Information Processing and Management 63 (2026) 104440

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A novel automated policy text evaluation framework integrating PMC into large language models

Xiaobin Lu, Zinan Yang, Chaoguang Huo *

School of Information Resource Management, Renmin University of China, China

ARTICLE INFO

Keywords:
PMC index model
Policy Automated evaluation
Policy informatics
Large language models
Science and technology policy

ABSTRACT

Automated policy text evaluation is a critical research topic in policy informatics. Previous methods rely predominantly on manual variable assignment, making them inadequate for large-scale policy evaluation, while their context-dependent indicators extracted from specific policies lack cross-domain applicability. To address this, we propose a novel policy text automated evaluation framework by redesigning three generalized first-level evaluation indicators that are applicable to policies in any domain, integrating Policy Modeling Consistency Model (PMC) into large language models (LLM), and constructing automated PMC-scoring models based on LLaMA-3-Chinese-8B and Qwen-2.5-7B respectively. Using Chinese S&T policies as examples, we construct the first Chinese policy evaluation dataset with 22,630 labeled policy samples and train four PMC indicator automated calculation models. Compared to the baselines, the model based on Qwen-2.5-7B achieves the best performance in the evaluation of policy character, with an F1-score of 80.41%. The model based on LLaMA-3-Chinese-8B achieves best performance in the evaluation of policy normativity and policy function, with F1-scores of 75.07% and 74.11% respectively. This enables the automated calculation of PMC indices and the generation of a multi-input-output table for comprehensive policy analysis. The application in biosafety policies and data governance policies validate the cross-domain applicability of the framework. As the first framework for automated PMC evaluation, our methodology provides an innovative approach for large-scale, cross-domain policy evaluation.

1. Introduction

Policy evaluation, as a critical component of the public policy process, serves as a core mechanism for assessing **policy formulation quality** and implementation effectiveness, as well as a benchmark for the rational allocation and efficient utilization of public resources (Yang et al., 2021; Wang et al., 2024). The process of policy evaluation is essentially to analyze and evaluate the quality of policy design or implementation effect through the establishment of a multidimensional evaluation framework and a system of standardized procedures (Estrada, 2011). The evaluation process is influenced by multiple factors. For example, as evaluation methods or objectives change, the content and results of the evaluation will also change accordingly. Therefore, ensuring accurate and interpretable evaluation results, necessitates adopting a rigorous evaluation paradigm grounded in data-supported objectivity. This evaluation methodology must also be cross-cutting and universal, with a system of evaluation criteria that can be applied to public

* Corresponding author at: School of Information Resource Management, Renmin University of China, No.59, Zhongguancun Street, Haidian District, Beijing 100872, China.
E-mail address: huochaoguang@126.com (C. Huo).

https://doi.org/10.1016/j.ipm.2025.104440
Received 21 August 2025; Received in revised form 16 September 2025; Accepted 7 October 2025
Available online 10 October 2025
0306-4573/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

policies in any industry or sector.

Among existing policy evaluation paradigms, the PMC index has gained significant traction as a scientifically validated and operationally robust framework for policy text evaluation. The PMC index is a quantitative paradigm used to evaluate and grade the quality and internal consistency of a policy document (such as a legal document, government strategy document, or legislative proposal). Its primary purpose is to move beyond subjective, qualitative analysis and provide a scientific, objective, and reproducible score that reflects how well-constructed a policy is. Since its introduction by Estrada in 2011, this paradigm has been extensively applied across diverse policy fields. For instance, in environmental protection (Liu et al., 2022), the PMC index effectively assesses the alignment between policy implementation and environmental conservation objectives. In the new energy vehicle sector (Yang et al., 2021), it demonstrates strong capability in evaluating the coherence between governmental support policies and industrial development goals. Emergency management studies (Li et al., 2022) further validate its utility in measuring the scientific rationale and implementation consistency of disaster response policies. Additional applications span green development (Ma et al., 2023), corporate social responsibility (Zhang et al., 2023), and COVID-19 pandemic management (Yimsuk et al., 2024). These cross-domain applications not only reinforce the model's adaptability but also highlight its advantages in evaluating the internal consistency of policies.

However, the previous PMC index method still exhibits certain limitations. A primary limitation stems from its reliance on manual variable assignment, as the PMC index requires the use of specific variables to evaluate a policy's strengths and weaknesses. This process is inefficient and restricts evaluations to a small number of policy samples, as demonstrated in Table 1. This constraint significantly hinders its broader application and scalability in large-scale policy evaluation across diverse domains.

To address these problems, we innovatively constructed a novel policy automated evaluation framework integrated PMC into large language models, leveraging the capabilities of large language models (LLMs), particularly those proficient in Chinese, to learn intricate features from policy texts. This approach frames the automated evaluation and scoring of policies as a text classification task. We demonstrate and validate our framework's effectiveness using Chinese science and technology policies as a case study. The aim is to provide a new automated tool for PMC quantitative evaluation, to provide an efficient method for larger-scale policy text evaluation, and to enrich the policy informatics methodology.

The contributions of this study are as follows:

We propose the first LLM-based automated PMC index evaluation methodology, overcoming the limitations of traditional manual classification and human-dependent scoring. Through this intelligent, automated evaluation framework, our study not only improves the efficiency of policy text evaluation, but also improves the model's adaptability and effectiveness in dealing with large-scale policy datasets.

We redesign the first-level evaluation indicators of the PMC index model, which differs from the PMC index evaluation system in other researches in that the modification is that the indicator system is cross-domain generalizable and applicable to any domain of policy, avoiding the use of indicators extracted from specific policy content.

We construct a high-quality Chinese policy dataset for the field of policy text evaluation, featuring comprehensive manual multi-dimensional annotations. This dataset will serve as a foundational resource for future research in this domain.

2. Research objective

Previous studies have focused on evaluating a limited number of policies within a particular domain and assigned variables or scores manually. In addition, in the PMC index model construction part, most studies used ROST data analysis software to assist in screening high-frequency keywords in policy texts (Gao et al., 2022), and extracted the themes, keywords, and key points of policy texts through rooting theory and coding to provide a reference basis for the determination of evaluation indicators. When it comes to in-depth understanding and analysis of policy texts, it still relies primarily on human interpretation. This approach is not only time-consuming and laborious, but also susceptible to research subjectivity, leading to issues of low efficiency and poor accuracy, and cannot guarantee comprehensive coverage of all relevant policies or consistent application of the PMC index model when dealing with large-scale policy datasets. Addressing these issues would facilitate the development of a new automated tool for high-volume policy

Table 1
Existing study on the PMC index model (partial).

Policy Field	Policy Samples	PMC Index Model Application	Source
Waste Sorting Management (WSMP)	22	Conducted quantitative evaluations of 22 WSMPs issued by central and provincial governments in China's Yangtze River Delta region (2013–2021) using text mining and PMC index modelling.	Liu et al. (2022)
New Energy Vehicles (NEV)	11	Developed PMC index surfaces and policy-specific PMC scores to systematically assess the alignment between 11 NEV promotion policies and industrial development objectives.	Yang et al. (2021)
Disaster Relief	10	Implemented an integrated text mining-PMC framework to evaluate the scientific rigor and implementation consistency of 10 representative disaster response policies in China.	Li and Guo (2022)
High-Tech Industries	7	Assessed 7 regional high-tech innovation policies in the Beijing-Tianjin-Hebei (BTH) region through hybrid PMC-text mining analysis to measure their multidimensional effectiveness.	Liu et al. (2022)
Corporate Social Responsibility (CSRP)	6	Designed a CSRP-specific PMC evaluation system to perform content analysis and quantitative benchmarking across 6 CSRP cases selected from multi-level governmental sources.	Zhang et al. (2023)
Green Development	21	Applied the PMC index framework to evaluate the internal consistency and goal achievement levels of 21 national green development policies through computational text analysis.	*Ma et al. (2023)*

text evaluation.

In sum, our research aims to develop a framework for automated evaluation of the PMC index based on an LLM, which can be used to address the problem that the PMC index model can only be applied to the evaluation of a small number of policy texts in the domain. Specifically, the research questions are stated as follows:

RQ1: How to construct a generalized PMC indicator framework applicable to policies in any domain?
RQ2: How to integrate PMC into a LLM to automatically calculate evaluation indicators∞

The remainder of this paper is organized as follows: Section 3 presents the literature review. Section 4 Methods, including the study design, data collection and labeling, and the construction of the core evaluation framework for the PMC index. Section 5 introduces and discusses the PMC automated evaluation model, including model training and performance validation. Section 6 summarizes the paper, articulates the theoretical and practical implications of our study, and outlines the limitations.

3. Literature review

3.1. Policy evaluation methodologies

Policy evaluation is a complex process involving the examination and analysis of policy systems and their implementation. By selecting scientific evaluation criteria and frameworks, it assesses policy effectiveness, thereby improving policy quality and providing a reference for refining future policy formulation and execution. Policy evaluation can be traced back to the period of World War I, where it was primarily applied in the fields of education and health, and the empirical-based method, which utilize experiments and mathematical analysis for evaluation, was widely adopted. Over time, policy evaluation gradually evolved into a relatively independent research field within social sciences. From the policy evaluation standards proposed by Suchman, which innovatively introduced five types of evaluation methods, to Poland's "3E" evaluation system based on economy, efficiency, and effectiveness (Suchman, 1967; Poland, 1974), policy evaluation research embarked on a developmental path evolving from empirical methods into a distinct discipline, continuously evolving on this foundation.

Currently, research on policy evaluation methods is focused on two main aspects. On one hand, it evaluates the effectiveness of policy implementation through measurement indicators. For example, Linder et al. (2020) proposed an algorithm to measure the similarity between the policy recommendations in two bills and conducted empirical research. Li et al. (2014) applied fuzzy mathematics to explore the internal coherence of China's circulation industry innovation policies. Wang (2014) used gray relational analysis to evaluate the performance of regional technological innovation policies in China. Chen and li (2021) applied causal reasoning and intelligent quantitative methods to assess the effectiveness of science and technology policies.

On the other hand, research primarily adopts content analysis or evaluates policy attributes from the perspective of policy tools. For example, Xu et al. (2022) classified science and technology policy texts into three categories: supply-driven, environmental, and demand-driven. Through coding and statistical analysis, they proposed comprehensive policy optimization suggestions based on intensity, completeness, and synergy. McWilliam et al. (2015) conducted content analysis of policy documents in Ontario, Canada, to determine whether policy tools existed to protect designated green infrastructure from the negative effects of residential encroachment. Lehtoranta et al. (2011) analyzed how policy tools impact the development of industrial symbiosis systems and environmental performance.

Although existing policy evaluation methods have made some progress in terms of empirical and text content analysis, there are still widespread issues such as inconsistent standards and strong subjectivity, making it difficult to comprehensively support the systematic evaluation of complex policy systems. In this context, the PMC index model, with its clear structure and rigorous logic, provides a more scientific and comparable policy evaluation method.

3.2. PMC index modelling

The Policy Modelling Consistency (PMC) Index Model, proposed by Estrada in 2011, is one of the most advantageous policy evaluation methods currently used both domestically and internationally. This model can determine the consistency level and strengths and weaknesses of any policy model. It allows for the quantification of the entire content of the policy text, evaluating it across multiple dimensions and indicators, providing a comprehensive understanding of the policy's strengths and weaknesses (Estrada, 2011). This approach significantly enhances the objectivity of the evaluation and enhances its accuracy. Since the introduction of the PMC index model in China, it has been widely and deeply applied in policy quantification research. Related research has shown an increasing richness of evaluation systems and an expanding field of application.

Research has also focused on model integration and indicator construction. Wang et al. (2019) innovatively combined the PMC model with autoencoder (AE) technology to construct a multi-level indicator system (9 first-level indicators and 38 second-level indicators) for the quantitative evaluation of 10 military-civilian integration policies, demonstrating the potential of the PMC model in structurally expressing complex policy relationships. In cross-national comparative studies, Zhu et al. (2024) constructed an evaluation system based on 74 Chinese and U.S. AI policy texts (19 first-level indicators and 42 second-level indicators), and selected six representative policies from each country to conduct a quantitative comparative analysis, which not only verified the cross-cultural applicability of the PMC model, but also demonstrated its strengths in evaluating policy competitiveness.

In terms of regional and industry policy analysis, the PMC model also demonstrates excellent adaptability and explanatory power.

Dai et al. (2021) focus on 16 policy texts on the theme of green development in China's Yangtze River Economic Belt, identifying the structure of the policy system and the distribution of its content through quantitative analysis, and providing empirical support for policy optimization in the region. Du et al. (2019) applied the PMC model to systematically analyze 120 innovation and technology policies within a large-scale text corpus, identifying structural issues in current policy design and suggesting directions for improvement. Hu et al. (2020) further integrated policy intelligence demand analysis to generate PMC surface maps for eight robotics industry policies, enhancing the visual representation of model results and providing intuitive support for structural assessments.

The PMC model continues to expand its applications into new fields. Liu et al. (2022) conducted a sample analysis of 10 textile industry policies to assess policy effectiveness, providing quantitative evidence for traditional manufacturing policy formulation. Notably, the PMC model is also widely used in international policy evaluation research. Yimsuk et al. (2024) systematically assessed the policy consistency and system structure of Thailand's COVID-19 control policies from 2020 to 2022, verifying the model's transferability in public health policy scenarios.

In summary, existing research indicates that the PMC index model not only possesses good versatility and flexibility, capable of adapting to the structural modeling needs of multiple types of policy texts, but has also gradually formed new trends in method integration, visualization enhancement, and cross-regional assessment in diverse applications. However, most studies still rely on policy texts to construct indicator systems and manually assign variables, and the model's automation level and semantic understanding capabilities need to be further improved.

3.3. Text classification

Text classification is the task of categorizing text into predefined groups by extracting and analyzing its features (Li et al., 2022). Looking back at the early history of machine learning and artificial intelligence, text classification technology was primarily applied in information retrieval systems. With technological advancements, it has found widespread application across various fields, including medicine, social sciences, healthcare, psychology, law, and engineering (Kowsarik et al., 2019). In natural language processing, the key tasks of text classification include sentiment analysis (Tai et al., 2015), topic tagging (Chen et al., 2019), question answering (Liu et al., 2015), news classification (Minaei et al., 2021), and dialogue act classification (Lee et al., 2016).

In the early stages of research on automated text classification, scholars mainly used machine learning methods such as Support Vector Machines (SVM), decision trees, Bayesian algorithms, and k-nearest neighbors (KNN) to improve classification accuracy (Zuo et al., 2024). For example, Zhang et al. (2011) used the Sogou corpus as the dataset, extracting data from 10 major categories, and performed automated classification using KNN and SVM classification algorithms. Kuo (2014) applied classifiers like Naive Bayes, SVM, and decision trees to propose a multilayer library literature classification model, achieving over 90% accuracy in the first-level classification according to the "Chinese Library Classification" system.

After 2016, with the rapid development of deep learning technologies such as neural networks and pre-trained models like BERT, automatic text classification entered the phase of deep learning algorithms. For instance, Luo et al. (2020) built a literature discipline classification model based on BERT and ERNIE, successfully classifying Chinese Social Sciences Citation Index literature into the Ministry of Education's first-level disciplines. Zhang et al. (2022) used the BERT model and constructed 26 automatic classifiers based on the logical framework of an automatic classification engine, enabling the automatic classification of scientific literature. Correa et al. (2020) aimed to develop a method to search in depth every web page on the Internet and formalize a machine learning classification model to improve the identification of data portals.

Owing to its powerful capabilities in feature selection and feature extraction (Hu et al., 2021), deep learning has been applied to the study of policy classification. Policy text classification can uncover associated knowledge, rules, and deeper semantic meanings from large datasets of text, aiding in policy prediction, optimization, and decision-making support. For example, Shen (2022) proposed a technology policy text classification method that combines the BERT model with keyword extraction techniques from the perspective of policy tools. Huo et al. (2024) trained an automatic policy tool classification model and mined 1,154 COVID-related policies in China, analyzing the co-evolution between policy goals and policy tools. Zhao et al. (2023) proposed a large-scale multi-label policy classification framework (Weak-PMLC) based on very weak supervision, in order to reduce the burden of annotating a large number of policies by human experts.

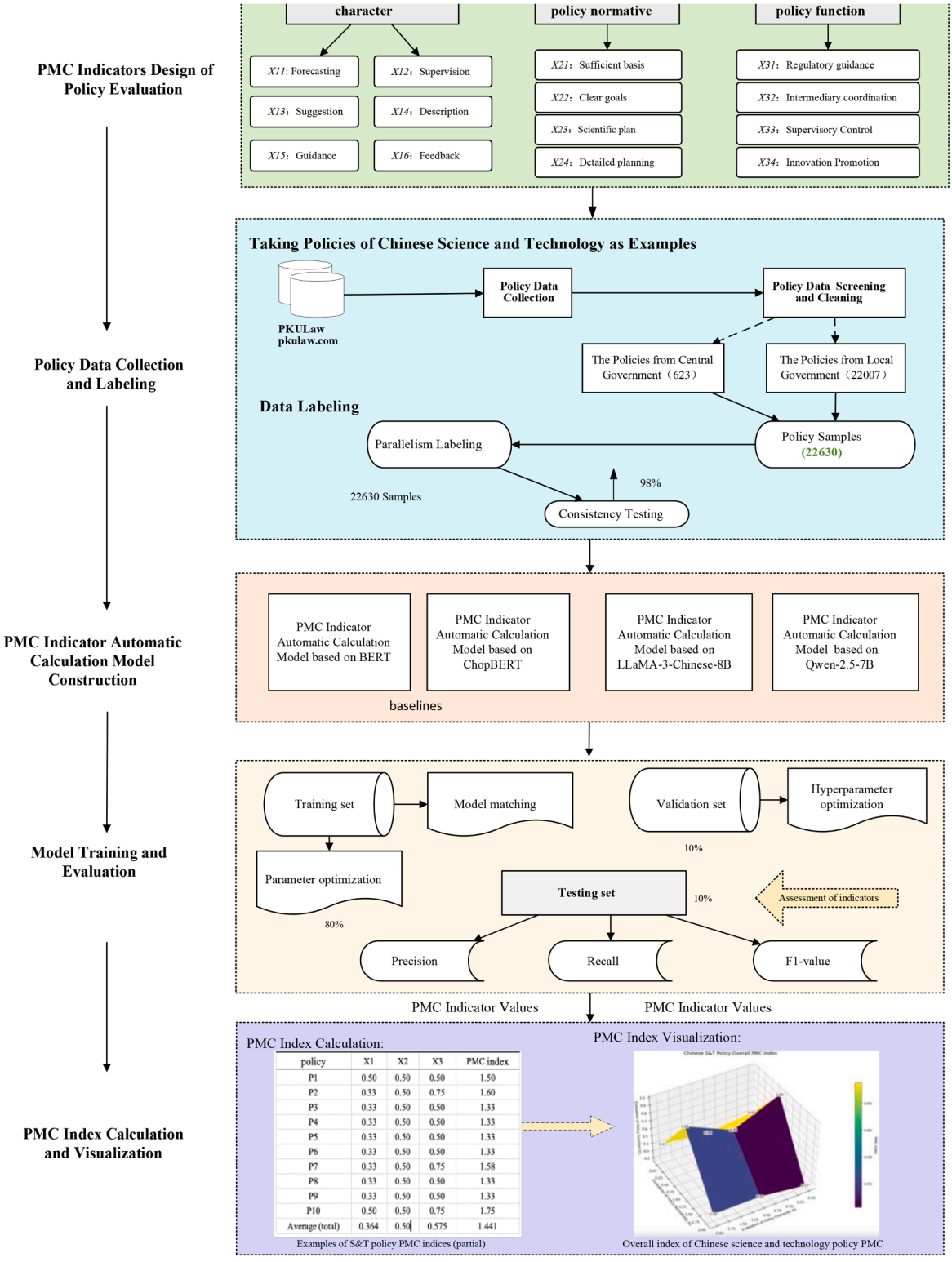
Large language modeling (LLM) also offers important opportunities and possibilities for quantifying political texts. Zhang et al. (2025) proposed a novel knowledge distillation framework (KOALA) based on the cooperation of large-scale language models, in response to the problems of complexity, subjectivity, data sparsity, and computational cost faced by policy style measures. Zhang et al. (2024) constructed Agent-Pro, an LLM-based agent with policy-level reflection and optimization that learns rich expertise from interactive experiences and gradually improves its behavioral policy. Yan et al. (2024) extracted representative scenarios of ChatGPT, explored the differences in users' perceptions of each scenario, and provided a policy-supporting model that offers insights into users' perceptions of ChatGPT in different scenarios.

3.4. Conclusion of the literature review

In summary, the existing literature exhibits several key limitations:

Firstly, PMC index models typically rely on manual variable assignment, which is labor-intensive. This high cost constrains most studies to small policy samples, thereby compromising the comprehensiveness of large-scale policy evaluation and threatening the accuracy and generalizability of the results.

Secondly, conventional PMC index model implementations often rely on non-exhaustive, "representative" policy sampling



policy	X1	X2	X3	PMC index
P1	0.50	0.50	0.50	1.50
P2	0.33	0.50	0.75	1.60
P3	0.33	0.50	0.50	1.33
P4	0.33	0.50	0.50	1.33
P5	0.33	0.50	0.50	1.33
P6	0.33	0.50	0.50	1.33
P7	0.33	0.50	0.75	1.58
P8	0.33	0.50	0.50	1.33
P9	0.33	0.50	0.50	1.33
P10	0.50	0.50	0.75	1.75
Average (total)	0.364	0.50	0.575	1.441

Fig. 1. The framework for automated evaluation of PMC index based on large language models.

workflows that are prone to selection bias. These methodological limitations undermine the model's consistency when scaled to analyze multi-domain, high-volume policy corpora. Addressing this requires constructing domain-generalized PMC evaluation datasets to support large-scale policy model pretraining and cross-sectoral validation.

Thirdly, the construction of the policy evaluation indicator system is the core of the application of the PMC index model. However, existing PMC evaluation systems often derive their primary and secondary indicators directly from the content of specific policy texts, lacking cross-domain generalizability (Wang et al., 2024).

Finally, although some text classification models such as BERT and its variants perform well in general text understanding tasks, they still have significant limitations when handling policy texts with complex structures and non-linear semantics. On one hand, the semantic structure of policy texts exhibits high levels of discreteness and dependency; On the other hand, traditional pre-trained models are constrained by their encoding mechanisms that primarily focus on local context, making it difficult to capture long-range semantic associations and implicit structural logic when modeling such texts.

To address these limitations, this study aims to develop the first automated PMC index evaluation framework leveraging large language models. We will construct a large-scale, manually annotated dataset of Chinese S&T policies to facilitate this research. Furthermore, we will redesign the PMC indicator system and its scoring rules to enhance domain generality, aiming to improve the efficiency, accuracy, and broader applicability of policy evaluation.

4. Methods

4.1. Research design

We constructed a framework for automated policy text evaluation based on several open-source large language models, as depicted in Fig. 1.

The first component involves the design of policy evaluation indicators. The PMC index model is a policy evaluation framework established grounded in the "Omnia Mobilis" hypothesis, which posits that all entities are in constant motion and interconnected (Estrada et al., 2008). This principle necessitates the inclusion of all relevant indicators while avoiding omissions in model construction. Guided by this theory, we developed a core PMC indicator system comprising three first-level indicators (Policy Character, Policy Normativity, and Policy Function) and 14 second-level indicators (e.g. forecasting, supervision, suggestion, sufficient basis, clear goals, regulatory guidance, intermediary coordination etc.).

The second phase involved policy data collection and labeling. We systematically collated science and technology policy texts issued by Chinese central and local governments between 1955 and 2022 from the PKULaw database. The raw data were then manually verified; non-normative documents were eliminated, resulting in a final corpus of 22,630 legislative policy texts for manual labeling.

The third step encompassed the construction of an automated PMC calculation model. We constructed four distinct PMC index automated calculation models: (1) BERT-PMC: a BERT-based PMC index calculation model. (2) ChpoBERT-PMC: a Chinese Policy BERT-based PMC index calculation model. (3) LLaMA-3-PMC: a PMC index calculation model based on LLaMA-3-Chinese-8B. (4) Qwen-2.5-PMC: a PMC index calculation model based on Qwen-2.5-7B. LLaMA-3-Chinese-8B and Qwen-2.5-7B both fall within the 10B parameter range, making them comparable. All models were trained and evaluated using multi-label data.

The fourth phase consisted of model training and evaluation. First, 80% of the data was used as the training set. Then, 10% was allocated as a validation set for hyperparameter optimization. Finally, the remaining 10% was held out as a test set to evaluate model performance. The primary evaluation metrics included accuracy, recall, and F1 score. This process followed standard data partitioning and phased optimization strategies to ensure the reliability of model training and the objectivity of evaluation.

The final stage involved PMC index calculation and visualization. We calculated the PMC index based on multi-input-output tables and ultimately generated a surface diagram to intuitively reveal the strengths, weaknesses, and overall characteristics of the policy.

4.2. PMC indicator design of policy evaluation

The indicator system is the core component of the PMC index model. Traditionally, constructing a PMC indicator system required researchers to extract high-frequency vocabulary from policy texts. These context-dependent keywords were then used to define evaluation indicators. For example, Zhao et al. (2024) established a PMC evaluation system for U.S. science and technology competition policies targeting China by analyzing word frequency and phrase co-occurrence. Their system comprised 9 first-level and 45 second-level variables, with three first-level variables (policy role receptor, policy tool, and policy role scope) being highly specific to that particular policy domain. Similarly, Dai et al. (2022) used the ROSTCM 6.0 tool to analyze high-frequency words in ecological protection compensation texts, forming a system with 10 first-level and 44 second-level variables. Three of these first-level variables (policy field, policy social benefits, and policy subjects) were directly derived from the content of those specific texts. This context-dependent indicator system construction method extracted from policy content, although it can accurately reflect the semantic features and functional structure of a specific policy text, is difficult to be directly transferred to the study of other policy domains, and the evaluation results are not commensurable.

To address these limitations, we fundamentally restructured the framework. Inspired by the conceptual approach of Ba et al. (2025), who proposed a multidimensional policy citation framework applicable across domains (three categories and 26 interpretable citation features), we redesigned a generalized PMC indicator framework suitable for cross-domain policy evaluation. The core innovation of this framework lies in its theoretical foundation. Instead of relying on text content, we return to the essential attributes of

policy itself, selecting three refined dimensions: evaluation of policy character X_1, evaluation of policy normativity X_2, and evaluation of policy function X_3.

- *Evaluation of policy character X_1 assesses whether the policy serves the roles of forecasting, supervising, advising, describing, guiding, or providing feedback (Estrada, 2011; Wang et al., 2024).*
- *Evaluation of policy normativity X_2 assesses whether the policy's basis is adequate, its goals are clearly defined, its plan is scientific, and its overall framework is reasonable (Zhang & Geng, 2015; Cai et al., 2021; Wang et al., 2024).*
- *Evaluation of policy Function X_3 assesses whether the policy has regulatory guidance, intermediary coordination, supervisory control, or innovation-driven functions. This is based on public policy theory, offering an understanding of the fundamental functions of public policy from a comprehensive perspective of guidance, coordination, control, and distribution, thereby deepening the understanding of the specific attributes of policy functions (Yan, 2002).*

The new framework is grounded in the underlying logic of policy science and focuses on attributes intrinsic to all policies. Specifically, X_1 evaluates the intrinsic rationality of policy design; X_2, the formal legitimacy of the text; and X_3, the practical effectiveness of implementation. This approach frees the evaluation from reliance on policy-specific keywords and provides a unified foundation for cross-sectoral policy assessment. Compared to the traditional approach of stacking 9-10 primary indicators, our framework's three dimensions form a rigorous and non-redundant logical sequence: policy design rationality (X_1) → textual formal normativity (X_2) → implementation functional effectiveness (X_3). This sequence corresponds to the key stages of the policy lifecycle: formulation, publication, and execution. This integrated structure forms a "policy quality transmission chain," providing a reusable core for policy optimization with broad applicability.

In addition, the second-level variables for X_1, X_2, and X_3 follow a binary scoring system [0,1]. Specifically, if the policy text contains relevant words or phrases that match the second-level indicators, the value for that second-level variable is set to 1; if the criteria are not met, the value is set to 0. The complete framework of PMC evaluation indicators is presented in Table 2.

4.3. Data collection and labeling

The policy data used in this study primarily is sourced from PKULaw[1], the China's most professional and advanced legal and regulatory search database. PKULaw includes over 2.99 million laws and regulations enacted in China since 1949. The specific steps for collecting policy data are as follows: Firstly, the document sources are set to "Central Regulations" and "Local Regulations," and the keyword "Science and Technology" is used to search for policies related to science and technology in a broad sense. Secondly, we extract and store the following metadata for each policy: title, full text, issue date, document number, and issuing agency. Finally, the policy documents are verified and screened for relevance. After collection, merging, filtering, and cleaning, a total of 22,630 regulations were obtained. Each policy's metadata and corresponding text were stored in the same dataset, and sorted by the order of issuance. Central policies are coded as S1, S2, S3… S623, and local policies are coded as R1, R2, R3… R22007. Some samples of the data are presented in Table 3.

During the data labeling process, we built an online platform to facilitate collaboration among all team members, enabling them to annotate the data independently. We performed five rounds of random sampling (300 policies per sample) to measure inter-annotator agreement. The Fleiss' Kappa score for each sample was above 0.98, indicating excellent consistency. This demonstrates the high quality and reliability of the annotated data in this study.

4.4. Construction of PMC indicator automated calculation model based on LLaMA-3-Chinese-8B

LLaMA-3 (Large Language Model Meta AI) is Meta AI's third-generation open-source large language model. It employs an enhanced Transformer decoder architecture and was pre-trained on 1.4 trillion tokens of multilingual data, demonstrating exceptional capabilities in semantic comprehension and text generation. Compared to its predecessors, LLaMA-3 achieves breakthroughs in parameter scale (up to 128 K context window), inference efficiency, and computational optimization through a Grouped Query Attention (GQA) mechanism. However, the base model lacks domain-specific adaptation for policy text analysis and is not inherently capable of parsing specialized policy evaluation metrics like the PMC index. Therefore, this study focuses on constructing an automated PMC evaluation model based on the LLaMA-3 framework. Due to GPU computational constraints, we selected LLaMA3-Chinese-8B, an open-source model with approximately 8 billion parameters (8B) that is specifically fine-tuned for Chinese language tasks. This selection ensured strong Chinese language processing capabilities within our computational budget. The specific processing process is shown in Fig. 2.

First, policy texts are encoded and tokenized. Policy documents were tokenized using LLaMA-3's native Byte-Pair Encoding (BPE) tokenizer to generate token ID sequences. To better handle domain-specific terminology in Chinese policy texts, we expanded the tokenizer's vocabulary with relevant terms. Input consistency is ensured by configuring maximum sequence lengths with adaptive padding (using the EOS token designated as pad_token) and strategic truncation through the Chinese-LLaMA-adapted AutoTokenizer implementation.

[1] PKULaw Legal Database (v6.0), Peking University Legal Artificial Intelligence Laboratory & Beida Yinghua Technology Co., Ltd., https://www.pkulaw.com/ (accessed January 18, 2025).

Table 2
The framework of policy evaluation indicators.

First-Level Indicators	Second-Level Indicators	Explanation of Second-Level Indicators	Evaluation Parameter	Source
X_1: Evaluation of Policy Character	X_{11}: Forecasting	Whether the policy has a forecasting tendency	X_{11} is normalized to [0,1]	Estrada et al., 2011; Wang et al., 2024
	X_{12}: Supervision	Whether the policy has a supervisory tendency	X_{12} is normalized to [0,1]	
	X_{13}: Suggestion	Whether the policy has a suggestive tendency	X_{13} is normalized to [0,1]	
	X_{14}: Description	Whether the policy has a descriptive tendency	X_{14} is normalized to [0,1]	
	X_{15}: Guidance	Whether the policy has a guiding tendency	X_{15} is normalized to [0,1]	
	X_{16}: Feedback	Whether the policy has a feedback tendency	X_{16} is normalized to [0,1]	
X_2: Evaluation of Policy Normativity	X_{21}: Sufficient Basis	Whether there is sufficient basis for policy formulation	X_{21} is normalized to [0,1]	Zhang & Geng, 2015; Cai et al., 2021; Wang et al., 2024
	X_{22}: Clear Goals	Whether the policy has clearly defined goals	X_{22} is normalized to [0,1]	
	X_{23}: Scientific Plan	Whether the policy has a scientifically designed plan	X_{23} is normalized to [0,1]	
	X_{24}: Detailed Planning	Whether the policy has detailed planning arrangements	X_{24} is normalized to [0,1]	
X_3: Evaluation of Policy Function	X_{31}: Regulatory Guidance	Whether the policy has regulatory guidance functions	X_{31} is normalized to [0,1]	Yan, 2002
	X_{32}: Intermediary Coordination	Whether the policy has intermediary coordination functions	X_{32} is normalized to [0,1]	
	X_{33}: Supervisory Control	Whether the policy has supervisory control functions	X_{33} is normalized to [0,1]	
	X_{34}: Innovation Promotion	Whether the policy has innovation promotion functions	X_{34} is normalized to [0,1]	

(Note: The evaluation parameters for each sub-indicator takes values in {0,1})

Table 3
Policy documents (For examples).

No.	Policy Title	Year
S72	Interim Regulations on the Science Prize of the Chinese Academy of Sciences	1955
S128	Implementation Guidelines for the Science and Technology Progress Award of the People's Republic of China (Trial)	1986
S243	The Ten-Year Plan for the Development of Science and Technology in the People's Republic of China and the " Eighth Five-Year Plan" Outline	1991
S378	National Science and Technology Award Regulations (2003 Revision)	2003
S512	Law of the People's Republic of China on the Promotion of the Transformation of Scientific and Technological Achievements (2015 Amendment)	2015
S623	Notice from the Ministry of Science and Technology on the Issuance of the " 14th Five-Year Plan for the Market of Technological Elements"	2022
R1712	Interim Administrative Measures for Scientific and Technological Research Achievements in Jiangsu Province	1980
R4273	Regulations on the Promotion of the Transformation of Scientific and Technological Achievements in Zhejiang Province	2004
R8389	Notice from the General Office of the Changchun Municipal People's Government on the Issuance of the " Long-Term Science and Technology Development Plan of Changchun (2006-2020)"	2006
R13421	Guidelines on the Development of Strategic Emerging Industries from the People's Government of Xuzhou	2010
R17460	Notice from the General Office of the Wuxi Municipal People's Government on the Issuance of Wuxi's " 13th Five-Year Plan" for Technological Innovation	2016
R22007	Notice on the Issuance of the " Implementation Opinion on Promoting the Development of Science and Technology Innovation Vouchers in the Yangtze River Delta" by the Shanghai Municipal Science and Technology Commission, Jiangsu, Zhejiang, and Anhui Provincial Science and Technology Departments	2022

The second is decoder feature extraction and context modeling. Tokenized sequences are processed through a multi-layer Transformer decoder, where autoregressive self-attention mechanisms iteratively extract hierarchical text features. Each decoder layer dynamically computes inter-token dependencies across policy documents, capturing long-range contextual patterns critical to regulatory intent interpretation. The resultant hidden states at the final layer serve as semantic representations for downstream tasks.

The third is multi-label classification with probabilistic mapping. A mean-pooling operation aggregates hidden states into document-level embeddings, which are projected via fully connected layers to match predefined PMC functional dimensions (*e.g., regulatory guidance, mediation coordination, supervisory control, innovation promotion*). Independent sigmoid activations per label dimension produce multi-hot predictions through threshold-based binarization. Numerical stability is maintained by directly optimizing raw logits with BCEWithLogitsLoss.

The fourth is parameter efficient fine-tuning and training optimization. We implement resource-conscious adaptation through a multi-pronged strategy: (1) A fully parameter-tuned mechanism is adopted, meaning that no layers of the model are frozen during

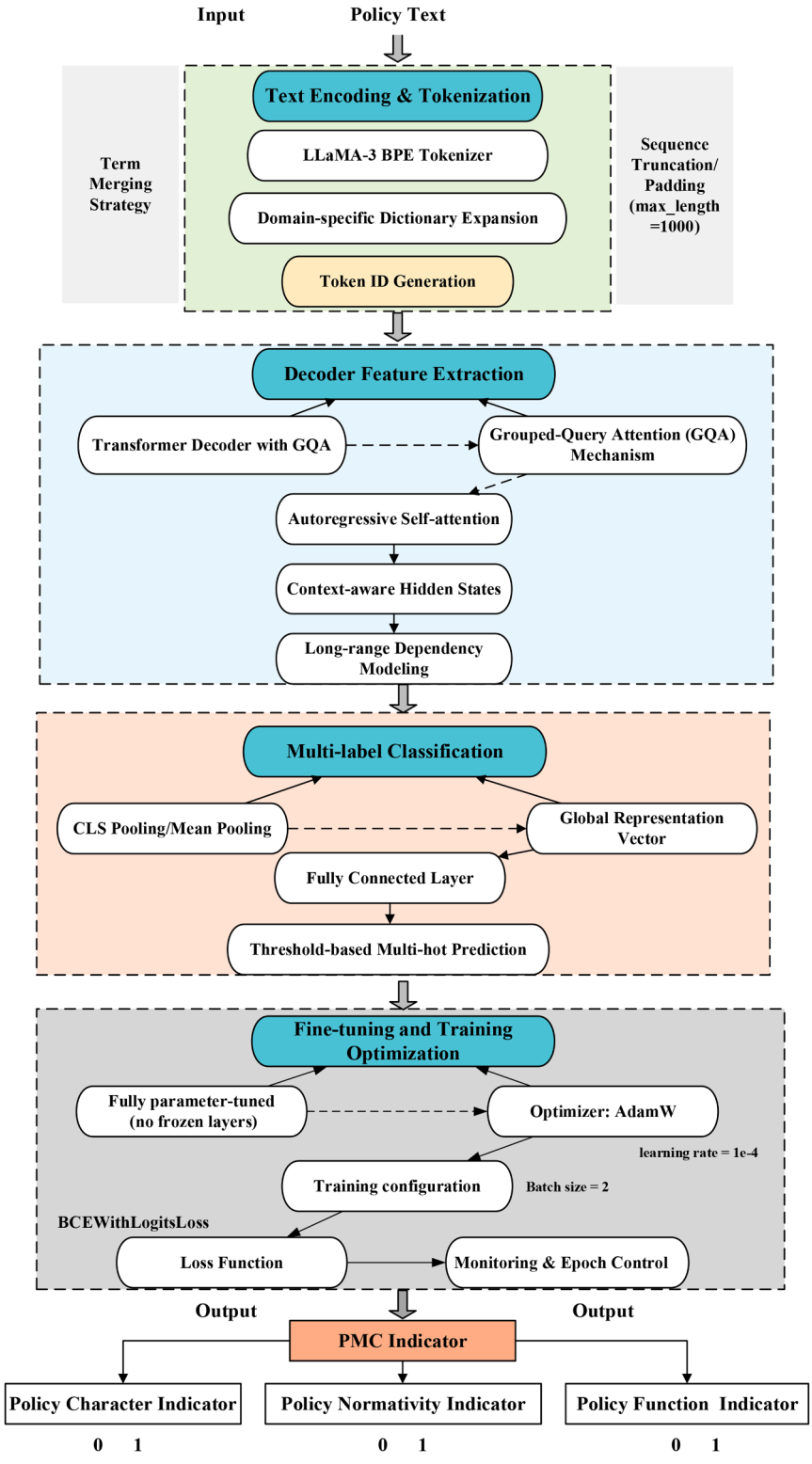


Fig. 2. Construction of PMC indicator automated calculation model based on LLaMA-3-Chinese-8B.

training, enabling joint optimization of low-level semantic representations and high-level task features to fully unleash the capabilities of the underlying large language model; (2) In terms of optimizers and precision strategies, the AdamW optimizer (with an initial learning rate of 1e-4) is combined with mixed-precision training (torch.bfloat16) to significantly reduce memory consumption while maintaining numerical stability; (3) For the training configuration, set batch_size=2 to adapt to the graphics card memory limit; (4) The loss function selected is BCEWithLogitsLoss (binary cross-entropy), suitable for multi-label classification tasks, effectively enhancing the learning capability for sparse labels; (5) Training monitoring mechanisms include dynamically evaluating the weighted F1 metric on the validation set and completing the entire fine-tuning process within three fixed training epochs without enabling early stopping, to balance training adequacy and computational cost. The PMC model codes constructed by LLaMA-3-Chinese-8B have been open-sourced on GitHub[2].

4.5. Construction of PMC indicator automated calculation model based on Qwen-2.5-7B

Qwen-2.5, developed by Alibaba Cloud, is a third-generation general-purpose large language model (LLM) that integrates a Mixture-of-Experts (MoE) architecture with a dynamic sparse attention mechanism. Pre-trained on 4.8 trillion tokens of multilingual high-quality corpora, it supports a 128K-token context window and multimodal input processing. Compared to its predecessors, Qwen-2.5 demonstrates significant advancements in parameter scale (up to 1.8 trillion activated parameters), inference efficiency, and cross-domain generalization capabilities. Notably, domain-adaptive pre-training (DAPT) enhances its ability to parse complex logical relationships inherent in policy texts. Benchmark evaluations confirm Qwen-2.5's state-of-the-art performance, establishing it as a leading NLP tool for analyzing structurally intricate long-form policy documents. Building on these strengths, this study proposes a PMC Indicator automated calculation model based on Qwen-2.5-7B, incorporating a policy-structured feature encoding module and task-oriented fine-tuning strategies to improve computational precision in multidimensional policy text evaluation, as detailed in Fig. 3.

First, policy texts are preprocessed and multi-tag encoding is performed. We extract domain expert annotations using Chinese comma delimiters (e.g., prediction, supervision, recommendation, description, guidance, feedback) are extracted via Chinese comma delimiters. A dictionary-mapping mechanism generates structured [text, label] pairs by converting categorical strings into binary multi-hot encoded vectors, ensuring precise alignment between textual content and functional labels.

The second is domain adaptive disambiguation and model initialization. We deploy Qwen-2.5's native BPE tokenizer with explicit EOS token padding for architectural compatibility. The pretrained backbone is initialized via AutoModelForSequenceClassification, configured with an output dimension matching label cardinality. To accommodate protracted policy documents, input sequences are truncated to 1,000 tokens, while BFloat16 mixed precision reduces GPU memory overhead by ~40% during inference.

The third is the multi-label training framework design. Use a random division strategy (80% training, 10% validation, 10% test) supports robust generalization. Batched processing pipelines (batch_size=2) are engineered via TensorDataset/DataLoader to respect GPU memory constraints. The training loop optimizes binary cross-entropy loss, where forward passes yield raw logits subsequently thresholded at 0.5 post-sigmoid for multi-label prediction. Performance is monitored through dynamically weighted precision, recall, and F1 metrics.

Finally, the model undergoes validation and performance optimization. Consistent dynamic thresholding is applied during validation/testing phases. Weighted macro-averaged metrics (precision/recall/F1) are computed. Final evaluation outputs test-set loss values and classification benchmarks, establishing reproducible protocols for end-to-end multi-label policy analytics frameworks. The PMC model codes constructed by Qwen-2.5-7B have been upload on GitHub[3].

5. Experiments and results

5.1. Evaluation metrics

Automated policy text evaluation, as formulated in this study, is a multi-label classification problem. Therefore, this study employs three commonly used evaluation metrics for classification tasks: precision, recall, and F1-score, to assess the performance of the proposed automated PMC calculation models. Precision measures the proportion of correctly identified positive instances among all instances predicted as positive. Recall measures the proportion of correctly identified positive instances among all actual positive instances. The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. The formulas for calculating each evaluation metric are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

[2] https://github.com/ChaoguangHuo/Policy_Automatic_Evaluation.
[3] https://github.com/ChaoguangHuo/Policy_Automatic_Evaluation.

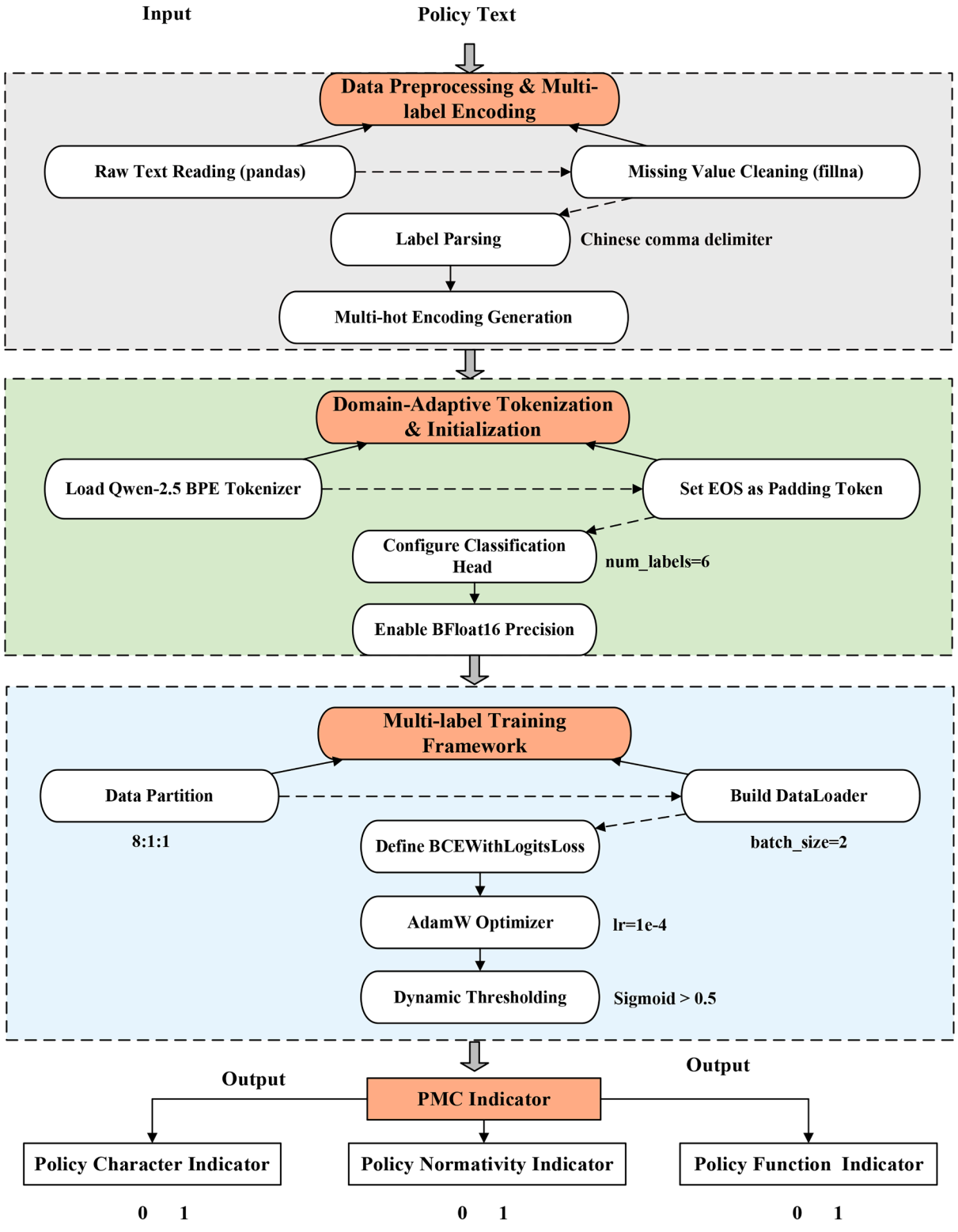


Fig. 3. Construction of PMC indicator automated calculation model based on Qwen-2.5-7B.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

5.2. Baselines and experimental setup

At the beginning of the experiment, we randomly sampled 100 samples from the dataset and evaluated them using the zero-shot capabilities of LLaMA-3-Chinese-8B and Qwen-2.5-7B. The results are shown in Table 4. Without any task-specific training, the performance of both LLaMA-3-Chinese-8B and Qwen-2.5-7B was low, with precision below 44% and recall under 34%. These results are insufficient for practical application, demonstrating that pre-trained LLMs without domain adaptation are inadequate for the specialized task of policy text evaluation. This underscores the necessity of constructing high-quality, domain-specific datasets and further fine-tuning the models to enhance their accuracy, which is the focus of this study.

For supervised learning baselines, and given the lack of prior work on automated PMC evaluation, we selected BERT and ChpoBERT (Chinese Policy BERT) as benchmark models. BERT employs bidirectional contextual modeling to holistically capture semantic relationships between words, demonstrating superior performance over traditional unidirectional models across NLP tasks. ChpoBERT is a Chinese policy-oriented pre-trained model (Shen et al., 2023) based on the BERT architecture. Its training on policy data allows it to better handle Chinese linguistic nuances such as polysemy and syntactic specificity. It has been shown to outperform generic BERT and LLaMA-2 on Chinese policy text tasks (Huo et al., 2024).

All experiments were conducted on a deep learning workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM), an Intel i9-14900 K processor (3.20 GHz), 128 GB of system RAM, and a 2 TB SSD. The 22,630 labeled samples were randomly divided into training, validation, and test sets in an 8:1:1 ratio. All models were trained with the number of epochs fixed at 3 and a batch size of 2 to accommodate GPU memory constraints. The models were trained sequentially.

Under this configuration, the complete training process required approximately 1,440 GPU hours. All training was conducted stably in a single-GPU environment. The detailed training time per model is presented in Table 5. The final fine-tuned models showed significant performance improvements, achieving both efficient inference speed and high multi-label classification accuracy. The final fine-tuned models achieved high multi-label classification accuracy while maintaining efficient inference speeds, demonstrating their strong potential for deployment in resource-constrained environments and practical policy analysis applications.

5.3. Results and discussion

To address RQ2 on integrating the PMC framework into LLMs for automated indicator calculation, our experimental results demonstrate that task-specific fine-tuning of modern LLMs is a highly effective methodology. The performance variations across models, as detailed below, highlight the critical importance of model architecture and training strategy in this integration process.

As shown in Tables 6, 7, and 8, different models exhibit varying performance across the three evaluation subtasks. Among them, Qwen-2.5-7B achieved the highest F1-score (80.41%) on the evaluation of policy character (X_1) subtask. This result suggests that the model, after task-specific fine-tuning, effectively balances precision and recall for the various labels within this subtask.

On the evaluations of policy normativity (X_2) and policy function (X_3) subtasks, LLaMA-3-Chinese-8B achieved F1-scores of 75.07% and 74.11%, respectively, demonstrating its advantage in understanding policy text structure and semantics. This advantage is likely due to the model's stronger capability in modeling long-range dependencies and cross-passage context, allowing it to identify normative clauses and functional expressions more accurately. These capabilities—localized semantic understanding, complex structure recognition, and multi-label discrimination—make it particularly suitable for analyzing the highly structured nature of policy texts.

In contrast, BERT and ChpoBERT performed poorly across all three subtasks, particularly on the multi-label classification of lengthy policy texts. This performance gap demonstrates that traditional PLMs like BERT struggle to model the structural complexity and semantic distribution of policy texts. Specifically, policy documents contain multiple parallel or overlapping functional modules (e.g., objectives, means, scope). These modules are scattered across different paragraphs or chapters, creating non-linear semantic dependencies. BERT-style models primarily capture local context, making it difficult for them to integrate long-distance dependencies and identify implicit structures, which ultimately hinders their performance on such tasks.

The fine-tuned Qwen-2.5-7B and LLaMA-3-Chinese-8B models demonstrated superior contextual awareness and domain adaptation, enabling them to more accurately distinguish heterogeneous policy labels, thereby enhancing overall evaluation performance. This performance gap not only is reflected in higher accuracy metrics but also empirically validates the advantages of fine-tuned large language models in structural understanding and semantic coverage for policy text evaluation.

The performance disparity between Qwen-2.5-7B and Llama-3-Chinese-8B on the three policy evaluation tasks stems from a fundamental alignment between each model's architectural strengths and the specific cognitive demands of the tasks. This alignment is rooted in their underlying architectures. LLaMA-3's design, featuring innovations like Grouped-Query Attention, is optimized for

Table 4
Performance of LLaMA-3-Chinese-8B and Qwen-2.5-7B with prompt method.

Model	Precision	Recall	F1-score
LLaMA-3-Chinese-8B	38.76%	33.93%	29.73%
Qwen-2.5-7B	43.39%	33.75%	34.78%

Table 5
Model training time.

Model	Actual training time	GPU hours
ChpoBERT	2 days	48 hours
BERT	2 days	48 hours
LLaMA-3-Chinese-8B	30 days	720 hours
Qwen-2.5-7B	27 days	648 hours

Table 6
Evaluation of policy character. Bold indicates best performance.

Model	Precision	Recall	F1-score
ChpoBERT	45.19%	68.25%	54.07%
BERT	44.90%	67.72%	53.65%
LLaMA-3-Chinese-8B	72.70%	76.30%	69.74%
Qwen-2.5-7B	**83.38%**	**81.52%**	**80.41%**

Table 7
Evaluation of policy normativity scores on various models. Bold indicates best performance.

Model	Precision	Recall	F1-score
ChpoBERT	32.30%	44.73%	37.51%
BERT	32.59%	44.77%	37.72%
LLaMA-3-Chinese-8B	**77.59%**	**73.55%**	**75.07%**
Qwen-2.5-7B	64.76%	82.23%	71.66%

Table 8
Evaluation of policy function scores on various models. Bold indicates best performance.

Model	Precision	Recall	F1-score
ChpoBERT	42.87%	69.11%	52.14%
BERT	41.13%	68.33%	51.35%
LLaMA-3-Chinese-8B	**80.18%**	**70.76%**	**74.11%**
Qwen-2.5-7B	73.32%	54.95%	59.89%

complex reasoning over long contexts. Qwen-2.5-7B's efficiency, potentially aided by its Mixture-of-Experts (MoE) design, makes it adept at pattern recognition and data-centric tasks. The architecture diagrams (Figs. 2 & 3) reflect these different emphases. Consequently, the evaluation of policy character (X1), which relies more on surface-level pattern recognition, aligns well with Qwen-2.5-7B's efficient design and strong baseline performance. In contrast, the evaluations of policy normativity (X2) and function (X3) require deeper abstract reasoning and understanding of logical coherence. LLaMA-3-Chinese-8B's architectural choices, such as its attention mechanisms, appear to provide a superior capacity for these more complex tasks, enabling it to map text to high-level theoretical functions more effectively.

Table 9
Multi-input-output table of all policies.

First-Level Indicators	Second-Level Indicators	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	...	P_{22630}
X_1	X_{11}	0	0	0	0	0	0	0	0	...	0
	X_{12}	0	0	0	0	0	0	0	0	...	1
	X_{13}	1	0	1	0	0	0	0	0	...	0
	X_{14}	1	1	1	1	1	1	1	1	...	1
	X_{15}	1	1	0	1	1	1	1	1	...	1
	X_{16}	0	0	0	0	0	0	0	0	...	0
X_2	X_{21}	1	1	1	1	1	1	1	1	...	1
	X_{22}	0	0	0	0	0	0	0	0	...	0
	X_{23}	1	1	1	1	1	1	1	1	...	1
	X_{24}	0	0	0	0	0	0	0	0	...	0
X_3	X_{31}	1	1	1	1	1	1	1	1	...	1
	X_{32}	0	0	0	0	0	0	0	0	...	0
	X_{33}	0	1	0	0	0	0	1	0	...	1
	X_{34}	1	1	1	1	1	1	1	1	...	1

(Notes: P1 represents the 1st policy, P2 represents the 2nd policy… P22630 represents the 22630th policy.)

5.4. Calculation and visualization of PMC index

Leveraging the above models, we obtained binary scores (0 or 1) for all secondary indicators of each policy. These scores were used to construct a multi-input-output table for all policies (Table 9).

Using this table and the formula below, we calculated the first-level indicator values (X_1, X_2, X_3) for each policy by averaging their constituent secondary indicators. For example, for P1 in Table 9, X_1 equal to (0+0+1+1+1+0)/6, namely 0.5, X_2 is 0.5, and X_3 is 0.5. Then we calculated the PMC of P1 is 1.5 (0.5+0.5+0.5). Following these steps, we calculated the PMC of all Chinese science and technology policies, as shown in Table 10. A higher PMC indicates a better rating and a higher consistency of the policy's secondary indicators.

$$X \sim N[0,1] \tag{1a}$$

$$X = \backslash\{X\_R\text{: } [0,1]\ \backslash\ \} \tag{2a}$$

$$X_i = \sum_{j=1}^{n} \frac{x_{ij}}{T(x_{ij})} \tag{3a}$$

$$PMC - Index = \sum_{i=1}^{y} \left[X_i \left(\sum_{j=1}^{n} \frac{x_{ij}}{T(x_{ij})} \right) \right] \tag{4}$$

(Notes: i=1, 2, 3, …, n, where i represents the first-level indicators, and j represents the second-level indicators.)

To clearly visualize the strengths and weaknesses of the policies, we employed a 3D surface plot (Hu et al., 2020) to represent the three primary indicators (X_1, X_2, X_3) for each policy. Fig. 4 presents the 3D surface plots for four representative policies: P1, P2, P3, and P22630. The three axes represent evaluations of policy character (X_1), policy normativity (X_2), and policy function (X_3), respectively. The shape of each surface allows for direct visual comparison of policy performance across these three dimensions.

As shown in Fig. 4, the four selected policies exhibit notable differences across the dimensions of evaluation of policy character X_1, evaluation of policy normativity X_2, and evaluation of policy function X_3, resulting in diverse surface patterns. For example, Policy 1, *Several Opinions on Further Promoting the Development of University Science Parks in Beijing*, records equal scores of 0.5 on X_1, X_2, and X_3. Its surface approximates an equilateral structure, indicating a relatively balanced performance across all three dimensions but lacking distinctive strengths in any single area. Policy 2, *Implementation Plan for Accelerating the Development of Science and Technology Services in Alxa League*, displays a clearly tilted surface, while it performs strongly on policy function (X_3), its score for policy character (X_1) is only 0.33, suggesting ambiguity in policy positioning that may weaken its implementation effectiveness. Policy 3, *Several Opinions on Promoting the Development of Science and Technology Intermediary Institutions in Beijing*, presents a trapezoidal distribution. With stable scores of 0.5 on both X_2 and X_3 but a comparatively lower X_1, this policy achieves expected outcomes in terms of normativity and function, yet its insufficient strength in policy character may undermine implementation. In contrast, Policy 22630, *Interim Measures for the Evaluation of Industrial Technology Innovation Strategic Alliances in Anhui Province*, produces a segmented surface pattern. Its X_3 score approaches the optimum (0.75), reflecting a systematic design in functional dimensions such as intermediary coordination and innovation promotion, while its X_1 and X_2 scores remain at relatively balanced levels.

It should be noted that the flat or segmented features observed in individual policy surfaces are not anomalies, but rather the result of discrete scoring across different indicators. When the scores for the three indicators are highly consistent, the surface tends to appear flat; when significant disparities exist across dimensions, the surface may exhibit a "two-plane" segmented structure. These surface patterns essentially reflect differences in the construction of fundamental policy attributes: some policies pursue overall balance, others emphasize a specific functional dimension, while still others demonstrate stronger normativity and functionality but relatively weaker policy character. Such differentiated patterns provide important insights into policy design orientations and their potential implications for implementation effectiveness.

We conducted a visual analysis of the overall distribution of 22,630 policy documents. The nine values presented in Fig. 5

Table 10
The PMC index of all science and technology policies.

policy	X_1	X_2	X_3	PMC Index
P1	0.50	0.50	0.50	1.50
P2	0.33	0.50	0.75	1.60
P3	0.33	0.50	0.50	1.33
P4	0.33	0.50	0.50	1.33
P5	0.33	0.50	0.50	1.33
P6	0.33	0.50	0.50	1.33
P7	0.33	0.50	0.75	1.58
P8	0.33	0.50	0.50	1.33
…	…	…	…	…
P22630	0.50	0.50	0.75	1.75
Average (total)	0.43	0.36	0.41	1.20
Variance (total)	0.0046	0	0.0131	0.0217

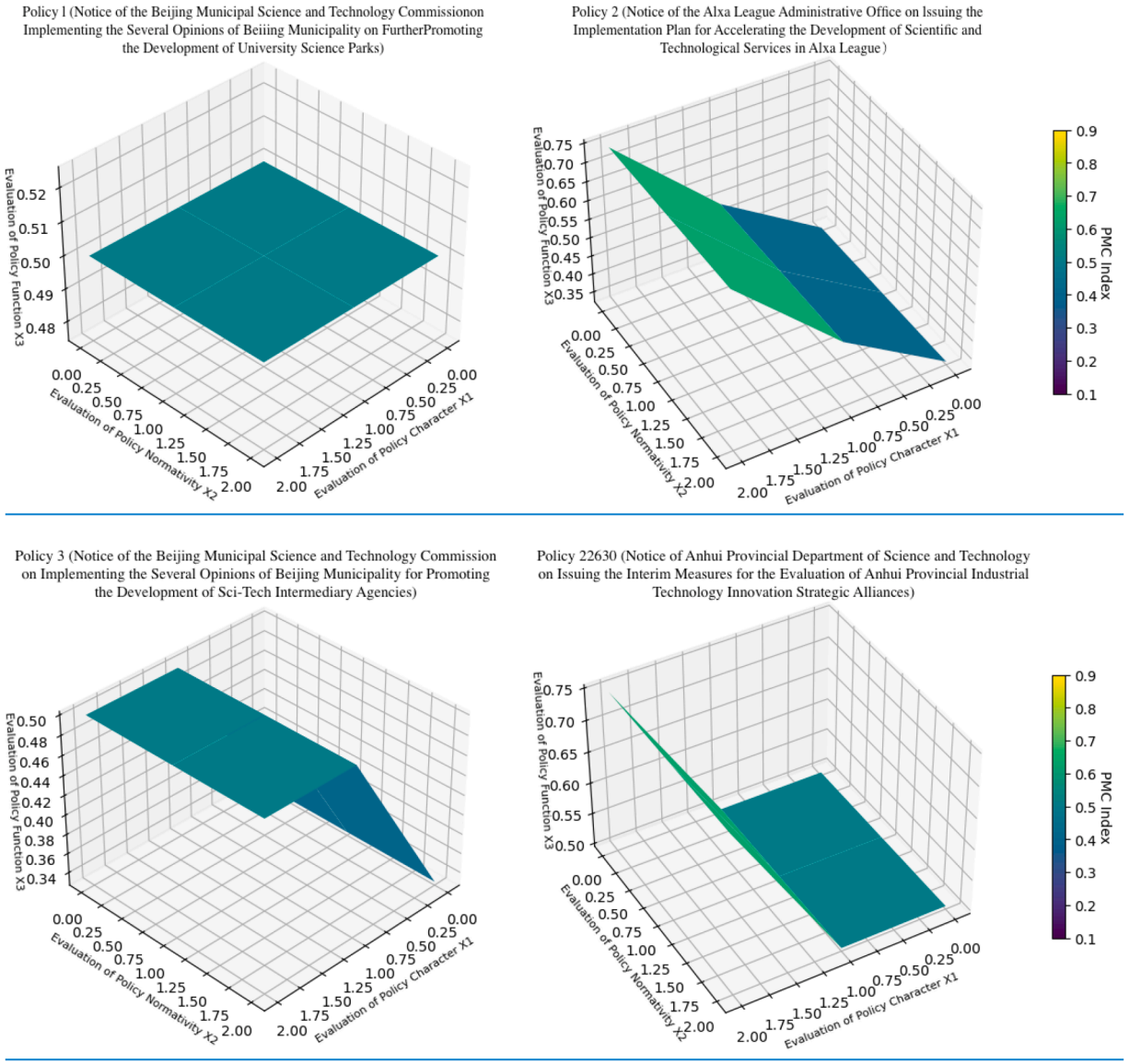


Fig. 4. Examples of PMC surface diagrams for Chinese science and technology policies.

correspond to the maximum, minimum, and mean for the evaluation of policy character (X_1), policy normativity (X_2), and policy function (X_3), clearly illustrating the overall distribution patterns across all dimensions. Specifically, the mean values of X_1, X_2, and X_3 were 0.43, 0.36, and 0.41, respectively, indicating some variation in overall levels, with the evaluation of policy normativity X_2 scoring the lowest. The maximum values of the three indicators reached 1.00, 0.75, and 1.00, approaching the theoretical optimum, reflecting that a small subset of policies—mainly central regulations and departmental rules—demonstrate high levels of design rationality, textual standardization, and functional effectiveness. In contrast, the minimum values were only 0.17, 0.25, and 0.25, indicating that some policies—typically operational or task-oriented documents with simple structures and limited coverage—perform poorly in fundamental attributes.

By examining these nine key values (maximum, minimum, and mean), the fluctuation characteristics and overall distribution patterns of each dimension can be clearly captured. The analysis indicates that, although a few high-quality policies approach the ideal level, the vast majority are concentrated in the mid-to-low range, particularly in terms of normativity evaluation (X_2). This range-based distribution not only reveals a significant structural imbalance in policy quality but also highlights the opportunities for improving policy coordination, systematization, and implementation consistency, providing directional guidance for the optimization of future policy texts and institutional design.

This section demonstrates the practical application and validation of the generalized PMC indicator framework proposed in Section 4.2 to address RQ1. The successful calculation and analysis of over 22,630 policies serve as a robust proof-of-concept, illustrating the

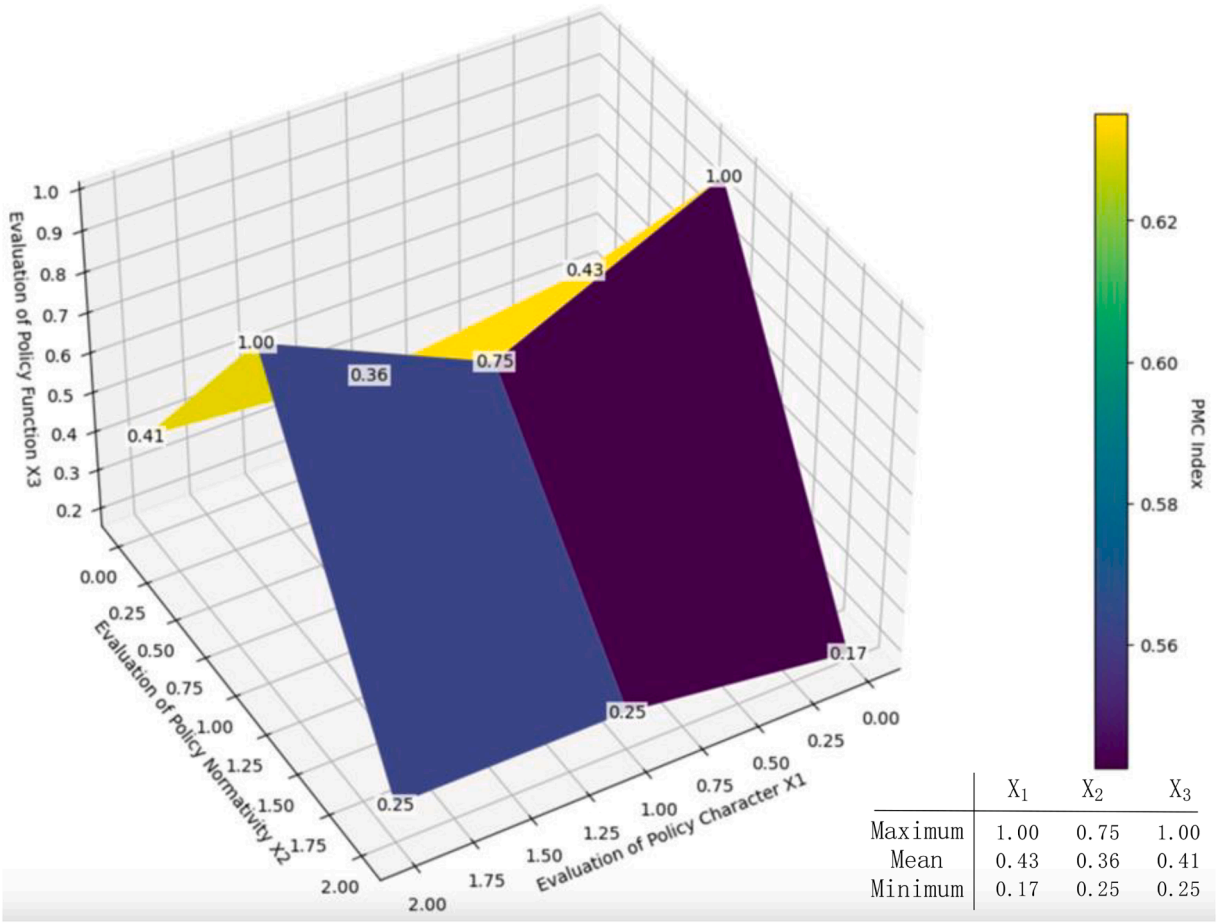


Fig. 5. The overall distribution of the PMC index for Chinese science and technology policies.

framework's core strength: its domain-agnostic nature. Unlike previous context-dependent approaches, our framework operates on a fixed set of universally applicable first-level (X_1, X_2, X_3) and second-level indicators, which are not extracted from the semantic content of specific policies but are derived from the fundamental attributes of policy itself. This design is the key to its cross-domain applicability. The findings on Chinese S&T policies are a specific instance of its application. The framework itself is designed for general use.

5.5. Application of the automated policy text evaluation framework

To validate the cross-domain applicability of the automated policy text evaluation framework, we directly applied the trained model to biosafety policies and data governance policies. Unlike previous approaches, which extracted context-dependent indicators

Table 11
The PMC index of Chinese biosafety policies.

policy	X_1	X_2	X_3	PMC Index
P1	0.66	0.75	0.75	2.16
P2	0.83	1.00	0.75	2.58
P3	0.66	0.75	0.75	2.16
P4	0.50	0.50	0.50	1.50
P5	0.50	0.75	0.75	1.33
P6	0.66	1.00	0.75	2.41
P7	0.33	0.50	0.50	1.33
P8	0.66	0.75	0.75	2.16
...	...	...	...	...
P2018	0.33	0.75	0.50	1.58
Average (total)	0.56	0.80	0.69	2.05
Variance (total)	0.0241	0.0303	0.0144	0.2230

from their policy content and scored manually, we performed automated evaluations for three types of indicators: policy character (X_1), policy normativity (X_2), and policy function (X_3). Based on the automated policy text evaluation framework, the PMC indexes were calculated.

The data for biosafety policies and data governance policies were sourced from PKULaw. For biosafety policy, we conducted a systematic search using keywords such as "biosafety" "biomanufacturing" "biomedical" and "human genetic resources". Each document was manually reviewed and filtered to eliminate those unrelated to biosafety. The data collection spanned from 2000 to 2024. After cleaning, a final set of 2,018 biosafety policy documents was obtained. These documents were fed into our trained model to calculate the PMC indexes for all biosafety policies, as shown in Table 11. Fig. 6 illustrates the evaluation results for a subset of these policies. For data governance policy, we used the dataset published by Huo and Huo (2023) in ASIS&T, which includes 1,097 policy documents. The evaluation results for a subset of data governance policies are shown in Fig. 7.

Experimental results demonstrate that the proposed automated policy text evaluation framework also exhibits strong evaluation efficacy in the domains of biosafety and data governance policy, further validating its cross-domain applicability. This robustness stems from the fact that its three evaluation indicators—Policy Character (X_1), Policy Normativity (X_2), and Policy Function (X_3)—are constructed based on fundamental attributes inherent to policy texts rather than domain-specific semantic knowledge. By leveraging large language models to accurately capture these domain-agnostic meta-attributes, our model generates consistent and reliable PMC index scores. Consequently, despite significant thematic differences across domains, shared fundamental policy attributes enable the model to perform effectively, solidifying the methodological validity of its cross-domain application.

6. Conclusion

The PMC index model is a widely recognized and scientifically sound method for policy text evaluation, demonstrating effectiveness in both domestic and international contexts. However, the existing PMC index model relies on manual variable assignment, which entails high labor costs and limits the assessment to a small number of policy samples. This approach is inadequate for evaluating large-scale policy corpora, as it cannot guarantee the comprehensiveness of policy coverage or the consistent application of the PMC index model across different policies. To address these limitations, this study innovatively develops an automated PMC index evaluation framework based on large language models (LLMs) to enhance efficiency and accuracy, providing an effective tool for large-scale, cross-domain policy evaluation.

6.1. Theoretical implications

This study achieves a critical breakthrough in computational policy science through the deep integration of large language models (LLMs) with the Policy Modeling Consistency (PMC) index framework, proposing an automated policy evaluation paradigm with significant methodological innovation and practical value. The concrete theoretical implications include:

Firstly, building on the foundational work of Estrada (2011), who established the PMC-Index as an instrument to evaluate policy modeling, our study extends this methodology by enabling its intelligent, automated application. This study intelligently enhances the PMC index model by constructing the first automated evaluation framework for policy text analysis through its deep integration with a large language model. This breakthrough overcomes the longstanding limitation of small-scale evaluation and provides a new paradigm for applying the PMC index model to large-scale policy texts.

Secondly, we have constructed the first specialized dataset for PMC index modeling, comprising 22,630 multi-dimensionally annotated samples from Chinese science and technology policy texts[4]. Based on this dataset, we conducted multi-model comparison experiments to evaluate mainstream LLMs (including LLaMA-3-Chinese-8B and Qwen-2.5-7B) on automated PMC evaluation tasks. This effort establishes benchmark F1-score results, providing reproducible evaluation standards and a comparison framework for subsequent research.

Thirdly, building on the work of previous scholars (Zhang et al., 2015; Cai et al., 2021; Yan, 2002; Wang et al., 2024), we have redesigned a generic PMC indicator framework that is independent of policy-specific content. The automatic calculation of PMC indices was achieved by mapping policy features onto a multidimensional vector space, finally enabling efficient quantitative policy evaluation.

6.2. Practical implication

Our findings have the following three practical implications:

Firstly, an intelligent policy text evaluation framework driven by large language models has been innovatively proposed. Using the Qwen-2.5-7B model, an F1-score of 80.41% was achieved in the policy character evaluation task, enabling the automated extraction and quantitative analysis of multi-dimensional policy features. This fundamentally breaks through the bottlenecks of low efficiency in manual annotation and the limited cross-domain scalability of the traditional PMC index model.

Secondly, by designing for compatibility with open-source modeling architectures (e.g., Qwen-2.5-7B and LLaMA-3-Chinese-8B), we have developed the first PMC automation framework that ensures methodological transferability in cross-domain policy

[4] https://github.com/ChaoguangHuo/Policy_Automatic_Evaluation.

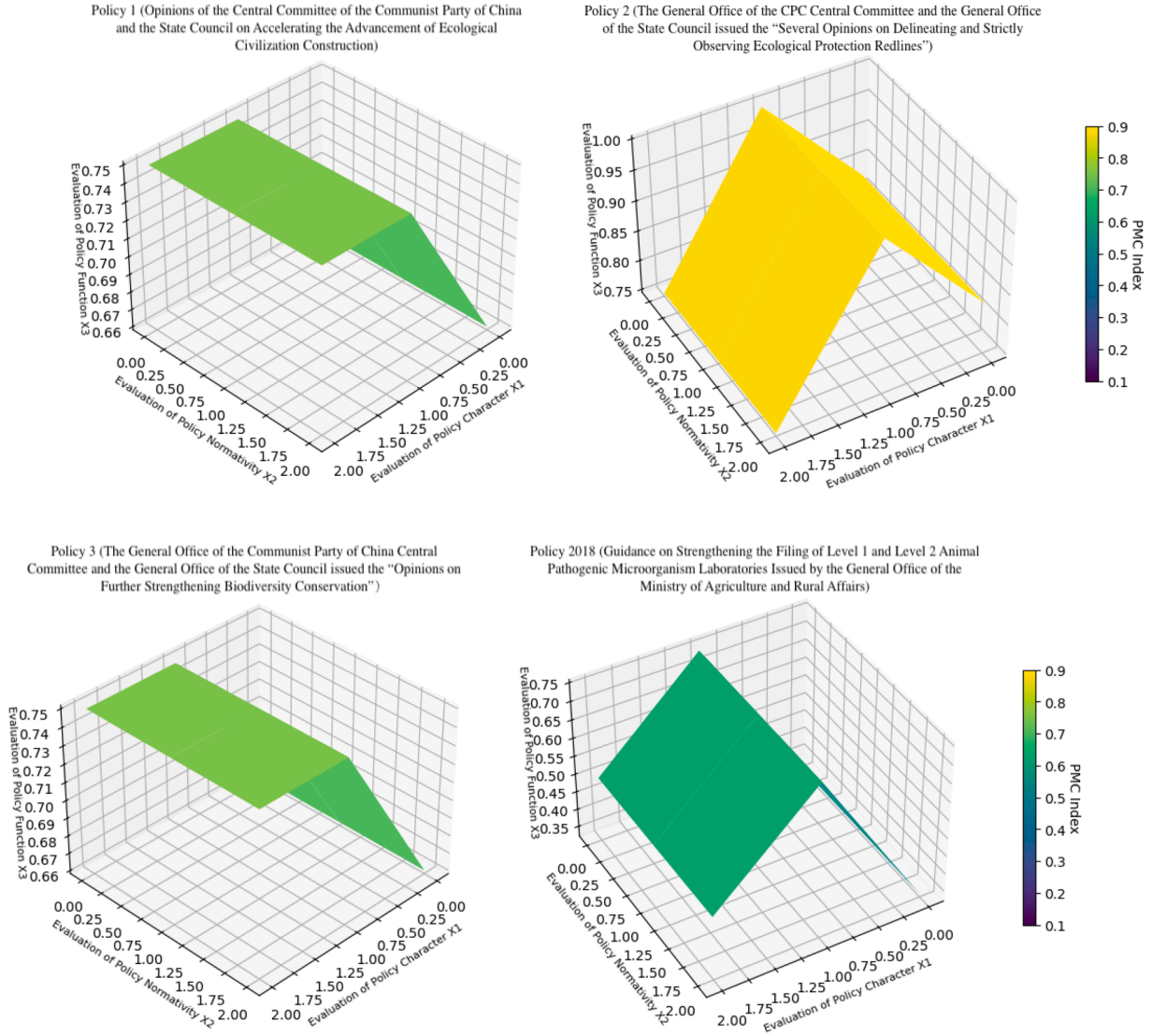


Fig. 6. Examples of PMC surface diagrams for Chinese biosafety policies.

evaluation. The subsequent release of our benchmark Chinese policy dataset and automated scoring rules will provide a repeatable and verifiable platform for future research.

Thirdly, we collected 22,630 Chinese S&T policies (1955-2022), automatically evaluated them using the PMC index, and transformed the results into visual indicators. This allows readers to intuitively identify synergies and flaws within policies. Our exploratory analysis reveals the significant potential of automated policy text evaluation methods for future research.

In summary, this study represents a landmark transition in policy evaluation from a "human-driven" to an "intelligence-augmented" paradigm. It also provides a scalable technical framework for emerging fields like policy text mining and multi-modal governance analysis, holding long-term significance for advancing the computational and empirical turn in policy science.

6.3. Limitations

The first limitation stems from computational constraints. Due to limited GPU resources, this study was limited to using models such as Qwen-2.5-7B and LLaMA-3-8B as base architectures. Future work with sufficient resources could explore larger-scale models (e.g., Qwen-2.5-72B or larger variants of LLaMA-3), potentially yielding improved performance. Furthermore, both Qwen-2.5 and LLaMA-3-Chinese are general-purpose LLMs, not models specialized for the policy domain. Consequently, their capability in processing policy-specific data may be constrained. Future work should involve continued pre-training or fine-tuning on massive policy-specific corpora to develop domain-specialized models for superior policy text comprehension.

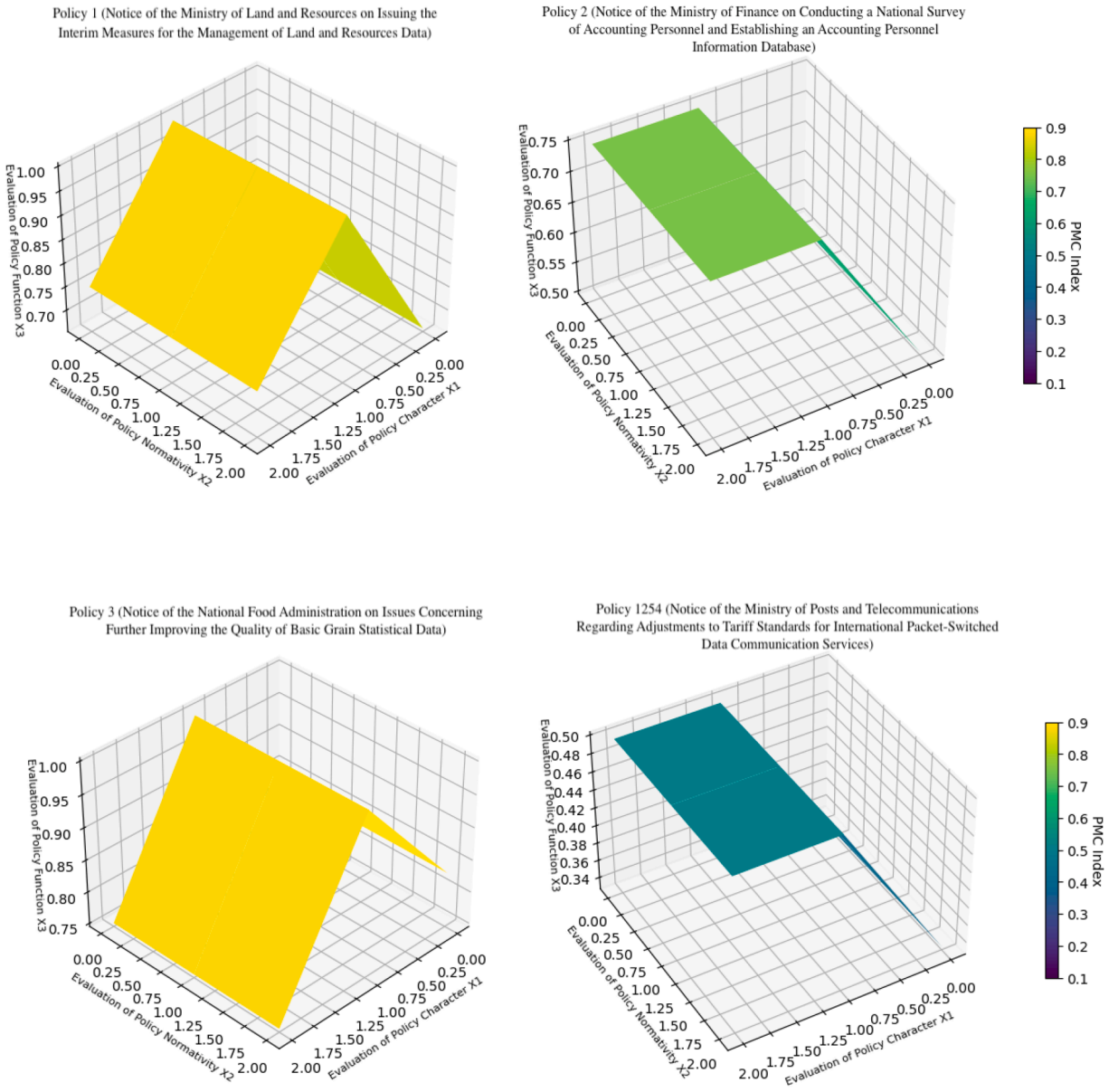


Fig. 7. Examples of PMC surface diagrams for Chinese data governance policies.

Author statement

This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. All study participants provided informed consent, and the study design was approved by the appropriate ethics. We have read and understood your journal's policies, and we believe that neither the manuscript nor the study violates any of these. There are no conflicts of interest to declare.

CRediT authorship contribution statement

Xiaobin Lu: Writing – original draft, Project administration, Methodology, Conceptualization. **Zinan Yang:** Writing – original draft. **Chaoguang Huo:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Education Ministry's Key Research Base for Humanities (Grant No. 22JJD870001) and National Natural Science Foundation of China (Grant No. 72374202).

Data availability

Data will be made available on request.

References

Ba, Z., Liu, L., & Xia, Y. (2025). Multidimensional policy citation features: Insights into policymakers' policy adoption decision-making. *Government Information Quarterly, 42*(1), Article 102004.
Cai, D., Chai, Y., & Tian, Z. (2021). Quantitative evaluation of Jilin province's digital economy policy texts based on the PMC index model. *Information Science, 39*(12), 139–145.
Chen, Y., & Li, G. (2021). Evaluation of the effectiveness of science and technology policies and their development trends. *Library & Information*, (6), 96–106.
Chen, J., Gong, Z., & Liu, W. (2019). A nonparametric model for online topic discovery with word embeddings. *Information Sciences, 504*, 32–47.
Correa, A. S., Jr, A. M., & Silva, F. S. C. D. (2020). A deep search method to survey data portals in the whole web: Toward a machine learning classification model. *Government Information Quarterly, 37*(4), Article 101510.
Dai, S., Zhang, W., Zong, J., Wang, Y., & Wang, J. (2021). How effective is the green development policy of China's Yangtze River economic belt? A quantitative evaluation based on the PMC-index model. *International Journal of Environmental Research and Public Health, 18*(14), 7676.
Dai, S., Zhang, W., & Lan, L. (2022). Quantitative evaluation of China's ecological protection compensation policy based on PMC index model. *International Journal of Environmental Research and Public Health, 19*(16), Article 10227.
Du, D., Yuan, L., & Gao, K. (2019). Evaluation on Beijing-Tianjin-Hebei science and technology innovation policy in 2010—2017. *Forum on Science and Technology in China, 10*, 100–109.
Gao, X., & Sun, Y. (2022). Quantitative evaluation of rural revitalization policy texts based on the PMC index model. *Statistics & Decision, 38*(19), 57–60.
Hu, J., Fu, W., Qian, W., & Tian, P. (2021). Research on policy text classification model based on topic model and attention mechanism. *Information Studies: Theory & Application, 44*(7), 159–165.
Hu, F., Qi, X., & Wang, X. (2020). Quantitative evaluation of robot industry policies based on the PMC index model: A case study of 8 robot industry policy intelligence items. *Journal of Intelligence, 39*(1), 121–129.
Hu, F., Wen, Z., Shen, J., & Yao, Y. (2020). Quantitative evaluation of big data policies from the perspective of the intelligence process: A case study of 11 national-level big data policies. *Forum on Science and Technology in China*, (4), 30–41.
Huo, F., & Huo, C. (2023). How China governs data: Evidence mined from the central government policy documents. *Proceedings of the Association for Information Science and Technology, 60*(1), 992–994.
Huo, C., Yin, Z., Yang, Y., Yang, W., & Ru, R. (2024). Research on automatic identification method of policy irony comments based on large language models. *Journal of the China Society for Scientific and Technical Information, 43*(12), 1414–1424.
Huo, C., Li, X., Zhang, C., & Huo, F. (2024). The mining of China's policies against COVID-19 based on a policy target and tool co-evolution framework. *Information Development*, Article 02666669241289931.
Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150.
Kuo, J. J. (2014). An automatic library data classification system using layer structure and voting strategy. *Lecture notes in computer scienceI* (pp. 279–287). Springer. https://doi.org/10.1007/978-3-319-12823-8_29
Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *ArXiv Preprint*. https://doi.org/10.18653/v1/N16-1062
Lehtoranta, S., Nissinen, A., Mattila, T., & Melanen, M. (2011). Industrial symbiosis and the policy instruments of sustainable consumption and production. *Journal of Cleaner Production, 19*(16), 1865–1875.
Li, J., & Chang, X. (2014). A synergistic study of innovation policies in China's distribution industry. *Journal of Business Economics*, (9), 5–16.
Li, Z., & Guo, X. (2022). Quantitative evaluation of China's disaster relief policies: A PMC index model approach. *International Journal of Disaster Risk Reduction, 74*, Article 102911.
Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology*, 1–41.
Linder, F., Desmarais, B., Burgess, M., et al. (2020). Text as policy: Measuring policy similarity through bill text reuse. *Policy Studies Journal, 48*(2), 546–574.
Liu, P., Qiu, X., Chen, X., Wu, S., & Huang, X. (2015). Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/D15-1280
Liu, Y., Li, J., & Xu, Y. (2022). Quantitative evaluation of high-tech industry policies based on the PMC-index model: A case study of china's Beijing-Tianjin-Hebei region. *Sustainability, 14*(15), 9338.
Luo, P., Wang, Y., & Wang, J. (2020). Automatic discipline classification for scientific papers based on a deep pre-training language model. *Journal of the China Society for Scientific and Technical Information, 39*(10), 1046–1059.
Ma, X., & Ruan, Y. (2023). How to evaluate green development policy based on the PMC index model: Evidence from China. *International Journal of Environmental Research and Public Health, 20*(5), 4249.
Mcwilliam, W., Brown, R., Eagles, P., & Seasons, M. (2015). Evaluation of planning policy for protecting green infrastructure from loss and degradation due to residential encroachment. *Land Use Policy, 47*, 459–467.
Minaei, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys (CSUR), 54*(3), 1–40.
Poland, O. F. (1974). Program evaluation and administrative theory. *Public Administration Review, 34*(4), 333–338.
Ruiz Estrada, M. A. (2011). Policy modeling: Definition, classification and evaluation. *Journal of Policy Modeling, 33*(4), 523–536.
Ruiz Estrada, M. A., Yap, S. F., & Nagaraj, S. (2008). Beyond the ceteris paribus assumption: Modeling demand and supply assuming omnia mobilis. *International Journal of Economic Research, 5*(2), 185–194.
Shen, Z., Li, Y., Ding, Q., Wang, J., & Bai, Q. (2022). Research on science and technology policy text classification based on BERT model. *Digital Library Forum, 2022*(1), 10–16.
Shen, S., Chen, M., Feng, S., Xu, Q., & Liu, J. (2023). ChpoBERT: A pre-trained model for Chinese policy texts. *Technical Information, 42*(12), 1487–1497.

Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service & social action programs*. Russell Sage Foundation.
Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv: 1503.00075*.
Wang, X. (2014). An empirical study on the performance evaluation of regional technological innovation policies–based on the perspective of correlation and gray correlation analysis. *Science Research Management, 35*(5), 38–43.
Wang, Z., Li, J., Wang, B., & Zhang, B. (2024). The policy evaluation based on an improved hierarchical game cross-efficiency model. *Journal of Management Sciences in China, 27*(2), 18–29.
Wang, J., Yang, Q., & Zhang, Y. (2019). Quantitative evaluation of civil-military integration policies based on PMC-AE index modeling. *Journal of Intelligence, 38*(4), 66–73.
Xu, G., Xue, M., & Guo, X. (2022). Evaluation of the administrative capacity of Sichuan, Shanxi and Chongqing governments in promoting S&T development - Based on the perspective of S&T policies. *Soft Science, 36*(2), 8–14.
Yan, W., Hu, B., Liu, Y. L., Li, C., & Song, C. (2024). Does usage scenario matter? Investigating user perceptions, attitude and support for policies towards ChatGPT. *Information Processing & Management, 61*(6), Article 103867.
Yan, Q. (2002). *Public policy studies*. Nanjing: Nanjing University Press.
Yang, T., Xing, C., & Li, X. (2021). Evaluation and analysis of new-energy vehicle industry policies in the context of technical innovation in China. *Journal of Cleaner Production, 281*(4), Article 125126.
Yimsuk, A., & Thammaboosadee, S. (2024). Evaluation of Thailand's COVID-19-related policies and their impact on the stock market using a PMC index model approach. *Cogent Social Sciences, 10*(1), Article 2285252.
Zhang, Y., & Geng, Z. (2015). The quantitative evaluation of regional science and technology innovation policy: Based on the index of PMC model. *Science and Technology Management Research, 35*(14), 26–31.
Zhang, Y., & Yang, J. (2011). Reseach on automatic classification for Chinese text based on KNN and SVM. *Information Science, 29*(9), 1313–1317.
Zhang, W., Tang, K., Wu, H., Wang, M., Shen, Y., Hou, G., ... & Lu, W. (2024). Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*.
Zhang, Y., Huang, B., Yuan, W., Jiang, Z., Peng, L., Chen, S., & Tan-Soo, J. S. (2025). Expert-level policy style measurement via knowledge distillation with large language model collaboration. *Information Processing & Management, 62*(4), Article 104090.
Zhang, Y., Wang, T., Wang, C., & Cheng, C. (2023). Quantitative evaluation of China's CSR policies based on the PMC-index model. *Sustainability, 15*(9), 7194.
Zhang, Z., Zhao, Y., & Liu, H. (2022). Construction of a practical application-oriented automatic classification engine for scientific literature. *Journal of Library Science in China, 48*(4), 104–115.
Zhao, J., Song, R., Yue, C., Wang, Z., & Xu, H. (2023). Weak-PMLC: A large-scale framework for multi-label policy classification based on extremely weak supervision. *Information Processing & Management, 60*(5), Article 103442.
Zhao, Y., Sun, S., & Zhang, J. (2024). A quantitative evaluation of the US science and technology competition policy toward China. *Journal of the China Society for Scientific and Technical Information, 43*(12), 1399–1413.
Zhu, X., Jiang, J., & Chen, Y. (2024). A comparative study of chinese and U.S. artificial intelligence policies based on the policy modeling consistency (PMC) index model. *Science and Technology Management Research, 44*(8), 20–30.
Zuo, L., Zhao, Z., & Wang, D. (2024). Research on the automatic classification of Siku Quanshu based on large language models. *Journal of Information Resources Management, 14*(5), 22–35.