Knowledge Organization

*Knowl. Organ.* **2026**; 53(6): 54147
https://doi.org/10.31083/KO54147

*Article*

# Enhancing Large Language Model Performance in Policy Tool Classification Through Coordinated Prompting, Rule Mapping, and Metadata

Yanhong Wang[1,2], Junping Qiu[2,*]

[1]School of Management, Hangzhou Dianzi University, 310000 Hangzhou, Zhejiang, China
[2]China Academy of Science and Education Evaluation, Hangzhou Dianzi University, 310000 Hangzhou, Zhejiang, China
*Correspondence: jpqiu@hdu.edu.cn (Junping Qiu)

Academic Editor: Natália Tognoli
Submitted: 22 May 2026 Revised: 31 July 2026 Accepted: 7 August 2026 Published: 23 November 2026



Dr. Yanhong Wang is an Assistant Researcher at the China Academy of Science and Education Evaluation, Hangzhou Dianzi University, China. She received her PhD from Nanjing University. Her research centers on the dynamics of media use behavior, reading behavior, and information access, with a particular focus on quantitative research methods and cohort analysis. She has published work on the changing trends of media use and reading behavior among various demographic groups in China. Her current research explores how digital transformations influence information acquisition habits across different age groups.



Professor Junping Qiu is a Professor and Researcher at the China Academy of Science and Education Evaluation, Hangzhou Dianzi University. He holds a Bachelor's degree from Wuhan University. With extensive experience in information science, his expertise spans user behavior, scientometrics, and digital media. His recent work investigates the driving mechanisms behind digital hoarding, online health information seeking, and knowledge exchange in social Q&A communities. He has published widely in leading journals on topics ranging from scientific collaboration to AIGC disinformation governance.

**Abstract**

The implicit semantics and overlapping boundaries of policy tools make policy tool classification a challenging multi-label text classification task. Related studies typically rely on single performance-enhancement strategies, which are insufficient for resolving semantic ambiguity in complex policy texts. This study presents a coordinated framework to improve the performance of large language models (LLMs) in policy tool classification. The proposed framework coordinates three complementary strategies, namely prompt engineering (augmented with few-shot learning), rule mapping, and policy metadata fusion. Instead of being treated as isolated, these strategies are organized as complementary task guidance mechanisms that jointly enhance semantic reasoning and label discrimination. Experiments are conducted on a manually annotated dataset of policy tools, and the framework is further evaluated across multiple mainstream LLMs. The fully integrated model achieves a substantial performance improvement over the baseline, boosting the F1 score to 77.39% (a 54.1% relative increase) while simultaneously reducing the Hamming loss from 0.3062 to 0.1279 (a 58.2% reduction). Fine-grained analysis reveals complementary effects across strategies, with metadata fusion showing particular effectiveness for context-sensitive classifications. The results of experiments on multiple LLMs demonstrate the robustness and generalizability of our framework. Policy tool classification is conceptualized as a multi-strategy semantic reasoning process rather than as a single-strategy text classification task. The proposed model-agnostic framework systematically integrates complementary task guidance strategies to improve policy knowledge organization and automated policy analysis, providing methodological support for intelligent policy information systems and information science research on document classification and knowledge organization.

**Keywords:** large language model; policy tool; prompt engineering; rule mapping; metadata; multi-label classification

**Copyright**: © 2026 The Author(s). Published by IMR Press.
This is an open access article under the CC BY 4.0 license.

**Publisher's Note**: IMR Press stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

Policy documents are an important type of semi-structured and conceptually complex information resources, encompassing heterogeneous policy tools, implicit semantic relations, and multi-level governance intentions. The extraction of such information and its organization into structured knowledge units is crucial for policy knowledge management, evidence-based decision support, and large-scale policy analysis. However, the semantic ambiguity and multi-label characteristics of policy tools greatly increase the difficulty of their automatic identification, making it a typical challenge in contemporary information extraction and knowledge organization research.

This study focuses on policy tool classification, which is formulated as a multi-label policy text classification task. Unlike conventional policy text classification, which assigns an entire document to a predefined thematic category, policy tool classification requires the identification of multiple policy tool types that may coexist within the same text segment. Consequently, this task combines the characteristics of text classification and fine-grained information extraction, requiring models to distinguish semantically similar policy tools while handling multiple labels.

Recent progress in large language models (LLMs) has transformed the landscape of text understanding and semantic processing. Owing to their powerful context modeling capabilities, LLMs have been increasingly applied to information science tasks such as entity extraction, document classification, and knowledge structuring. However, the direct application of general LLMs to the identification of multi-label policy tools in specific domains is challenging, mainly because (i) the categories of policy tools usually have implicit or overlapping semantic boundaries (which results in a relatively high degree of information uncertainty) and (ii) LLM performance largely depends on how task guidance and domain-specific information are incorporated during inference. No single strategy, whether relying solely on prompt engineering, few-sample examples, rule-based constraints, or metadata, can fully cover the multi-dimensional semantic conditions required for policy tool identification. Each strategy has its inherent limitations: prompt engineering can activate task awareness but lacks precise domain knowledge guidance; few-sample learning provides reference examples but is prone to selection bias; rule-based mapping imposes strict constraints but cannot handle unaddressed edge cases; metadata fusion provides document-level background information but struggles to distinguish fine-grained labels when used alone. The complementarity between the advantages and limitations of these strategies constitutes the core research motivation for multi-strategy synergy.

To address these challenges, we developed a progressive coordinated multi-strategy framework for LLM-based policy tool identification. This framework combines prompt engineering, few-shot learning, rule mapping, and policy metadata fusion to provide complementary task guidance signals that jointly improve semantic reasoning and multi-label discrimination, thus establishing a richer information environment for LLM reasoning. These strategies are not viewed as isolated engineering components but conceptualized as mutually reinforcing guidance mechanisms that reduce semantic ambiguity and improve classification robustness in complex policy texts.

This research makes the following contributions. First, we establish a theoretical foundation for coordinated multi-strategy frameworks by demonstrating how prompt engineering, rule mapping, and metadata fusion jointly address semantic ambiguity and multi-label classification in policy tool identification, grounding our approach in both knowledge integration and policy tool classification theory. Second, we propose a progressive framework that systematically integrates coordinated prompt engineering, few-shot learning, rule mapping, and metadata fusion. This framework consistently outperforms both the baseline model and single-strategy approaches, with experiments across multiple mainstream LLMs confirming its robustness and model-agnostic applicability. Third, our framework provides an effective solution for automated policy tool extraction and policy knowledge organization, offering methodological support for intelligent policy analysis and evidence-based decision support in domains characterized by implicit semantics and multi-label structures.

Unlike previous studies, which primarily improved policy tool classification through feature engineering, deep neural architectures, or fine-tuning pre-trained language models, this study focuses on the systematic integration of complementary task guidance strategies to enhance the semantic reasoning capability of LLMs. This perspective shifts the emphasis from optimizing model architectures to improving inference through coordinated strategy integration, thereby providing a more general and transferable framework for LLM-based policy text classification.

## 2. Literature Review and Theoretical Foundation

### *2.1 Automatic Policy Tool Classification Methods*

Early research on policy tool identification primarily relied on expert manual interpretation and content analysis methods, with dictionary- and rule-based approaches becoming increasingly adopted with the development of computational social sciences. However, the dictionary method is constrained by the completeness and domain adaptability of dictionaries and therefore struggles to address the intricate semantics and contextual nuances inherent in policies. This limitation can be addressed using machine learning techniques. The related research frequently employs traditional classifiers such as support vector machines (SVMs) and random forests while utilizing bag-of-words models and Term Frequency–Inverse Document Frequency (TF-IDF) as text features. Traditional machine learning methods

heavily depend on feature engineering but exhibit a limited capacity to capture deep semantic features and therefore encounter performance bottlenecks.

Inspired by deep learning technologies, Li et al. (2021) employed convolutional neural networks to automatically recognize policy tools. Zhao and Li (2022) optimized the Bidirectional Encoder Representations from Transformers (BERT) model to facilitate the automatic classification of these tools. Huo et al. (2023) developed a more sophisticated automatic classification model based on WordBERT and Bidirectional Long Short-Term Memory (BiLSTM) architectures. Li et al. (2024) improved the processing capability of representative deep learning models in identifying technology policy tools based on Easy Data Augmentation. Hu et al. (2024) achieved notable performance improvements in policy tool identification by fine-tuning pre-trained language models while incorporating decision enhancement methods from ChatGPT.

Although deep learning substantially improves semantic representation, existing methods remain constrained by limited domain knowledge utilization, insufficient contextual reasoning, and weak capability for multi-label semantic discrimination. These limitations inspire the exploration of LLM-based approaches.

### *2.2 Applications of LLMs in Policy Text Analysis*

LLMs—exemplified by ChatGPT, DeepSeek, and LLaMA (Large Language Model Meta AI) and characterized by billions of parameters and extensive knowledge and reasoning capabilities acquired from vast datasets—offer promising solutions to the aforementioned challenges. These models demonstrate proficiency in zero/few-shot learning, comprehend complex instructions, and generate coherent text, thus being well suited for tasks involving policy content parsing and classification.

The potential applications of LLMs in policy analysis are drawing increasing attention. Rodriguez et al. (2024) explored and assessed the use of LLMs as an effective alternative for identifying and categorizing the various data practice disclosures present within privacy policy texts. Shen et al. (2025) proposed the Policy RAG (Retrieval-Augmented Generation) framework, which integrates large-scale language models with policy information retrieval mechanisms to enhance the efficiency of retrieving and utilizing policy texts. Lu et al. (2026) incorporated a Policy Modeling Consistency Model (PMC) into LLMs while developing automated PMC scoring systems. Churchill et al. (2026) used OpenAI's GPT model to annotate public policy texts and benchmarked the performance of two prompting strategies applied to distinct GPT models against human annotations, achieving a consistency rate of 86.25% between GPT and human outputs. Liu and Sun (2025) utilized GPT-4 to analyze themes within education policies related to coding, achieving an effectiveness surpassing those of traditional natural language processing methods by over 25% and sentiment analysis outcomes more closely aligned with expert human judgments.

Although LLMs possess strong generalization capabilities, their performance in domain-specific classification tasks often depends on effective task guidance. Rather than modifying model parameters, recent studies have improved LLM performance by integrating complementary inference strategies. Representative approaches include prompt engineering, which specifies task objectives through natural language instructions (Liu et al., 2023); few-shot learning, which enables in-context semantic generalization through representative demonstrations (Brown et al., 2020); symbolic rule guidance, which introduces explicit domain constraints into the reasoning process (Ribeiro et al., 2020; Sivarajkumar, 2025); and retrieval- or context-based augmentation, which supplements model reasoning with external contextual information (Lewis et al., 2020). Collectively, these approaches enrich task-specific knowledge without modifying model parameters and have become important paradigms for improving LLM performance in specialized domains.

Existing studies have primarily investigated these strategies independently, paying limited attention to their coordinated application in policy text classification. However, policy tool classification requires not only semantic understanding but also explicit policy rules, representative examples, and contextual policy information to resolve semantic ambiguity and identify multiple policy functions within the same text. Consequently, no integrated-strategy framework capable of coordinating complementary strategies for policy tool classification is available, which has inspired the development of the multi-strategy synergy approach proposed herein.

### *2.3 Theoretical Foundation of Multi-Strategy Knowledge Integration*

Policy tool classification is fundamentally a knowledge-intensive semantic understanding task. Unlike conventional text documents, policy documents often contain implicit expressions, overlapping policy objectives, and multiple policy tools embedded within a single provision. According to the policy tool framework proposed by Hood (1983) and developed by Schneider and Ingram (1990), policy tools represent functional mechanisms of government intervention rather than explicit lexical categories. Consequently, the identification of policy tools requires understanding policy intentions beyond surface textual semantics while resolving semantic ambiguity and multi-label relationships. Thus, the effectiveness of policy tool classification depends not only on language understanding but also on the integration of diverse forms of policy knowledge.

The theoretical basis of this study is grounded in knowledge integration theory, which argues that complex decision-making is achieved through the integration of het-

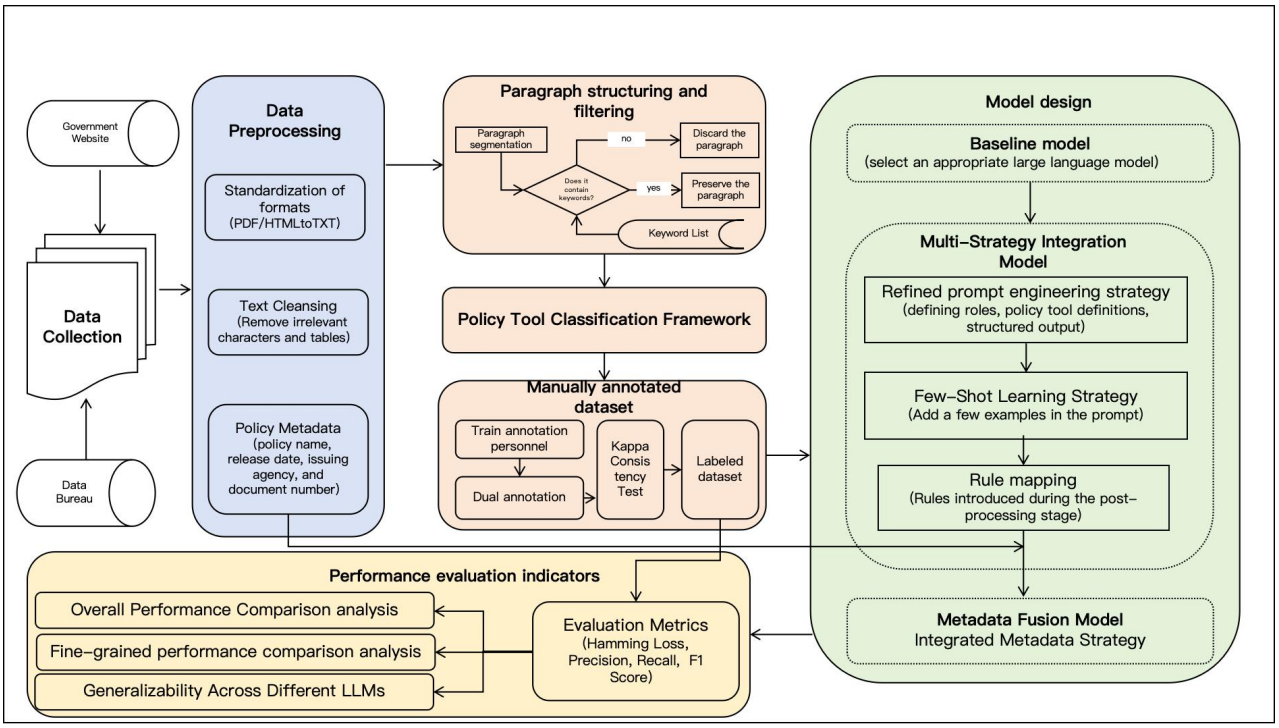


**Fig. 1. Flowchart**. LLMs, large language models.

erogeneous yet complementary knowledge rather than reliance on a single knowledge source (Grant, 1996). Knowledge integration allows different types of knowledge to exhibit mutual complementarity by providing distinct but coordinated cognitive functions. This perspective has been widely adopted in knowledge management research and extended to artificial intelligence, in which case the integration of heterogeneous knowledge sources enhances semantic representation, reasoning capability, and decision robustness (Kraaijenbrink and Wijnhoven, 2008; Zhao et al., 2020). Following this perspective, LLM-based policy text classification can be viewed as a knowledge integration process in which different knowledge sources jointly support semantic understanding instead of independently optimizing model performance.

Building upon this theoretical foundation, we herein conceptualize coordinated prompt engineering (encompassing few-shot learning), rule mapping, and policy metadata fusion as three complementary strategy modules within a unified coordinated framework. Prompt engineering provides task knowledge by explicitly defining classification objectives and decision boundaries. Few-shot learning contributes exemplar knowledge that allows the model to infer category semantics from representative instances. Rule-based mapping introduces symbolic policy knowledge to constrain category boundaries and reduce semantic ambiguity. Policy metadata fusion supplements contextual knowledge by incorporating policy attributes that are difficult to infer from textual content alone. Rather than functioning independently, these four knowledge sources collaboratively support semantic discrimination and multi-label reasoning, thereby enhancing the capability of LLMs to identify policy tools in complex policy documents. This theoretical framework provides the conceptual basis for the proposed multi-strategy synergy approach.

## 3. Methods

The process of enhancing LLM performance in policy tool classification through the coordinated multi-strategy framework is illustrated in Fig. 1.

### *3.1 Dataset*

This study draws on 113 authoritative policy documents issued by China's central and provincial local governments between January 2020 and April 2025. In April 2020, the "Opinions of the Central Committee of the Communist Party of China and the State Council on Establishing a More Complete Mechanism for the Market-Based Allocation of Factors" first listed data as a production factor, which marked the official launch of the policy for data factor marketization. The selected timespan was chosen to capture the context before and after this event.

The analysis was deliberately confined to policies issued at the central and provincial administrative levels, while municipal, county-level, and non-governmental documents (e.g., industry association guidelines or enterprise white papers) were excluded. This scope definition helped focus on interactions between central and local levels in terms of top-level institutional design and implementation at the provincial level.

**Table 1. Classification of policy tools.**

| Type | Explanation and description |
|---|---|
| Regulatory tool | The data element market involves core issues such as data security, privacy protection, and fair competition. It is necessary to draw red lines through mandatory rules to provide a basic framework for market operation. |
| Incentive tool | The core means of market-based resource allocation is economic incentives. Through fiscal, taxation and financial means, the enthusiasm of market entities should be directly mobilized to solve the problems of "market failure" and "insufficient investment". |
| Service tool | The development of the data element market requires infrastructure and public service support. By building public platforms and providing common technical services, the government can significantly reduce the costs and difficulties for market entities to participate in data trading and utilization. |
| Information tool | In the data element market, information asymmetry is prominent. The government can provide clear signals and knowledge to market entities by issuing guidelines, white papers, best practices and other means, promoting information sharing and trust building. |
| Organizational/cooperative tool | The circulation of data elements involves multiple entities such as the government, enterprises, and research institutions. The government needs to step in to coordinate relations, establish cooperation mechanisms and cultivate industry organizations, so as to form a collaborative and co-governed industrial ecosystem. |
| Environmental creation tool | The data element market is a brand-new and complex system. The government needs to create a favorable institutional environment, credit environment and social and cultural environment for the long-term healthy development of the entire market through macro means such as top-level design, strategic planning and cultural construction. |

Policy texts were retrieved from two complementary sources, namely (i) official government websites (those of the State Council, central ministries, commissions (https://www.gov.cn), and provincial (autonomous region, municipality) governments) and (ii) professional legal and policy databases (Peking University Law (https://www.pkulaw.com) and the policy text database of China National Knowledge Infrastructure).

Policy documents had to be issued by the central or provincial government, the effective date had to lie between January 2020 and April 2025, and the documents had to directly involve the marketization of data elements or contain specific chapters on the development of the data element market. Internal administrative notices, meeting minutes, and duplicates were excluded.

#### 3.1.1 Data Preprocessing

The format of policy documents was standardized by converting PDF/HTML to TXT, and the text was then cleaned by removing irrelevant characters and tables. Policy text metadata, such as policy name, release time, file number, and issuing institution, were acquired and brought to a common standard, e.g., by unifying the release time format of policies from different information sources and standardizing the names of policy-issuing institutions.

Policy texts were subjected to paragraph processing and screening. Given that these texts are generally long and complex, direct entire-text analysis by LLMs may lead to computational resource wastage and interference from irrelevant information. Policy unit screening at the paragraph level retained complete contextual information. Paragraph segmentation afforded 14,180 fragments. Subsequently, a keyword table for the marketization policy of data elements was constructed, and paragraph-level preliminary screening was performed to extract research topic–related paragraphs and rapidly filter out unrelated ones. After manual proofreading, 584 relevant data points were obtained and used for the subsequent data annotation of policy tool labels and LLM recognition.

The 113 policy documents represented the source corpus, whereas the effective sample size for classification was 584 paragraph-level policy units.

#### 3.1.2 Policy Tool Classification Framework

To identify policy tools in policy texts, one should first establish a tool classification system. Owing to differences in research subjects and objectives, current policy tool classification standards vary, which results in diverse classification systems. Hood (1983) introduced the influential NATO (Nodality, Authority, Treasure, Organization) framework, classifying policy tools into information, authority, treasure, and organization. McDonnell and Elmore (1987) proposed a four-category (mandates, inducements, capacity building, and system-changing instruments) policy tool framework. These frameworks provide important theoretical foundations for understanding how governments intervene through different policy tools.

However, no single classification framework is universally applicable. Building upon existing policy tool classification frameworks and considering the characteristics of policies pertaining to the marketization of data elements, we adopted a six-category (regulatory, incentive, service,

**Table 2. Reliability test results for policy tool coding.**

| Policy tools | Regulatory tool | Incentive tool | Service tool | Information tool | Organizational/ cooperative tool | Environmental creation tool |
|---|---|---|---|---|---|---|
| Kappa value | 0.830 | 0.817 | 0.828 | 0.820 | 0.826 | 0.808 |
| Quantity | 124 | 46 | 269 | 134 | 73 | 241 |

**Table 3. Keyword mapping table.**

| Tool types | Keywords | Tool types | Keywords |
|---|---|---|---|
| Regulatory tool | "Compliance", "Requirements", "Implementation", "Supervision", "Approval", "License", "Legal Provisions", "Must", "Standards" | Information tool | "Data sharing", "information", "dissemination", "standards development", "openness", "information disclosure" |
| Incentive tool | "Subsidy", "Preferential Treatment", "Reward", "Incentive", "Reduction or Exemption", "Funding" | Organizational/ cooperative tool | "Collaboration", "Cooperation", "Linkage", "Synergy", "Joint Efforts", "Joint Construction", "Coordination" |
| Service tool | "Construction", "Provide", "Service", "Platform", "System", "Support", "Assistance", "Training" | Environmental creation tool | "Planning, Design", "Policy", "Environment, Ecology", "Improvement", "Layout", "Strategy", "Mechanism", "Cultivation", "Development" |

information, organizational/cooperative, and environmental creation) tool classification system, with relevant definitions and classification criteria presented in Table 1.

#### 3.1.3 Manually Indexed Datasets

The experimental samples were manually labeled to verify the effectiveness of LLM-generated labels, evaluate their consistency with manually generated labels, and verify the impact of LLMs implementing different strategies on the results of label generation. In the first stage, two researchers holding PhD degrees in information resource management independently annotated policy tools after completing standardized training on annotation guidelines. Inter-rater reliability was assessed using Cohen's kappa coefficient ($\kappa$; Cohen, 1960), which measures agreement between annotators while accounting for chance agreement (Table 2). The average $\kappa$ was 0.822, suggesting a high level of inter-rater agreement (Landis and Koch, 1977). Discrepancies in annotation results were resolved through discussion between the two annotators until a consensus was achieved.

### *3.2 Model Design*

Guided by the theoretical rationale of multi-strategy synergy, we developed a progressive framework consisting of three models with increasing levels of strategy integration. Instead of modifying the internal LLM architecture, the proposed framework incrementally incorporated complementary task guidance strategies to examine their individual and combined contributions to policy tool classification. Each successive model extended the previous one by introducing an additional strategy, which enabled the systematic evaluation of how prompt engineering, few-shot learning, rule mapping, and metadata fusion jointly improve semantic reasoning and multi-label classification performance.

#### 3.2.1 Baseline Model (Model-1)

The baseline model was designed to evaluate the original ability of LLMs to identify policy tools with zero samples, thereby providing a comparative baseline for subsequent performance optimization.

DeepSeek-R1-Distill-Llama-70B from DeepSeek was selected as the benchmark model for two reasons. First, the investigated policy texts were in Chinese, and the DeepSeek model has been trained on a substantial amount of Chinese data, thus showing accuracy in understanding Chinese semantics. Second, this model was distilled from Llama-70B, which is known for its robust language comprehension abilities. Compared with the original Llama-70B model, the distilled version reduces computational costs while maintaining high performance levels. Moreover, the distilled model exhibits strong contextual understanding and adaptability within coordinated multi-strategy frameworks, thus being an ideal choice for tasks involving the identification of Chinese policy tools and providing a solid foundation for future optimization and expansion.

Fragments of policy texts to be analyzed were directly input into the model, and an extremely simple task instruction (Prompt 1) was used to analyze the text output without supplying any examples or background knowledge. The outputs were mapped to six predefined categories of policy tools through string matching techniques. The results were entirely dependent on the semantic understanding capabilities of the model, reflecting its initial performance prior to optimization.

**Prompt 1:**

"Please analyze the following policy texts and identify the types of policy tools used within them. Please select from one of six categories: regulatory

tool, incentive tool, service tool, information tool, organizational/cooperative tool, or environmental creation tool."

3.2.2 Multi-Strategy Integration Model (Model-2)

The multi-strategy integration model introduced three strategies: prompt engineering, few-shot learning, and rule mapping.

(1) Refinement of the engineering strategy for prompt words. Initially, the role of the model was defined as that of a senior policy researcher, with its identity and mission objectives clearly articulated at the prompt outset. Furthermore, the model was emphasized to be familiar with the policy domain concerning the marketization of data elements and endowed with professional knowledge and analytical capabilities allowing it to comprehend mission requirements from a domain expert's perspective. Additionally, each type of policy tool was explicitly defined within the prompt to assist the model in accurately distinguishing between different tool types. Finally, output format requirements were distinctly outlined in the prompt, stipulating that the results are to be presented in a standardized JSON format to ensure uniformity and facilitate the parsing of output content (see Prompt 2).

**Prompt 2:**

"You are a senior policy researcher, skilled in using quantitative and qualitative methods to identify policy tool types, and familiar with key policy texts such as the 'Opinions of the Communist Party of China Central Committee and the State Council on Building a Data Infrastructure System to Better Leverage the Role of Data Elements.'"

Please first determine if the sentence contains policy tool behavior. If so, please identify the most suitable tool type (multiple selections allowed) and support your judgment with a single sentence.

Policy Tool Types:

- Regulatory Tool (Laws and Regulations, Standards, Supervision, Approval and Licensing)
- Incentive Tool (Fiscal Subsidies, Tax Incentives, Special Funds)
- Service Tool (Public Services, Technical Support, Platform Construction)
- Information Tool (Data Sharing, Standard Setting, Information Release)
- Organizational/Cooperative Tool (Chain Leader Enterprises, Cross-Regional Collaboration, Government-Enterprise Cooperation)
- Environmental creation tool (Top-Level Design, Strategic Planning, Business Environment, Innovation Ecosystem)

Please output the result in JSON format, including:

{"sentence": "...", "tool_type": ["Type 1", "Type 2"], "evidence": "Briefly explain the reasons and cite keywords"}

(2) Learning strategy utilizing a limited number of samples. Representative cases were integrated into prompts as contextual examples, accompanied by question–answer pairs for each type of policy tools.

(3) Rule reinforcement strategy. Recognition accuracy was further improved by building a keyword mapping library (Table 3) and introducing a rule validation mechanism in the post-processing stage.

**Example:**

"sentence": "All parties involved in the Trusted Data Space must comply with the provisions of the Cybersecurity Law, the Data Security Law, the Personal Information Protection Law, and other relevant laws and regulations, and implement requirements for data classification and grading protection, dynamic perception, risk identification, monitoring and early warning, emergency response, and governance and supervision, and establish a Trusted Data Space Security Management System."

"tool_type": ["Regulatory Tool", "Service Tool"],

"evidence": "The phrases 'must comply with legal provisions' and 'implement...requirements' clearly reflect legal regulations and regulatory requirements, classifying them as regulatory tools; 'establish a Trusted Data Space Security Management System' involves technical support and system construction, exhibiting characteristics of a service tool."

3.2.3 Metadata Fusion Model (Model-3)

Building upon the multi-strategy integration model, we introduced a metadata fusion component incorporating metadata from policy documents and thus provided supplementary contextual clues (including but not limited to policy name, release date, document number, and issuing institution) for core tasks. Policy name elucidated the theme and target area while assisting the model in determining appropriate tool types, release time reflected agenda priorities alongside historical context, and document numbers indicated levels of authority associated with the released policies, thereby refining potential classifications for tool types. Publishing institutions reflected functional tendencies, indicating the field and coordination level of policy tools. Through the synergy of multidimensional metadata, the model could comprehend policy intentions from various perspectives, which enabled a more accurate identification of tool types.

Each fragment of policy text was associated with policy name, publication date, document number, and metadata pertaining to the publishing institution. These text fragments and their associated metadata were integrated into a unified JSON format. The simultaneous dynamic integration of metadata such as policy name, release time, document number, and releasing organization into prompts (e.g., Prompt 3) aided the model in understanding policy intention and tool selection logic with greater precision. Additionally, these fields were incorporated into several sample examples.

**Prompt 3:**

"Policy Background:

- Policy Name: {item.get("source", "unknown")}

- Release Date: {format_release_date(item.get("release_date", "unknown"))}

- Document Number: {item.get("document_no", "unknown")}

- Issuing Agency: {clean_issuing_agency(item.get("issuing_agency", "unknown"))}

"You are a senior policy researcher,..."

Based on the following policy texts and their titles, release dates, and issuing agencies...

..."

The inclusion of metadata allowed the model to analyze policy intentions across multiple dimensions, including temporal context, thematic direction, authority level, and institutional functions, thus increasing the accuracy of analysis and reasoning.

## 4. Results

### *4.1 Performance Metrics*

In multi-label classification tasks, model performance evaluation necessitates a comprehensive consideration of multiple dimensions to effectively measure the ability to predict different label combinations. Herein, Hamming loss (HL), precision, recall, and the F1 score were used to quantitatively analyze model outcomes.

HL represents the ratio of labels used to assess mispredictions by the model within each sample, with lower values indicating reduced prediction error rates and enhanced performance.

$$HL = \frac{1}{M}\sum_{i=1}^{M}\frac{|\ Y_i \Delta \hat{Y}_i\ |}{|L|} \tag{1}$$

where $M$ is the number of samples, $Y_i$ is the set of true labels, $\hat{Y}_i$ is the set of predicted labels, $L$ is the total set of all possible labels, and $\Delta$ denotes the symmetric difference between these sets.

Precision, recall, and the F1 score are defined as

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

where TP is the number of samples correctly predicted as positive by the model, FP is the number of samples incorrectly predicted as positive by the model, and FN is the number of samples incorrectly predicted as negative by the model.

### *4.2 Overall Performance Comparison Analysis*

To systematically assess the impact of different optimization strategies on multi-label policy tool identification, we compared three model families: (i) Model-1 (baseline); (ii) Model-2 (multi-strategy integration), which incorporates prompt engineering, few-shot learning, and rule mapping; and (iii) Model-3 (metadata fusion), which augments the multi-strategy integration model with policy metadata. The main performance indicators (HL, precision, recall, and F1 score) of these models are listed in Table 4.

The F1 score and HL of Model-1 were determined as 0.5023 and 0.3062, respectively. Thus, although this model had a certain zero-shot recognition ability, the prediction results were noisy, and accuracy and recall were unbalanced, hindering compliance with the requirements of practical applications. Introducing prompt engineering (Model-2A) delivered the first major gain (F1 = 0.6703). Adding few-shot learning (Model-2B) further improved performance to reach a local peak (F1 = 0.7322) and markedly lowered HL to 0.1518, thus decreasing the incidence of multi-label errors. After the introduction of rule mapping (Model-2C), recall was maximized (0.8172) at the cost of reduced accuracy, which resulted in the F1 score being basically the same as that of Model-2B. This behavior indicates that the rules effectively captured more positive examples while introducing a certain number of misjudgments. Building on these advances, Model-3 attained the best overall results (F1 = 0.7739, HL = 0.1279) with a good balance between precision (0.7307) and recall (0.8226), which underscores the value of contextual policy metadata for accurate tool classification.

The superiority of the Model-2 over the baseline model validated the effectiveness of improving semantic

**Table 4. Comparison of model performance**.

| Model | | Performance indicators | | | |
|---|---|---|---|---|---|
| | | Hamming loss | Precision | Recall | F1 score |
| Model-1: Baseline Model | | 0.3062 | 0.4414 | 0.5828 | 0.5023 |
| Model-2: Multi-Strategy Integration Model | 2A: prompt | 0.2651 | 0.6264 | 0.7209 | 0.6703 |
| | 2B: prompt + few-shot | 0.1518 | 0.6895 | 0.7806 | 0.7322 |
| | 2C: prompt + few-shot + Rule Mapping | 0.1590 | 0.6620 | 0.8172 | 0.7315 |
| Model-3: Metadata Fusion Model | | 0.1279 | 0.7307 | 0.8226 | 0.7739 |

recognition ability through structured prompts, few-shot learning, and rule mapping. Furthermore, the metadata fusion model achieved optimal performance. The policy context information introduced by this model effectively enhanced the ability to distinguish semantically ambiguous regions in policy texts and hence, robustness and generalization ability in complex and polysemous policy tool classification tasks.

The results of three-stage model optimization fully validated the effectiveness of the proposed multi-strategy optimization framework. By combining prompt engineering, few-shot learning, rule mapping, and policy metadata fusion, we markedly enhanced performance in multi-label policy tool recognition tasks, demonstrating the crucial role of structured semantic constraints and contextual metadata in the customized application of LLM tasks.

### *4.3 Fine-Grained Performance Comparison Analysis*

The fine-grained analysis of different types of policy tools (Table 5) revealed differences in model performance across different categories and the effectiveness of improvement strategies.

Model-1 had serious problems in identifying "incentive tools" and "organizational/cooperative tools", as reflected by its extremely high recall (>0.88) and low accuracy (~0.21). Thus, this model wrongly predicted a large amount of text from categories other than these two, featuring extremely blurred classification boundaries. "Service tools" presented the opposite problem, namely high precision but low recall, i.e., the model made accurate judgments but missed many positive examples. These findings confirm that policy tool classification is characterized by heterogeneous semantic complexity rather than uniform classification difficulty.

Prompt word introduction (Model-2A) provided clear classification definitions and thus improved F1 scores for all tool categories. For example, the F1 score of "incentive tools" increased from 0.339 to 0.530, which indicates that explicit instructions effectively narrowed the prediction range of the model. Few-shot learning (Model-2B) was the most effective single reinforcement strategy, effectively balancing accuracy and recall. For instance, the F1 score of "informational tools" increased from 0.584 to 0.699, and that of "organizational/cooperative tools" increased from 0.495 to 0.638. Thus, a small number of examples provided specific decision-making contexts for the model, greatly improving the clarity of classification boundaries. Upon the incorporation of rule mapping (Model-2C), recall further increased for "service tools", "organizational/cooperative tools", and "environmental creation tools", which demonstrates that explicit domain rules effectively captured policy-specific linguistic patterns that may not be fully represented through statistical learning alone.

Model-3 achieved the best combination of accuracy and recall in the vast majority of tool categories (Fig. 2).

For instance, the F1 scores of "regulatory tools" and "information tools" reached 0.794 and 0.766, respectively, both being the highest. In particular, the recall of "informational tools" increased to 0.910 without a notable accuracy decline, which indicates that metadata can strongly prompt the emergence of such tools. However, the accuracy of "organizational/cooperative tools" dropped to 0.390, although the corresponding recall was very high. This indicates a complex association between this type of tool and metadata features or confusion with other tools (especially "service tools"), which is a key area requiring further optimization.

Overall, the results of fine-grained analysis demonstrate that different strategies contributed to different aspects of semantic understanding instead of providing uniform performance improvements. Prompt engineering primarily clarified task objectives and category definitions, few-shot learning strengthened semantic generalization through representative examples, rule mapping introduced explicit domain constraints, and metadata fusion supplied document-level contextual information that complemented textual semantics. The sequential integration of these strategies allowed the LLM to progressively reduce semantic ambiguity and improve the accuracy of multi-label discrimination across heterogeneous policy tool categories, which explains the superior overall performance of the proposed coordinated multi-strategy framework.

### *4.4 Generalizability Across Different LLMs*

To evaluate whether the proposed coordinated multi-strategy framework is specific to a single LLM or can be generalized across different foundation models, we performed additional experiments using three representative LLMs (DeepSeek-R1-Distill-Llama-70B, GPT-4o, and Qwen-Turbo). For each model, baseline performance was compared with that achieved after applying the proposed framework (Table 6).

**Table 5. Comparison of model performance for each label.**

| Model | Tool type | Precision | Recall | F1-score |
|---|---|---|---|---|
| Model-1 | regulatory tool | 0.617 | 0.630 | 0.624 |
| | incentive tool | 0.207 | 0.941 | 0.339 |
| | service tool | 0.752 | 0.382 | 0.506 |
| | information tool | 0.408 | 0.639 | 0.498 |
| | organizational/cooperative tool | 0.217 | 0.882 | 0.349 |
| | environmental creation tool | 0.647 | 0.625 | 0.636 |
| Model-2A: prompt | regulatory tool | 0.652 | 0.815 | 0.725 |
| | incentive tool | 0.470 | 0.608 | 0.530 |
| | service tool | 0.813 | 0.747 | 0.778 |
| | information tool | 0.436 | 0.885 | 0.584 |
| | organizational/cooperative tool | 0.363 | 0.779 | 0.495 |
| | environmental creation tool | 0.789 | 0.597 | 0.680 |
| Model-2B: prompt + few-shot | regulatory tool | 0.761 | 0.761 | 0.761 |
| | incentive tool | 0.569 | 0.647 | 0.606 |
| | service tool | 0.733 | 0.845 | 0.785 |
| | information tool | 0.639 | 0.770 | 0.699 |
| | organizational/cooperative tool | 0.547 | 0.765 | 0.638 |
| | environmental creation tool | 0.708 | 0.750 | 0.728 |
| Model-2C: prompt + few-shot + rule mapping | regulatory tool | 0.676 | 0.772 | 0.721 |
| | incentive tool | 0.552 | 0.627 | 0.587 |
| | service tool | 0.723 | 0.880 | 0.794 |
| | information tool | 0.588 | 0.738 | 0.655 |
| | organizational/cooperative tool | 0.466 | 0.794 | 0.587 |
| | environmental creation tool | 0.709 | 0.837 | 0.768 |
| Model-3 | regulatory tool | 0.738 | 0.859 | 0.794 |
| | incentive tool | 0.613 | 0.745 | 0.673 |
| | service tool | 0.865 | 0.832 | 0.848 |
| | information tool | 0.661 | 0.910 | 0.766 |
| | organizational/cooperative tool | 0.390 | 0.882 | 0.541 |
| | environmental creation tool | 0.849 | 0.764 | 0.804 |

**Table 6. Generalizability of the proposed framework across different large language models.**

| LLM | Strategy | Hamming loss | Precision | Recall | F1 |
|---|---|---|---|---|---|
| DeepSeek-R1-Distill-Llama-70B | Baseline | 0.3062 | 0.4414 | 0.5828 | 0.5023 |
| DeepSeek-R1-Distill-Llama-70B | Proposed Framework | 0.1279 | 0.7307 | 0.8226 | 0.7739 |
| GPT-4o | Baseline | 0.2723 | 0.4927 | 0.5768 | 0.5314 |
| GPT-4o | Proposed Framework | 0.1458 | 0.6776 | 0.7488 | 0.6987 |
| Qwen-Turbo | Baseline | 0.3738 | 0.3600 | 0.4744 | 0.3608 |
| Qwen-Turbo | Proposed Framework | 0.1718 | 0.6393 | 0.8220 | 0.7192 |

Note: Precision, Recall, and F1 are reported as Macro-averaged scores. The Macro-F1 is calculated as the unweighted mean of the per-class F1 scores, rather than directly applying the formula to the aggregated Macro-Precision and Macro-Recall.

The proposed framework consistently improved classification performance across all evaluated LLMs, thus featuring strong generalizability. Although the three models exhibited different baseline capabilities, each substantially benefited from the integration of prompt engineering, few-shot learning, rule mapping, and metadata fusion.

DeepSeek-R1-Distill-Llama-70B achieved the best overall performance after the proposed framework was applied, with the F1 score increasing from 0.5023 to 0.7739 and HL decreasing from 0.3062 to 0.1279. Thus, multi-strategy synergy effectively enhanced semantic reasoning for policy tool classification. GPT-4o also substantially benefited from the proposed framework, with the macro F1 score increasing from 0.5314 to 0.6987 and HL decreasing from 0.2723 to 0.1458. These results demonstrate that the framework remained effective even when applied to high-

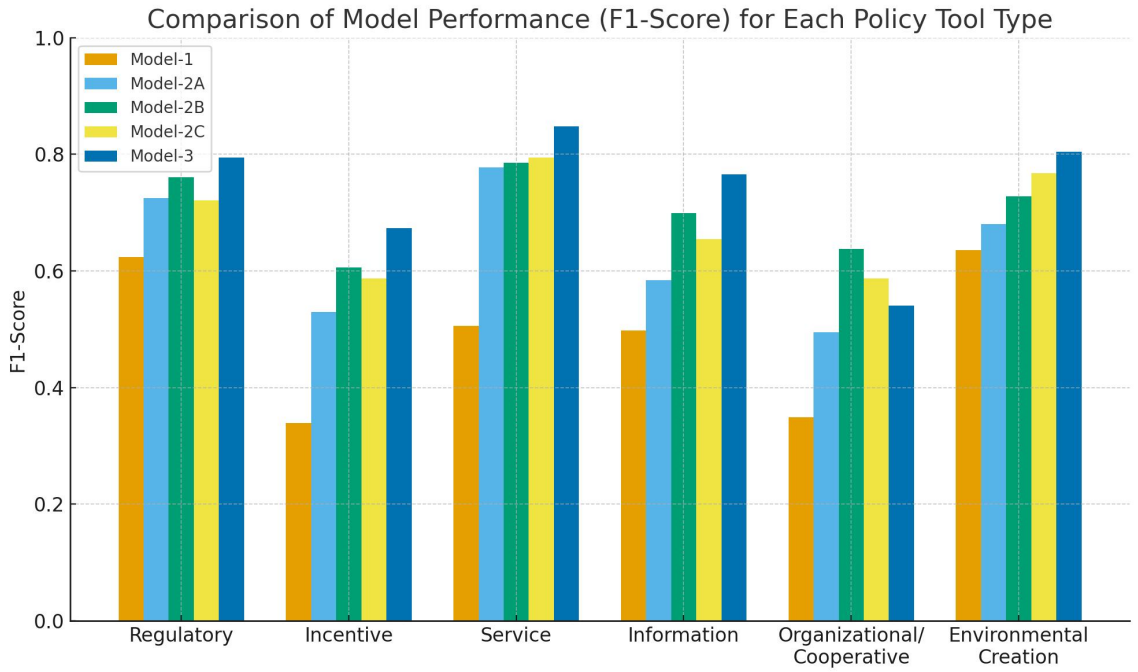


**Fig. 2. Comparison of model performance (F1 score) for each policy tool type**.

performance proprietary LLMs. Qwen-Turbo exhibited the largest relative improvement, with the macro F1 score increasing from 0.3608 to 0.7192 and HL decreasing from 0.3738 to 0.1718. This finding suggests that LLMs with weaker baseline performance more strongly benefit from explicit task guidance and contextual information.

Overall, although the absolute performance varied across LLMs, the proposed framework improved precision, recall, and macro F1 score while reducing HL for every evaluated model. These findings indicate that the performance gains primarily originated from the complementary effects of multi-strategy synergy rather than from the intrinsic capability of any particular LLM. Thus, the proposed framework features strong model-agnostic characteristics and can be readily transferred to different LLMs for policy tool classification tasks.

## 5. Discussion

The coordinated integration of multiple complementary strategies substantially improved the performance of LLM-based policy tool classification. Rather than functioning merely as independent enhancement techniques, prompt engineering, few-shot learning, rule mapping, and metadata fusion contributed complementary information signals that collectively reduced semantic ambiguity and strengthened the interpretive stability of LLMs. The progressive improvements observed across the three models, together with the consistent performance gains across different LLMs, indicate that the effectiveness of the proposed framework arises from strategy integration rather than from the capability of any individual model.

From a theoretical perspective, our findings support the rationale of the proposed multi-strategy synergy. Prompt engineering structures the task space through explicit instructional cues, while few-shot examples calibrate the model's sensitivity to contextual pattern variations. Rule mapping imposes domain-consistent boundaries through symbolic logic, and metadata fusion further anchors the model's output in document-level contextual attributes. The results of fine-grained analysis show that these strategies contribute differently across policy tool categories, highlighting their complementary rather than redundant roles in addressing the semantic complexity of policy texts.

From the perspective of information science, our findings suggest that policy tool identification is not purely a text classification task but represents a task-guided semantic reasoning process in which multiple forms of contextual information jointly support knowledge extraction from complex policy texts. The proposed framework illustrates how complementary guidance strategies can effectively organize task instructions, representative examples, symbolic constraints, and contextual metadata to improve semantic interpretation without modifying the underlying LLM architecture. This provides a practical strategy for enhancing intelligent information organization in domains characterized by implicit semantics and multi-label structures.

The proposed framework demonstrates practical relevance to automated policy text processing and knowledge management. By integrating multiple strategies, the model markedly improves the accuracy and reliability of policy tool extraction, a crucial step in policy indexing, re-

trieval, and evidence-based policy analysis. Strategy fusion is particularly beneficial in complex policy scenarios where tools are implicitly embedded or expressed through non-standard terminology. Therefore, this framework provides a methodological foundation for building more transparent, accurate, and scalable policy information systems.

## 6. Conclusions

This study demonstrates that a coordinated framework integrating prompt engineering (with few-shot learning), rule mapping, and metadata fusion can substantially improve LLM performance in multi-label policy tool classification, achieving a 54.1% F1 improvement over the baseline while remaining model-agnostic. By systematically coordinating these three strategy modules, the framework effectively resolves semantic ambiguity and handles multi-label characteristics in policy texts, all without altering the underlying LLM architecture. Using policies on the marketization of data elements as the empirical context, we demonstrated that the proposed framework consistently outperformed the baseline model and progressively improved classification performance as additional strategies were incorporated. The framework also exhibited strong generalizability across different LLMs, i.e., its effectiveness was not dependent on a specific foundation model.

This study makes three main contributions. First, it provides a theoretical perspective on a coordinated multi-strategy framework by explaining how different task guidance strategies complement one another to improve semantic discrimination in policy tool classification. Second, it proposes a progressive framework that allows LLMs to effectively utilize task instructions, representative examples, symbolic rules, and contextual metadata, offering a practical approach for complex multi-label classification tasks. Third, our findings demonstrate the potential of multi-strategy synergy for intelligent policy analysis, contributing to policy knowledge organization, automated policy tool extraction, and evidence-based decision support.

Despite these contributions, several limitations remain. First, although the proposed framework substantially improved overall performance, certain policy tool categories, particularly organizational/cooperative tools, were difficult to distinguish because of their implicit and overlapping semantic characteristics. Second, the experiments were conducted using GPT-4o and Qwen-Turbo as baseline models. We acknowledge that both versions are relatively outdated given the rapid evolution of LLM capabilities. However, these models were selected to maintain temporally consistent baselines. Meanwhile, our core contribution, the coordinated integration of prompting, rule mapping, and metadata, is model-agnostic and independent of any version-specific features. Future work should validate the framework on more recent model versions to assess its continued effectiveness. Third, the current framework adopted a fixed strategy integration process in which all strategies were manually designed and sequentially incorporated. Future studies may explore adaptive strategy selection or dynamic weighting mechanisms that automatically determine the most appropriate guidance strategy according to the characteristics of the input text. Fourth, the experiments were conducted using Chinese policy documents on the marketization of data elements. Although cross-model experiments demonstrated good robustness across different LLMs, further research is needed to evaluate the framework on other policy domains, alternative policy tool classification schemes, multilingual policy corpora, and cross-national policy datasets.

Future research may also investigate automatic rule induction, retrieval-based contextual enhancement, and interpretable reasoning mechanisms to improve the transparency and scalability of LLM-based policy text classification. Extending the proposed framework to broader information organization tasks involving implicit semantics and multi-label structures would further demonstrate its applicability to intelligent knowledge organization and document analysis.

## Availability of Data and Materials

Data will be made available on request.

## Author Contributions

YW performed the research, processed the experimental data, and wrote the manuscript. JQ designed the research framework, provided the conceptual structure, and revised the manuscript. Both authors contributed to editorial changes in the manuscript. Both authors read and approved the final manuscript. Both authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.


## Acknowledgment

Not applicable.

## Funding

This research was supported by the Leading Talent Cultivation Special Project of Zhejiang Province Philosophy and Social Science Planning (Grant No. 26YJRC011ZD-1YB) and the 78th Batch of General Projects of the China Postdoctoral Science Foundation, Funding Number: 2025M783237. The funding body had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.


## Conflicts of Interest

The authors declare no conflicts of interest.

## Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used DeepSeek tools to assist in developing the multi-source

knowledge fusion framework and as core classifiers within the experimental apparatus, as described in the Methodology and Results sections. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## References


Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020; 33: 1877–1901.

Churchill A, Pichika S, Xu C, Liu Y. GPT models for text annotation: An empirical exploration in public policy research. Policy Studies Journal. 2026; 54: e70034.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20: 37–46.

Grant RM. Prospering in Dynamically-Competitive Environments: Organizational Capability as Knowledge Integration. Organization Science. 1996; 7: 375–387. https://doi.org/10.1287/orsc.7.4.375

Hood CC. The Tools of Government. Macmillan: London. 1983.

Hu Z, Li P, Wang J, Xiong X. Research on Policy Tools Classification Based on ChatGPT Augmentation and Supervised Contrastive Learning. Computer Engineering and Applications. 2024; 60: 292–305. https://doi.org/10.3778/j.issn.1002-8331.2308-0354 (In Chinese)

Huo C, Huo F, Wang W, Yu Q, Yang G. Automatic classification method of policy tools based on WordBERT and BiLSTM. Documentation, Information & Knowledge. 2023; 40: 129–138. https://doi.org/10.13366/j.dik.2023.03.129 (In Chinese)

Kraaijenbrink J, Wijnhoven F. Managing heterogeneous knowledge: a theory of external knowledge integration. Knowledge Management Research & Practice. 2008; 6: 274–286. https://doi.org/10.1057/kmrp.2008.26

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33: 159–174.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020; 33: 9459–9474.

Li MN, Wang L, Lai HP. Identification of China's S&T policy properties based on deep learning. Science Research Management. 2024; 45: 1–11. https://doi.org/10.19571/j.cnki.1000-2995.2024.02.001 (In Chinese)

Li N, Jiang E, Zhu Y, Liu T. Policy tool identification method and empirical research based on deep learning. Library and Information Service. 2021; 65: 115–122. https://doi.org/10.13266/j.issn.0252-3116.2021.07.011 (In Chinese)

Liu A, Sun M. From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews. AERA Open. 2025; 11. https://doi.org/10.1177/23328584251374595

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys. 2023; 55: 1–35. https://doi.org/10.1145/3560815

Lu X, Yang Z, Huo C. A novel automated policy text evaluation framework integrating PMC into large language models. Information Processing & Management. 2026; 63: 104440. https://doi.org/10.1016/j.ipm.2025.104440

McDonnell LM, Elmore RF. Getting the Job Done: Alternative Policy Instruments. Educational Evaluation and Policy Analysis. 1987; 9: 133–152. https://doi.org/10.3102/01623737009002133

Ribeiro MT, Wu T, Guestrin C, Singh S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 4902–4912). 2020. https://doi.org/10.18653/v1/2020.acl-main.442

Rodriguez D, Yang I, Del Alamo JM, Sadeh N. Large language models: a new approach for privacy policy analysis at scale. Computing. 2024; 106: 3879–3903. https://doi.org/10.1007/s00607-024-01331-9

Schneider A, Ingram H. Behavioral assumptions of policy tools. The Journal of Politics. 1990; 52: 510–529.

Shen S, Feng SY, Wu N, Zhao ZX. Retrieval-augmented generation of policy texts based on large language models. Data Analysis and Knowledge Discovery. 2025; 9: 37–48. https://doi.org/10.11925/infotech.2096-3467.2024.0670 (In Chinese)

Sivarajkumar S. Natural Language Processing for Information Extraction and Predictive Modeling in Healthcare [Doctoral dissertation]. University of Pittsburgh: Pittsburgh (PA). 2025.

Zhao J, Li C. Research on the Classification of Policy Instruments Based on BERT Model. Discrete Dynamics in Nature and Society. 2022; 2022: 6123348. https://doi.org/10.1155/2022/6123348

Zhao X, Jia Y, Li A, Jiang R, Song Y. Multi-source knowledge fusion: A survey. World Wide Web: Internet and Web Information Systems. 2020; 23: 2567–2592. https://doi.org/10.1007/s11280-020-00811-0