LLM-Enhanced Theme Identification and Classification of Urban Housing Policy Texts

Xiaodie Yuan[1], Wenkang Zhang[2], Chengle Zhou[1] and Chunshan Zhou[1*]

[1]School of Geography and Planning, Sun Yat-sen University, 510000 Guangzhou, China
[2]School of Economics and Management, Fuzhou University, 350108 Fuzhou, China

Abstract. Urban housing development has always been a key area of focus in urban studies. Accurate understanding of housing policy texts is of great significance for the sustainable development of urban housing. Different from traditional text analysis methods, we have proposed a machine learning (ML) framework driven by a large language models (LLMs) to quickly and accurately extract policy text themes and classify them. We used Mengzi-BERT-Base (MBB) for semantic encoding of policy documents, used GPT-4o to extract policy keywords and grasp the macro logic, and finally integrated the features of the two using a random forest model to output classification results. The results show that this framework divides housing policy texts into 4 core themes, with an overall accuracy rate of 70.25%, and the accuracy rate of individual themes is 75.56%, indicating that this framework has certain application value in urban policy text analysis.

1 Introduction

Housing is not only the most fundamental physical carrier of the residential function in a city, but also influences and shapes the urban social structure and class mobility through its resource allocation [1]. As a typical high-density city in southeastern China, Guangzhou has continuously introduced a variety of policies such as purchase restrictions, loan restrictions, talent housing, urban renewal, and centralized land supply during the transition from "incremental expansion" to "stock optimization" over the past decade, profoundly reshaping the urban housing spatial pattern [2, 3]. However, the continuous inflow of population under the background of rapid urbanization has led to increasingly acute problems such as imbalance in housing supply and demand, regional housing price differentiation, and mismatch in the space of affordable housing, and it is urgent to reveal the underlying mechanism through the analysis of policy texts.

Mainstream methods for analyzing policy texts mainly include manual qualitative analysis and traditional machine learning. Traditional policy interpretation relies on manual in-depth reading. When dealing with massive municipal government bulletins, departmental normative documents, supporting implementation rules, and press conference transcripts, it not only consumes a lot of time and effort but is also prone to selection bias due to the researcher's subjective stance. At the same time, housing policy texts are highly specialized, covering terms from multiple fields such as land, finance, taxation, planning, and population, and there are combinations, conflicts, and iterations of cross-departmental policy tools, which further increase the difficulty of semantic analysis [4]. Traditional machine learning methods can handle large volumes of text, have high classification efficiency, and are highly repeatable, but they rely on high-quality labeled data, have weak ability to identify abstract concepts, and the model training requires professional knowledge and resources [5].

In recent years, large language models (LLMs) have demonstrated outstanding performance in general semantic understanding and reasoning tasks, providing a new technical path for large-scale, high-precision, and interpretable policy text analysis. Scholars have begun to apply them to the auxiliary analysis of policy texts, and the results show that LLMs exhibit strong semantic understanding and reasoning capabilities in text analysis, suitable for handling abstract, complex, and cross-domain policy concepts [6,7]. By collaborating with humans and designing high-quality prompt frameworks, LLMs not only improve analysis efficiency but also enhance the interpretability and scalability of research [8,9]. Moreover, the introduction of the random forest model (RFM) in ensemble learning can more robustly complete the task of topic classification [10].

At present, policy analysis faces challenges such as large-scale text, highly professional terminology, and complex cross sectoral policy linkages. The emergence of large language model provides a new idea for automated policy analysis. Its powerful semantic understanding and generation ability can effectively extract policy topics and cluster analysis. This study focuses on the applicability of LLM in housing policy text analysis, especially its recognition ability in policy topics and clustering, in order to make up for the shortcomings of traditional content

* Corresponding author: zhoucs@mail.sysu.edu.cn

analysis methods. This study focuses on the applicability of LLM and random forest integrated framework in housing policy text analysis, especially its ability to accurately identify and robustly classify policy topics, so as to make up for the shortcomings of traditional content analysis methods.

2 Methodology

The methodological framework of this study aims to systematically evaluate the efficacy of the "Mengzi BERT Base (MBB) + GPT-4o + RFM" hybrid architecture in the identification and classification of housing policy text topics. Specifically, MBB is responsible for fine-grained semantic encoding of policy texts, GPT-4o focuses on extracting policy keywords to grasp the macroscopic thematic logic, and finally, the RFM integrates the features of the two and outputs the classification results, forming an end-to-end policy text analysis process. The overall architecture of the research method is shown in Fig. 1.

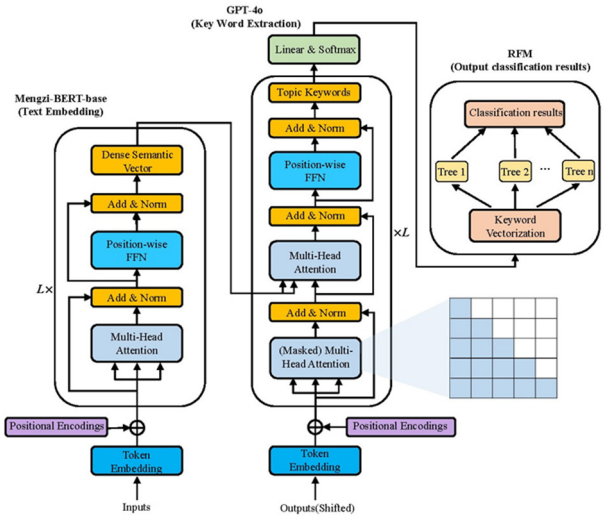

Fig. 1 Schematic Diagram of the Research Methodology Framework.

2.1 Data Collection and Preprocessing

To ensure the timeliness and representativeness of the research samples, this study systematically collected public policy documents related to the housing sector issued by authoritative departments from 1989 to 2025. Ultimately, a policy text corpus was formed, containing over 770,000 Chinese characters and 16,000 sentences. The text preprocessing process mainly includes the following three steps:

• Text cleaning and standardization: Uniformly convert the collected heterogeneous format documents such as PDF and DOC into plain text format, and remove non-text content such as tables, headers, footers, page numbers, and legal provision citation numbers.

• Professional dictionary enhanced word segmentation: By using the Jieba word segmentation tool and injecting professional terms, the negative impact of incorrect segmentation on semantic understanding has been reduced.

• Stop word filtering: A stop word list has been constructed to filter out common words with high frequency but extremely low information content, such as "this method", "Notice", "regulation", etc.

2.2 Model Selection and Combination Strategy

In terms of the model strategy, this study innovatively adopts a hybrid embedding method to integrate the macro-level semantic understanding capability of the general LLM with the micro-level semantic encoding capability of the specialized domain pre-trained model. Specifically, GPT-4o is selected as the general base model, leveraging its powerful generation and context understanding capabilities to grasp the overall intention and logical connection of the policy [9]; at the same time, MBB is chosen as the Chinese pre-trained model for text recognition to accurately capture the subtle semantic differences and professional terms in the policy text [11]. The core strategy lies in constructing the final hybrid feature vector through vector concatenation, and the specific process is as follows:

- Let the input policy text be T, which is encoded by GPT-4o to generate a text vector $V_{\text{GPT}} \in \mathbb{R}^{d_1}$ with dimension d_1;

- The same text T is encoded by MBB to generate a semantic vector $V_{\text{BERT}} \in \mathbb{R}^{d_2}$ with dimension d_2.

- The two vectors are concatenated to obtain a hybrid feature vector V_{hybrid} with dimension $d_1 + d_2$, as shown in the formula below:

$$V_{\text{hybrid}} = \text{Concat}\left(V_{\text{GPT}}, V_{\text{BERT}}\right) = \left[V_{\text{GPT}}^T, V_{\text{BERT}}^T\right] \quad (1)$$

where $\text{Concat}(\bullet)$ denotes the vector concatenation operation, and $\bullet^T$ denotes the transpose of the vector.

2.3 Clustering Approach and Evaluation

In the clustering and analysis phase, the study mainly adopts the RFM for supervised classification. This model belongs to the ensemble learning method, which improves classification performance by constructing multiple decision trees and fusing their prediction results. It is characterized by noise resistance and low risk of overfitting, making it suitable for the high-dimensional hybrid feature vector classification scenario in this study. In model training, hyperparameters such as number and maximum depth of decision trees are optimized through grid search to balance model complexity and generalization ability. The specific formula is as follows:

$$\hat{y} = \arg\max_{c \in C} \sum_{t=1}^{T} \mathbb{I}\left(h_t\left(V_{\text{hybrid}}\right) = c\right) \quad (2)$$

Where $\hat{y}$ denotes the final classification result; C represents the set of classes; T stands for the number of decision tree; $h_t\left(V_{\text{hybrid}}\right)$ is the prediction result of the t -

th decision tree for the hybrid feature vector V_{hybrid}; and $\mathbb{I}(\cdot)$ denotes the indicator function.

2.4 Validation

This study employs principal component analysis to reduce the dimension of the mixed feature vectors. The overall accuracy (OA) is used to reflect the overall classification performance of the model, the average accuracy (AA) represents the average level of performance of each category, and the category accuracy (CA) measures the classification accuracy of a single category. The Kappa coefficient is used to evaluate the consistency between the classification results and the real labels to reflect the robustness of the model.

3 Results and discussion

3.1 Cluster Analysis and Key Findings

This study classified the housing policy documents in Guangzhou and ultimately identified four core themes. These themes reflect the main direction and key areas of Guangzhou's housing policies (Table 1). Specifically, Class 0 focuses on the operation of housing provident funds and financial credit tools for regulating the housing market; Class 1 covers housing system reform and policy dynamic adjustments, including related mechanisms such as land supply; Class 2 emphasizes the construction planning and fair distribution management of affordable housing; Class 3 mainly regulates the order of the housing rental market and the process of property ownership transactions.

Table 1. Classification Results.

Class ID	Theme Name	Representative Documents	Core Keywords
0	Housing Finance (HF)	Regulations on Housing Provident Fund of Guangzhou	Provident fund loans, Credit risks, Financing support
1	Housing System Reform (HSR)	Provisions on the Listing of Purchased Public Housing in Guangzhou	Housing system reform, Policy pilots, System optimization
2	Affordable Housing (AH)	Measures for the Reserve of Affordable Housing Land in Guangzhou	Allocated construction area, Anju Project, Waiting list system
3	Housing Market (HM)	Notice on Further Standardizing Information Disclosure at Commercial Housing Sales Sites	Rental subsidies, Property rights registration, Transaction supervision

As shown in Table 2 and Figure 2, the accuracy verification results show that the OA of the model is 70.25%, the AA is 68.92%, and the Kappa coefficient is 0.5927. This indicates that the model exhibits moderate performance in the classification of housing policy texts. Compared to the ideal accuracy of over 80% for the 4-classification task, this performance has significant room for improvement. The current results indicate that although the model can distinguish the four core themes, its cross-category discrimination ability is still insufficient. For class-specific accuracy: Class 0 achieved high accuracy owing to focused semantics on financial tools and provident fund management with clear conceptual boundaries; Class 1 involves cross-domain adjustments and semantic overlap with Classes 2 and 3, leading to more misclassifications; Class 2 has a higher misclassification rate due to thematic overlap and blurred semantic boundaries; Class 3 performs best, forming an independent semantic domain centered on "rental" and "ownership transactions" with high conceptual concentration.

Table 2 Model Performance Validation Results.

OA	AA	Kappa	CA			
			Class 0	Class 1	Class 2	Class 3
70.25 %	68.92 %	0.5927	72.34%	61.11%	75.56%	66.67%

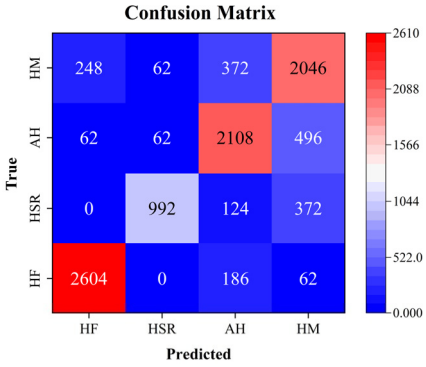

Fig. 2 Confusion Matrix.

3.2. Insights from the Policy Clusters

Firstly, the various thematic categories reflect the key focus of housing policies as they transition from "incremental expansion" to "stock optimization". Class 0 indicates that the policies rely on financial tools to regulate the housing market, providing financial support for the stage of stock optimization. Class 1 demonstrates that the policies adapt to the dynamic changes of the urbanization process, with mechanisms such as land supply playing a crucial role in reshaping the housing market landscape. The frequent appearance of Class 2 reflects Guangzhou's attention to the housing needs of disadvantaged groups, but it also highlights the prominent problem of mismatch in the supply of affordable housing, which aligns with the sharp housing contradiction mentioned in the background. The emergence of Class 3 theme indicates that Guangzhou is alleviating the imbalance between housing supply and demand by regulating the rental market, which is in line with the characteristic of a high-density urban population continuously flowing in. Secondly, through the LLM-driven machine learning approach, this paper not only accurately identified the four core themes of the policy text, but also further revealed the institutional roots behind the policy - for instance, in Class 1, how the land supply strategy, through the adjustment of land use nature,

promotes changes in the housing market pattern. The research shows that the text analysis framework based on LLM can effectively interpret the implicit logic of the policy text, providing a data-driven technical path for revealing the internal mechanism of housing issues and supporting the sustainable development of urban housing.

3.3 Challenges

Although this study validates LLMs' potential in policy text analysis, it faces challenges. First, terminology ambiguity: polysemous and context-dependent terms in policy texts may cause clustering biases, requiring domain knowledge graphs or context enhancement mechanisms for disambiguation. Second, framework generalization: the current classification, designed for Guangzhou's housing policies, lacks tested adaptability to other cities or fields, and cross-domain transfer requires adjusted embedding strategies and metrics. Third, computational constraints: full GPT-4o's high memory usage for long texts reduces efficiency, limiting real-time analysis of large policy libraries. Fourthly, there exists significant potential for optimizing the model's classification accuracy. The current OA of 70.25% represents a moderate performance level in the four-class classification task. Future enhancements could be achieved through the implementation of domain attention mechanisms, the application of SMOTE oversampling for balanced sample distribution, and the integration of ensemble models such as XGBoost. These challenges also delineate specific directions for improvement: by incorporating multimodal data fusion and dynamic modeling, along with domain knowledge integration, sample optimization, multi-model ensemble techniques, and lightweight computational solutions, the framework's robustness and scalability can be further enhanced.

4 Conclusion

This study systematically evaluated the application efficacy of LLMs in the analysis of housing policy texts in Guangzhou. Firstly, LLMs can efficiently and accurately extract implicit themes from a vast amount of policy texts, verifying their outstanding potential in policy semantic understanding and complex concept identification. Specifically, the hybrid model composed of GPT-4o and MBB achieved an accuracy of 75.56% in policy text theme identification and classification, demonstrating a high level of accuracy. This confirms that the macro-context grasping ability of the general-purpose LLM and the micro-semantic parsing ability of the specialized domain pre-trained model are key paths to achieving high-quality policy text clustering. Secondly, this study successfully transformed the LLM-driven analysis method from theoretical conception into an operational empirical research framework, providing a stable and reliable new paradigm for automated and large-scale policy evaluation, significantly surpassing the limitations of traditional analysis methods that rely on manual interpretation in terms of efficiency and scalability. Future research will focus on three directions: multimodal fusion analysis, dynamic evolution tracking, and lightweight model optimization.

Acknowledgement

Thanks to all editors and reviewers. This work was supported by the National Natural Science Foundation of China (No.42371208).

References

1. Zhang, G., Yuan, X., Zhou, C., & Jin, W. From "access to housing" to "access to decent housing": A systematic literature review of China's housing security system. Cities, **162**, 105921. (2025)
2. Zheng, H., Zhang, R., & Wu, J. Value of qualification to buy a house: Evidence from the housing purchase restriction policy in China. Cities, **135**, 104197. (2023)
3. Yuan, X., Dai, X., Zou, Z., He, X., Sun, Y., & Zhou, C. Multi-source data-driven extraction of urban residential space: a case study of the Guangdong–Hong Kong–Macao Greater Bay Area Urban Agglomeration. Remote Sens., **16**(19), 3631. (2024)
4. Browne, J., Coffey, B., Cook, K., Meiklejohn, S., & Palermo, C. A guide to policy analysis as a research method. Health Promot. Int., **34**(5), 1032-1044. (2019)
5. Ahmad, M., Aftab, S., Bashir, M. S., & Hameed, N. Sentiment analysis using SVM: a systematic literature review. I Int. J. Adv. Comput. Sci. Appl., **9**(2). (2018)
6. Zhang, W., Chen, Y., Liu, X., & Zhang, H. A large language model-based approach to building the knowledge graph for master plan: A case study in Guangzhou, China. Land Use Policy, **159**, 107807. (2025)
7. Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. LLM-assisted content analysis: Using large language models to support deductive coding. arxiv preprint arxiv:2306.14924. (2023)
8. Pelaez, S., Verma, G., Ribeiro, B., & Shapira, P. Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents. Quant. Sci. Stud., **5**(1), 153-169. (2024)
9. Singgalen, Y. A. Analyzing an Interest in GPT 4o through Sentiment Analysis using CRISP-DM. J. Inf. Syst. Inform., **6**(2), 882-898. (2024)
10. Biau, G. Analysis of a random forests model. J. Mach. Learn. Res., **13**(1), 1063-1095. (2012)
11. Zhou, C., He, Z., Zou, L., Li, Y., & Plaza, A. HSACT: A hierarchical semantic-aware CNN-Transformer for remote sensing image spectral super-resolution. Neurocomputing, **636**, 129990. (2025)