits significant

advantages in semantic reasoning and modality fusion, while BLOOM-Z shows relative weaknesses in execution stability and response control. The results indicate that GPT-4 possesses stronger generalization and transfer capabilities across diverse fiscal tasks, making it more suitable for intelligent government decision-making scenarios in complex contexts.

TABLE 2. COMPARISON OF FISCAL TASK ACCURACY RATES ACROSS MODELS (UNIT:%)

Model	QA Classification	Forecasting Accuracy	Summary ROUGE-L	Semantic Matching	Average Accuracy
GPT-3.5	86.5	84.7	78.9	84.3	85.9
GPT-4	92.4	89.6	85.8	91.2	91.2
BLOOM-Z	81.0	80.5	74.1	84.0	82.4

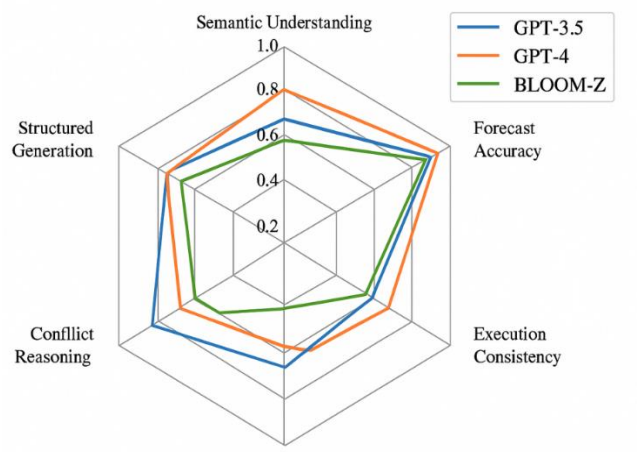

FIGURE 2. PERFORMANCE RADAR OF LLMS IN FISCAL TASKS

6 CONCLUSION

This paper explores the application of language models in intelligent fiscal decision-making and policy execution. It systematically constructs a task framework integrating semantic understanding, predictive modeling, execution parsing, and conflict reasoning. Through multidimensional evaluation, the model's accuracy, stability, and generalization capabilities are validated. Future research may further incorporate cross-modal information fusion and supervised reinforcement mechanisms to enhance the model's ability to analyze logical chains and implicit constraints in complex fiscal scenarios, thereby advancing the continuous evolution of fiscal governance systems toward greater intelligence and precision..

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

REN, Luqing

Columbia University, New York, USA.

REFERENCES

[1] Barbosa, E. D., & Paramo, S. J. (2024). Enhancing board of director decision-making: The impact of government support on risk management and nonfinancial

performance. Business Strategy & Development, 7(3), e413-e413.

[2] Mienye, D. I., Jere, N., Obaido, G., & Others. (2025). Large language models: An overview of foundational architectures, recent trends, and a new taxonomy. Discover Applied Sciences, 7(9), 1027-1027.

[3] Aman, S. S., Kone, T., N'guessan, G. B., & Others. (2025). Learning to represent causality in recommender systems driven by large language models (LLMs). Discover Applied Sciences, 7(9), 960.

[4] Qiu, J., Fang, Q., & Kang, W. (2025). Towards controllable and explainable text generation via causal intervention in LLMs. Electronics, 14(16), 3279.

[5] A, M. A., M, E., M, S., & Others. (2021). Factors influencing financial performance of the government. Academy of Accounting and Financial Studies Journal, 25(3), 1-15.