Creation of the Chinese Adaptive Policy Communication Corpus

Anonymous ACL submission

Abstract

We introduce **CAPC-CG**, the *Chinese Adaptive Policy Communication (Central Government) Corpus*, the first open dataset of Chinese policy directives annotated with a five-color taxonomy of clear and ambiguous language categories, building on Ang's theory of adaptive policy communication. Spanning 1949–2023, this corpus includes national laws, administrative regulations, and ministerial rules issued by Chinese top authorities. Each document is segmented into paragraphs, producing a total of 3.3 million units. Alongside the corpus, we release comprehensive metadata, a two-round labeling framework, and a gold-standard annotation set developed by expert and trained coders. Inter-annotator agreement achieves a Fleiss's kappa of $K = 0.86$ on directive labels, indicating high reliability for supervised modeling. We provide baseline classification results with several large language models (LLMs), together with our annotation codebook, and describe patterns from the dataset. This release aims to support downstream tasks and multilingual NLP research in policy communication.

1 Introduction

China's one-party political system is politically centralized yet administratively decentralized (Landry, 2008; Lieberthal and Oksenberg, 1988a; Ang, 2016). The Communist Party and the State Council communicate instructions through written policy directives, which local government actors then interpret and implement. These directives—including National Laws, Administrative Regulations, and Ministerial Rules—function like the formal nervous system, sending signals from central to local authorities across policy domains and regions.

As Ang (2016, 2023) demonstrates, these directives operate along a spectrum of clarity and ambiguity. Some directives clearly authorize or forbid, while others deliberately employ vague or flexible language to allow local experimentation and adaptation. Capturing this adaptive logic requires systematic, large-scale annotation, yet no open dataset has previously existed. In this paper, we present the first such resource: **the Chinese Adaptive Policy Communication Corpus (CAPC-CG)**,[1] an open corpus of central-government policy directives spanning 1949–2023.

Inspired by expert-annotated corpus construction such as OPP-115 (Wilson et al., 2016), **CAPC-CG** provides annotations based on a five-color taxonomy of directive clarity and ambiguity. The corpus is released together with code, guidelines, and baseline models to support reproducible NLP research.

Building on Ang's original taxonomy of three directive types (*Red* = *prohibiting*, *Black* = *authorizing*, *Grey* = *ambiguous*), we extend her schema into five categories by adding *Yellow (mandating with high pressure)* and *Charcoal (clearly urges flexible action)*. This expanded schema is designed to be computationally codable while remaining faithful to the theoretical framework. We trained Chinese-proficient researchers to annotate a gold-standard training dataset and fine-tuned Large Language Models (LLMs) to scale annotation across the whole corpus. Inter-annotator agreement for directive typing reaches $\kappa = 0.86$, indicating reliable labels despite the complexity of these directives (Bhowmick et al., 2008; Richie et al., 2022). Our contributions are threefold:

1. A large-scale, computable corpus of Chinese central government issued policy directives,with normalized metadata, revision handling to remove duplicates, and paragraph-level segmentation. This resource addresses the acute data scarcity in Chinese policy corpora and enables multilingual NLP research at scale.

[1] Dataset and Scripts

2. We introduce a five-color schema grounded in expert knowledge of Chinese bureaucracy and policymaking. Based on it, we perform multi-dimensional annotation, including topic labels, named entities, and attributes, yielding a gold-standard set of about 6,000 paragraphs with high inter-annotator agreement.

3. We provide evaluation splits, scripts, and baseline models using LLMs, together with a hybrid human–LLM pipeline that scales labeling beyond the gold-standard set. All data, codes, and guidelines are released under a permissive license, establishing **CAPC-CG** as a standardized testbed for diachronic analysis, cross-regime comparisons, information extraction, and downstream applications such as compliance monitoring and policy-impact assessment.

In short, **CAPC-CG** integrates adaptive communication theory from political science with innovative NLP methods. Figure 1 illustrates our workflow for creating this corpus.

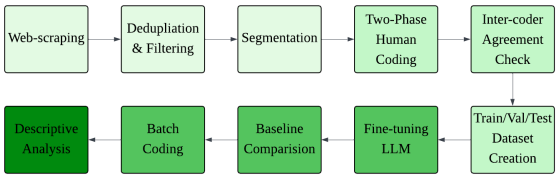

Figure 1: Overall Workflow

2 Theoretical Foundation: Adaptive Policy Communication

Although several large-scale Chinese datasets have been developed, they are mainly based on Internet and social media content without much quality control (Wang et al., 2025a; Li et al., 2024; Zhang et al., 2023; Yuan et al., 2023, 2021; Xu et al., 2020). In contrast, resources for political and legal documents remain scarce: existing datasets lack the high-quality raw texts and rigorous annotations to trace policymaking at the national level (Ahrens et al., 2018; Xiao et al., 2018; Yao et al., 2022; Duan et al., 2019; Wang et al., 2025b; Li et al., 2025; Xiao et al., 2021). This noticeable gap significantly constrains computationally linguistic, legal, and politically scientific research on China's policymaking, legal analysis, and policy NLP research.

Among scholars of Chinese bureaucracy, the long-standing impression has been that "the higher levels give vague directives" (Lieberthal and Oksenberg, 1988a, p. 340) or pronounce "catchy slogans that... direct the flow of policy implementation at all levels" (Brandt and Rawski, 2008, p. 16). Yet this conventional view raises a puzzle: How can a vast country like China, facing immense challenges and numerous fine-grained tasks, be effectively governed through slogans and vague expressions from the top?

Ang (2016) advanced a revisionist account she calls *adaptive policy communication*. In her book, *How China Escaped the Poverty Trap*, she characterizes post-1979 Chinese governance as a system of "directed improvisation," in which central authorities signal and steer rather than dictate, while local governments improvise, producing diverse paths of policy implementation and development. A key mechanism of "directing" is that central authorities deliberately mix ambiguous and clear mandates.

More precisely, she highlights three varieties of directives: *Black* (clearly states what can be done), *Red* (clearly states what cannot be done), and *Grey* (ambiguous about what can or cannot be done). Black authorizes, Red prohibits, while Grey permits flexibility. This categorical mix enables the center to alternate between granting discretion and enforcing discipline. Earlier on, Ang (2023) made an initial attempt at operationalizing these three categories using a pilot sample of central regulations. In this project, we refine classification and annotation methods using a hybrid of LLMs and trained human coders, and scale the analysis to a much larger dataset.

To achieve precision, we segment the vast corpus of highly variegated and lengthy bureaucratic texts into smaller, analyzable paragraphs, ultimately producing a gold-standard, expert-labeled data for NLP research. We also share a rigorous and replicable workflow that combines raw data and human annotation, helping mitigate bias and inaccuracy in LLM applications (Baumann et al., 2025).

3 Corpus Creation and Structure

This section details how we created **CAPC-CG**, from our document collection, annotation schema, to the resulting metadata-rich dataset. Figure 2 illustrates our workflow step by step.

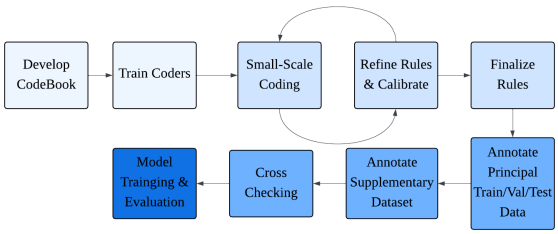

Figure 2: Workflow for Corpus Creation

Year	1949-2023
Policy Documents	337,056
Policy Paragraphs	3,275,477
Annotated Data Practices	2,647,695
Mean TextLength	162.50
Median TextLength	108.00

Table 1: Summary Statistics

3.1 Scope of Policy Directives

Our analysis focuses on policy directives, defined as formally issued documents in which a higher-level authority addresses state or public-sector actors with actionable implications—that is, the bureaucrats are expected to notice, interpret, and respond to the directives. Such directives form the "vertical formal channels" of bureaucratic communication, as Oksenberg (1974) noted. Central directives define national priorities and directions, then as they travel down the hierarchy, each level of government further interprets and details their implementation of the instructions.

To construct our dataset, we collected three main type of central directives. 1) **National Laws** and decisions promulgated by the National People's Congress (NPC), the legislative branch. Among the directives, these laws carry the highest level of authority. 2) **Administrative Regulations**, which are typically issued by the State Council (the highest office of the executive branch), or the General Office of the Communist Party Central Committee, or are co-issued by the State Council and relevant ministries. 3) **Ministerial Rules** issued by central-level ministries and commissions.

This dataset spans 1 October 1949 to 13 March 2023, that is, from the founding of the People's Republic of China through the end of President Xi Jinping's second term. It comprehensively includes all the top leaders since founding, from Mao Zedong, Deng Xiaoping, Jiang Zemin, Hu Jintao, to Xi Jinping, who is currently still the General Secretary of the Party and President.

For each document, we parsed and created a standardized metadata record with the following fields: *title*, *validity status*, *hierarchical level of legal force*, *issuing body*, *reference number*, *date of promulgation*, and *date of entry into force*. The product is a massive longitudinal corpus. Spanning multiple leadership eras, it contains 337,056 unique policy documents. Table 1 presents summary statistics of the corpus, including its scale before and after annotation at the paragraph level.

Next, we transformed the raw text into a structure that captures our five-category schema: *Black, Red, Grey, Yellow, Charcoal* (Table 2).[2] Appendix A provides our codebook that details the classification rules.

We segmented all documents into paragraphs using an ensemble of rule-based algorithms and LLM (see the detailed segmentation method in Appendix B). Because the structure of Chinese policy directives is highly variegated, a single passage may include multiple signals, while sometimes a series of paragraphs may correspond to one consistent signal. To tackle this variation, we apply an LLM to segment each document into units bounded by a coherent signal. We assigned annotators to manually validate 4,600 randomly sampled segmented paragraphs and found that 97.74% were correctly segmented. Following this validation, we removed units that consisted of non-actionable statements, which annotators labeled as *Neutral* (N) in the first stage of annotation (see Appendix A). This procedure produced a final set of 2,647,695 paragraphs, which serves as the foundation for analysis.

3.2 Annotation Scheme and Process

We created a gold-standard dataset through a rigorous, multi-stage annotation procedure carried out by three domain experts (Marcinkiewicz, 1994; Tseng et al., 2020), who annotated paragraph-level texts over a three-month training period. As preparation, all three coders read a diverse sample of policy directives carefully and analyzed representative cases of the five categories listed in Table 2. Then, we organized a workshop dedicated to explaining and testing out the codebook. Every week, the annotators coded small samples of roughly 500 paragraphs, compared their annotation results, then collectively discussed the results and refined their

[2]Colors are employed solely as mnemonic devices; the analytical model encodes the abstract categories themselves.

Colour	Governmental Intent	Signal to Local Actors	Example
Red (R)	Prohibits	"No, you *cannot*."	No agency shall intercept or misappropriate the subsidy funds for any reason or in any form.
Grey (G)	Ambiguous	"You *may* or *may not*."	This matter will be left to 'relevant departments' to weigh the pros and cons and decide for themselves.
Charcoal (C)	Urges flexible means	"Yes, do it *flexibly*."	Actively promote this technology through policy experiments and tailoring to local contexts.
Black (B)	Authorizes	"Yes, you *can*."	The government encourages, supports, and steers the development of the non-public sector.
Yellow (Y)	Pressures or mandates	"Yes, you *must*."	This target must be achieved by any means necessary and will determine bureaucratic performance.

Table 2: Five-Category Taxonomy of Policy Directives.

interpretation of the coding rules. By the end of the training period, our annotators reached $K = 0.864$ of inter-coder reliability. Figure 3 shows inter-coder agreement across three rounds of coding. While the expected agreement under random sampling remained largely unchanged, Fleiss' Kappa improved substantially from 0.670 to 0.864 after iterative training and calibration, indicating that the three annotators achieved high reliability and that our dataset is robust.

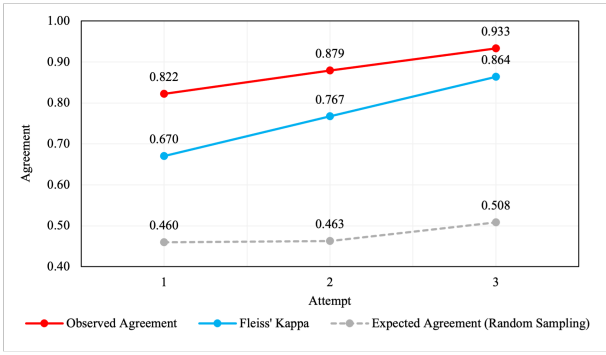

Figure 3: Inter-Coder Agreement Improvement

Once we felt confident about the annotators' grasp of the codebook, they proceeded to code paragraphs at scale. Each week, we assigned 1,000 randomly sampled paragraphs and repeated this process without sample replacement. By the end of the period, they annotated 6,000 samples, ensuring an even distribution across all labels.

Two-Round Labeling Workflow Each paragraph is processed through a two-step classification procedure (Figure 4):

Step 1. Level-1 screening — Annotators categorize the paragraph into one of three high-level labels: *W*, *R*, or *N*. *W* represents all affirmative directives, including *B*, *C*, *G*, and *Y*; *R* represents directives indicating prohibition; while *N* is used for paragraphs that contain no directives, consisting solely of descriptive or contextual information.

Step 2. Level-2 refinement — If a paragraph is labeled as *W* at Level-1, Annotators will specify the exact color using one of the following codes: *B*, *C*, *G*, or *Y*. The paragraph is labeled as *B* when it demonstrates authorization, *C* when it encourages flexible means to implement, *G* when the directive is ambiguous and open to interpretation, and *Y* when it demands immediate action with a sense of urgency or coercion. If directives cannot be classified into any of the four specified colors, the paragraph will be labeled as *U* to indicate uncertainty.

Label *U* We reserve *U* for genuinely unresolved cases after applying the above rules. *U* works as a last resort and will only be used when necessary.

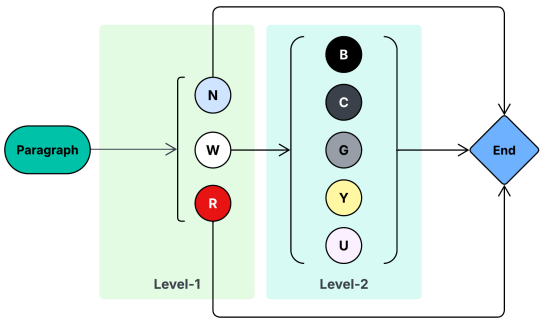

Figure 4: Two-Round Labeling Workflow

This rigorous, two-stage protocol balances theoretical nuance with operational reliability, enabling

the creation of a high-quality training set for subsequent NLP models.

4 Composition of the CAPC-CG

The **CAPC-GC** corpus is organized into a relational database designed to facilitate multi-level analysis. It consists of a two-level model implemented across three pairs of tables, corresponding to Laws, Regulations, and Rules.

Each pair comprises a `documents_*` table and a corresponding `paragraphs_*` table. The `documents_*` tables serve as a comprehensive repository for document-level metadata. Each row corresponds to a unique policy document, identified by a primary key, `DocumentID`, and stores essential fields such as the document's `Title`, `IssuingDepartment`, and `IssueDate`. The `paragraphs_*` tables contain the core units of analysis. Each row represents a single paragraph, identified by its own unique primary key, `ID`, and stores the paragraph's raw text (`Paragraph_content`), its sequential order, and a rich set of annotations, including the final directive classification (`Label_Color`) and Named Entity Recognition (NER) results.

Within each pair, the two tables are linked via the `DocumentID` field. In the `paragraphs_*` table, `DocumentID` acts as a foreign key that references the primary key of its corresponding `documents_*` table. This establishes a clear one-to-many relationship, allowing researchers to seamlessly query individual directives within the context of their original source document. The complete schema is detailed in Table 3.

5 Prediction of Directives Color

While our manual annotation yielded highly accurate data, it is not feasible to scale this method to more than 3.3 million documents. For datasets of immense size, it is essential to automate annotation and classification (Tan et al., 2024). A few recent studies(Sun et al., 2025) attempted to annotate data using LLM APIs with prompt-based instructions; however, existing studies and our own experiments demonstrate that this approach can be highly unreliable (Dong et al., 2022; Baumann et al., 2025). In our experience, even three researchers with knowledge of Chinese policymaking required a rigorously designed codebook and extensive training to achieve a satisfactory level of agreement (Artstein et al., 2008). After comparing multiple strategies, we find that fine-tuning LLMs on a carefully constructed, human-annotated gold-standard dataset provides the most reliable and efficient solution. In this section, we present our experiments on policy classification using LLMs under different settings, and report baseline results that both validate the reliability of our dataset and establish a benchmark for future research. This approach offers a cost-effective, scalable, and robust pathway for applying LLMs to similar tasks in computational social science.

5.1 Experimental Setup

Our aim was to create an automated pipeline for classifying policy paragraphs according to our two-level annotation schema. All experiments were conducted using our human-annotated gold-standard dataset, which contains 6,000 samples, divided into training (80%) and validation (20%) sets. For a final evaluation, we used a separate test set of 1,900 samples, which was collaboratively constructed and validated by our three domain experts.

Our experimental process followed four stages. First, we conducted extensive prompt engineering to find the most effective instructions for our classification tasks. The optimal prompt, detailed in Appendix C, was identified through iterative testing on OpenAI models and applied in all subsequent experiments. Second, we established baselines using traditional machine learning models (SVM and XGBoost) and benchmarked a range of state-of-the-art LLMs for the zero-shot Level-1 classification task to identify a model that balances performance and cost-efficiency. Third, using the selected model, we compared zero-shot, few-shot, and fine-tuning methods to determine the most effective strategy (Brown et al., 2020). Finally, we applied this optimal strategy to the more granular Level-2 (Color) classification task.

5.2 Results and Analysis

5.2.1 Level-1 Directive Classification (R/W/N)

To ensure practical deployment under constraints of compute independence, low latency, and budget efficiency, we restrict our comparison to cost-effective models with stable API access. Although very large open-source models can yield higher raw accuracy, their inference cost and latency render them unsuitable for our large-scale pipeline. Within this design space, we evaluate Level-1 (R/W/N) zero-shot performance (Table 4). GPT-4o-mini attains the best overall metrics (Kappa = 0.618; Acc.

Field	Explanation
From `documents_*` tables	
DocumentID	Primary key for the document tables. A unique integer identifier for each policy document.
Title	Official title of the policy document.
Validity	Legal validity status (e.g., effective, repealed).
Content	Full, raw text of the document.
TextLength	Total number of characters in the document.
EffectivenessLevel	Hierarchical level of legal force (e.g., National Law, Administrative Regulation).
IssuingDepartment	Government body or bodies that issued the document.
DocumentNumber	Official issuance number of the document.
IssueDate	Date on which the document was promulgated.
ImplementationDate	Date on which the document took effect.
From `paragraphs_*` tables	
ID	Primary key for the paragraph tables. A unique integer identifier for each paragraph.
DocumentID	Foreign key linking the paragraph to its parent document.
Paragraph_content	Raw text of the specific paragraph.
order	Sequential position of the paragraph within its parent document.
topics	Topic labels assigned to the paragraph.
NER / NER_DICT	Final, cleaned, and validated NER results.
Label_L1	Level-1 label
Label_L2	Level-2 label

Table 3: Schema Description For the Relational Database

= 0.793; M-F1 = 0.692) at substantially lower cost and latency. Balancing throughput, responsiveness, and unit cost, we therefore adopt GPT-4o-mini for subsequent experiments and large-scale inference.

Next, we compared the performance of zero-shot, few-shot (k=3/6), and fine-tuning methods using GPT-4o-mini. The results, presented in Table 5, unequivocally demonstrate the superiority of fine-tuning. Our fine-tuned GPT-4o-mini model achieves a Cohen's Kappa of 0.841 and a Macro-F1 score (Sokolova and Lapalme, 2009) of 0.844, substantially outperforming the best few-shot baseline (Kappa = 0.686). This represents a 22.5% improvement in Kappa, with particularly strong gains on the minority 'R' class, validating our strategy of leveraging the gold-standard dataset for model specialization.

5.2.2 Level-2 Color Classification (B/C/G/Y)

Building on the success of our Level-1 model, we applied the same fine-tuning methodology to the more nuanced four-class Level-2 task, which classifies affirmative directives into specific colors (B, C, G, Y). The fine-tuned GPT-4o-mini model continued to exhibit high performance and reliability on this challenging task. As detailed in Table 6, the model achieved an overall accuracy of 87.5% and a multi-class Cohen's Kappa of 0.833, indicating a strong agreement with our expert annotators. The model performed consistently well across all four categories, demonstrating its capability to capture the subtle distinctions in policy language required for our fine-grained analysis. This result confirms that our fine-tuning approach is robust and effective for both levels of our annotation schema.

Model	Kappa	Acc.	R-F1	W-F1	N-F1	M-F1
XGBoost	0.153	0.487	0.233	0.538	0.488	0.420
SVM	0.175	0.508	0.260	0.539	0.527	0.442
DeepSeek-V3	0.481	0.733	0.438	0.613	0.815	0.622
Qwen3-235B	0.570	0.780	0.561	0.705	0.834	0.700
GPT-4o-mini	**0.618**	**0.793**	**0.482**	**0.721**	**0.873**	**0.692**

Table 4: Performance comparison of traditional ML models and LLMs with Zero-shot on the Level-1 task.

Method	Kappa	Acc.	R-F1	W-F1	N-F1	M-F1
Zero-shot	0.618	0.793	0.482	0.721	0.873	0.692
Few-shot (k=3)	0.686	0.830	0.489	0.786	0.898	0.724
Few-shot (k=6)	0.659	0.816	0.512	0.763	0.883	0.539
Fine-tuning	**0.841**	**0.910**	**0.677**	**0.905**	**0.952**	**0.844**

Table 5: Comparison of methods for Level-1 classification using the **GPT-4o-mini** model. Fine-tuning shows substantial improvement.

6 Data Analysis

Our corpus facilitates a wide range of analyses into the nuances of Chinese policy communication. To

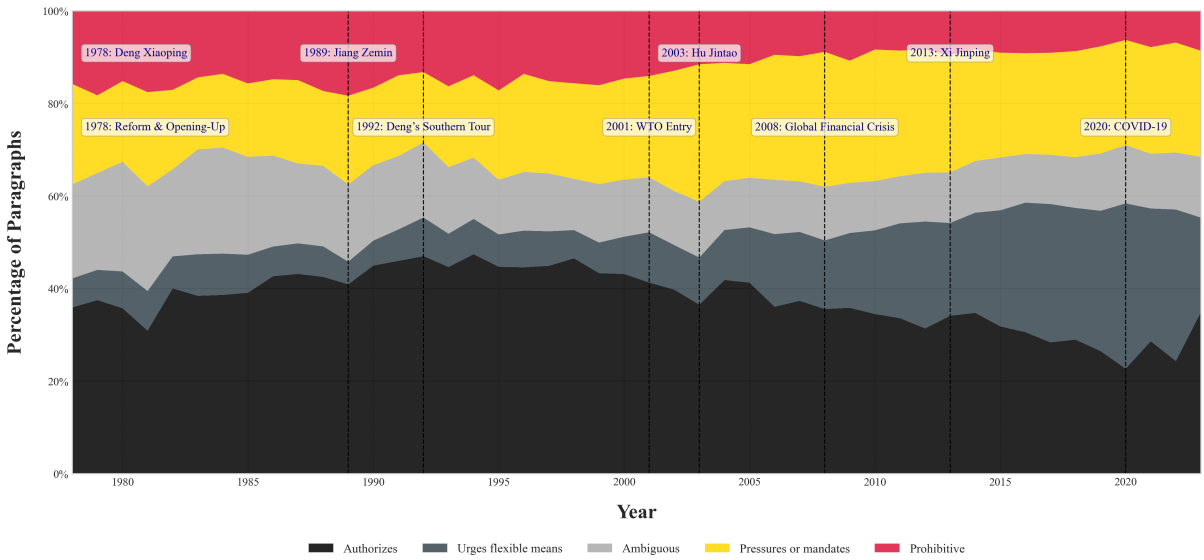

Figure 5: Temporal Trend of Policies (1978-2023)

Metric	Score
Cohen's Kappa (4-class)	0.833
Overall Accuracy	87.5%
Accuracy per Class:	
Class 'B' (Authorizing)	92.0%
Class 'C' (Flexible)	85.0%
Class 'G' (Ambiguous)	84.0%
Class 'Y' (Mandating)	89.0%

Table 6: Performance of the fine-tuned GPT-4o-mini model on the Level-2 (Color) classification task.

demonstrate its utility, we present two preliminary findings.

We begin with temporal evolution. Figure 5 shows that certain patterns have remained quite consistent: for example, *Red* directives make up only a small portion, compared to *Black*—especially after market liberalization policies were introduced by Deng Xiaoping in 1979. Moreover, this pattern appears to be influenced more by major events, such as economic shifts or public health crises, than by leadership transitions.

Figure 6 visualizes the topic distribution derived from our LDA model through heatmap, revealing a strong thematic concentration for certain policy domains (y-axis) across several action-oriented topics (x-axis). Notably, the 'Environmental Protection' corpus is almost entirely dominated by the 'Construction' topic, while the 'Infrastructure' domain is also heavily characterized by 'Construction', as well as 'Development' and 'Management.' This concentration suggests a primary textual focus on tangible, project-based activities within these policy areas. In stark contrast, social policy domains such as 'Educational Equity' and 'Social Security' exhibit negligible weights across all these topics, indicating that their discourse is thematically distinct and likely represented by other latent topics in the model.

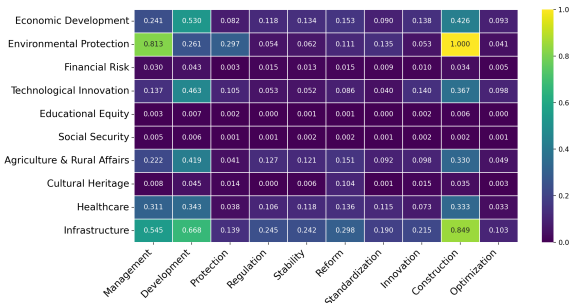

Figure 6: The distribution of action-oriented topics across different policy domains

Figure 7 visualizes the distribution of governmental intents (x-axis), as defined in Table 2, across 17 policy domains. The analysis reveals that specific intents are strongly correlated with particular types of domains. Notably, the intent to "urge flexible means" (C) is almost exclusively dominant in technology-driven domains such as 'Digital Transformation' (0.91) and 'Science & Technology Innovation' (0.91), suggesting a governance model that prioritizes local experimentation.

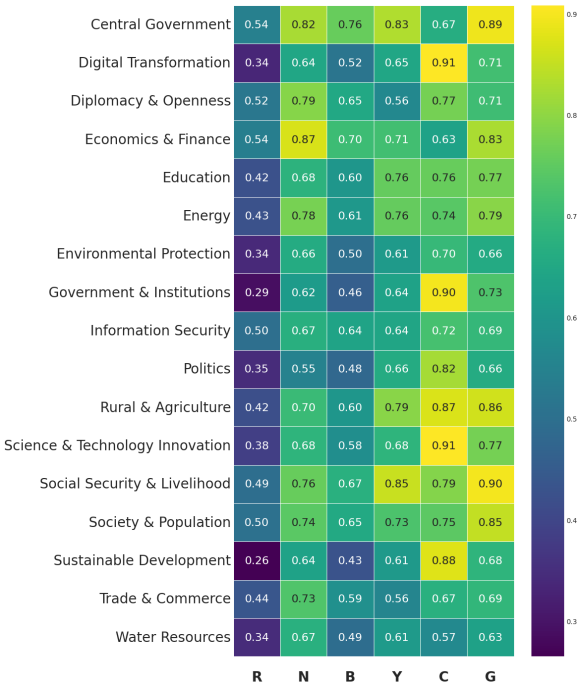

Figure 7: Heatmap of governmental intents across policy domains

7 Future Directions

As a new and foundational resource, our corpus can open up at least three avenues for future research.

7.1 Corpus Expansion and Enrichment

As an essential first step, the **CAPC-CG** corpus covers central-level documents. But future work can collect directives issued by provincial and municipal governments to enable analysis of multi-level policy communication and diffusion. In the future, we also hope to integrate supplementary materials, such as leaders' speeches and state media commentary, to provide a richer context for analyses.

7.2 Advancements in NLP for Legal and Policy Text

Our experiments highlight the limitations of current state-of-the-art LLMs in capturing the nuances of Chinese policy directives. This presents a key challenge for the NLP community: developing models that are not only fluent but also deeply attuned to the subtleties of political language. Future research could explore novel architectures, representation learning techniques, or fine-tuning strategies to model varied policy signals and strategic ambiguity. Our corpus provides an ideal testbed for such innovations through our rigorous annotation.

7.3 Empirical resource for adaptive political economy

Beyond NLP, our dataset serves as a high-resolution instrument to empirically advance the emerging field of "adaptive political economy" (APE) (Ang, 2016, 2023), which studies how actors communicate under uncertainty. APE contrasts conventional principal-agent models that assume clear goals among central authorities, who seek only to control rather than to learn and adapt (Kiewiet and McCubbins, 1991). One surprise in the Chinese context is that even an authoritarian government can sometimes prefer ambiguity and flexibility to strict commands. With our dataset, researchers can ask why and when Chinese central authorities choose to be ambiguous to encourage policy experimentation.

8 Conclusion

This paper introduces and validates **CAPC-CG**, which, to our knowledge, is the most comprehensive paragraph-level corpus of Chinese policy directives for NLP research, built around a novel taxonomy of five nuanced signals. By implementing a theoretically-grounded annotation framework and leveraging fine-tuned Language Models, we created a first-of-its-kind computable dataset.

We hope releasing this dataset will spark studies of policy language in different political systems and the training of advanced models. By bridging political science and NLP, we hope **CAPC-CG** can empower a new generation of research in comparative and adaptive policy communication.

Limitations

We acknowledge the following limitations, which also point to avenues for future research.

First, our corpus is subject to the inherent challenges of archival availability and opacity in contemporary China. Although we have systematically compiled the most comprehensive collection possible from public and official databases, the complete historical record is not fully accessible. Notably, numerous documents from politically sensitive periods, such as the Cultural Revolution (1966-1976), are intentionally withheld from the public domain or remain in undigitized archives. Consequently, while our dataset represents the fullest publicly available record, it cannot capture the entirety of the CCP's internal policy discourse during certain historical junctures. Our analysis should

therefore be understood as pertaining to China's public-facing policy landscape.

Second, our gold-standard dataset was annotated by a small, dedicated team of three domain experts rather than a large crowdsourced workforce. This decision was necessitated by the significant domain knowledge required to interpret Chinese policy language and the complexity of our annotation schema, which demanded an intensive, multi-month training period to ensure reliability. We deliberately prioritized the quality, depth, and consistency of annotations over the sheer quantity of annotators. The high inter-annotator agreement achieved (Fleiss' Kappa = 0.864) validates this expert-driven approach and the reliability of our data, though we acknowledge that a larger pool of annotators could have captured a wider range of interpretations.

Finally, while our experiments establish strong performance baselines using state-of-the-art API-based Large Language Models, they do not provide an exhaustive comparison with a broader range of open-source LLMs and some Transformer-based models. Our primary goal was to construct and publicly release the **CAPC-CG** corpus, and the modeling experiments were designed mainly to illustrate its utility rather than to perform a comprehensive benchmarking exercise. Evaluating a wider array of open-source and fine-tuned models will strengthen our demonstration of the baseline. We hope that our released dataset will serve as a standard testbed for such comparative analyses, supporting more extensive exploration by the computational social science and NLP community.

References

Kathleen Ahrens, Huiheng Zeng, Shun Han Rebekah Wong, and 1 others. 2018. Using a corpus of english and chinese political speeches for metaphor analysis.

Yuen Yuen Ang. 2016. *How China escaped the poverty trap*. Cornell University Press.

Yuen Yuen Ang. 2023. Ambiguity and clarity in china's adaptive policy communication. *The China Quarterly*, 257:20–37.

Artstein, Ron, Massimo Poesio, and Massimo. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825*.

Plaban Kumar Bhowmick, Anupam Basu, and Pabitra Mitra. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65.

Loren Brandt and Thomas G Rawski. 2008. *China's great economic transformation*. Cambridge university press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, and 1 others. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *China national conference on Chinese computational linguistics*, pages 439–451. Springer.

Roderick D. Kiewiet and Mathew McCubbins. 1991. *The Logic of Delegation: Congressional Parties and the Appropriations Process*. University of Chicago Press, Chicago.

Pierre F Landry. 2008. *Decentralized Authoritarianism in China: the Communist Party's control of local elites in the post-Mao era*. Cambridge University Press.

Tianxin Li, Xigang Ke, and Hui Shi. 2025. Topic modeling and evolutionary trends of china's language policy: A lda-arima approach. *PLoS One*, 20(5):e0324644.

Xiaodong Li, Yuwei Zhao, and 1 others. 2024. Pretraining a chinese-centric large language model. *arXiv preprint arXiv:2404.04167*.

Kenneth Lieberthal and Michel Oksenberg. 1988a. *Policy Making in China*, volume 1. Princeton University Press, Princeton, NJ.

Kenneth Lieberthal and Michel Oksenberg. 1988b. *Policy making in China: Leaders, structures, and processes*. Princeton University Press.

Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, 273:31.

Michel Oksenberg. 1974. Methods of communication within the chinese bureaucracy. *The China Quarterly*, 57:1–39.

Russell Richie, Sachin Grover, and Fuchiang Rich Tsui. 2022. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st workshop on biomedical language processing*, pages 275–284.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437.

Bolun Sun, Yifan Zhou, and Haiyun Jiang. 2025. Empowering users in digital privacy management through interactive LLM-based agents. In *The Thirteenth International Conference on Learning Representations*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Tina Tseng, Amanda Stent, and Domenic Maida. 2020. Best practices for managing data annotation projects. *arXiv preprint arXiv:2009.11654*.

Ming Wang, Rui Chen, and 1 others. 2025a. Opencsg chinese corpus: High-quality corpora for chinese large language model training. *arXiv preprint arXiv:2501.08197*.

Xiaoyu Wang, Hong Zhao, Weichong Zhang, and Fang Wang. 2025b. Improving the discourse parsing of chinese government documents: a policy-based reinforcement learning method. *Information Technology and Management*, pages 1–23.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and 1 others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Li Xu, Hai Hu, Xuan Zhang, Lu Li, Chenjie Cao, Junjie Li, and Kai Sun. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *arXiv preprint arXiv:2003.01355*.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. Leven: A large-scale chinese legal event detection dataset. *arXiv preprint arXiv:2203.08556*.

Weizhe Yuan, Zhen Wu, Can Xu, and 1 others. 2023. Wanjuan: A comprehensive multimodal dataset for advancing chinese large models. *arXiv preprint arXiv:2308.10755*.

Yuxiao Yuan, Tianyi Zhang, and 1 others. 2021. Wudaocorpora: A super large-scale chinese corpus for pretraining large models. *AI Open*, 2:65–68.

Wenxuan Zhang, Zhengyuan Liu, and 1 others. 2023. Chinesewebtext: A large-scale high-quality chinese web text dataset. *arXiv preprint arXiv:2311.01149*.

A Codebook and Annotation Process

A.1 Purpose and Scope

To enhance comprehension, this appendix outlines a concise codebook and workflow for annotating *policy directives* in formally issued Chinese government documents.

Document universe We cover three main types of central policy documents spanning 1949 to 2023: (i) National Laws, (ii) Administrative Regulations, and (iii) Ministerial Rules.

A.2 Conceptual Framework

Policy communication mixes clear commands with ambiguous guidance. Following Ang (2023) and Lieberthal and Oksenberg (1988b), we treat vertically issued directives as *signals* along five colors:

- **Red (R)**: Clear prohibition of state actions or state workers' actions; often accompanied by penalties for violations.
- **Black (B)**: Clear authorization, approval, or assignment of responsibility without urgency or ambiguity.
- **Yellow (Y)**: Clear mandate with urgency or top-down pressure; often accompanied by binding performance targets, statutory timeline, or assignment of performance reviews.
- **Charcoal (C)**: Clear mandate with flexible implementation (e.g. experiments and local adaptation).
- **Grey (G)**: Ambiguous/ hedged intents (e.g., delegation, caution, and exploratory proposals).

In addition to this five-color taxonomy, we introduce the label **N** to designate paragraphs that contain no directives for state actions. These paragraphs typically include background information, ideological slogans, private laws and regulations, personnel notices, and transmittal notes.

A.3 Coding Principles

1. **Recipient principle**: All directives should be read from the perspective of state actors. Private laws and regulations directed solely at individuals, enterprises, or social groups without assigning bureaucratic tasks are not considered directives.
2. **Clarity vs. specificity**: Code for clarity of intent, not specificity of detail. A directive may express a clear purpose while remaining vague in its implementation, or contain extensive details yet lack a clear underlying intent.
3. **Avoid keyword-only tagging**: Keywords like *must*, *flexible*, or *forbid* can serve as helpful cues, but they should never be used mechanically or treated as the deciding criteria.

A.4 Step-by-Step Workflow

This section provides a practical workflow for applying the two-round annotation process.

Step 1: Assign Level 1 Label (W / R / N) Determine whether the paragraph belongs to:

- W (White): Directive for actions (B, Y, C, G)
- R (Red): Prohibition
- N (None): No directive

Step 2: If W, identify B / Y / C / G / U If the paragraph clearly fits one category, label it as R, B, Y, C or G. If the paragraph contains mixed signals, apply the **Proportion Method**:

- Break the paragraph into individual sentences
- Assign a color signal to each sentence
- Label the entire paragraph based the color signal with the largest proportion

Step 3: Use U (Unsure) only when necessary

- Assign U only if the paragraph cannot be confidently labeled after applying the first two steps.
- Use U as a last resort. Do not assign U to avoid difficult judgments.

A.5 Data Schema

For each paragraph record, store:

- Document ID and Paragraph ID
- Level 1 label: W/R/N
- Level 2 label: B/Y/C/G/U
- Optional notes (brief rationale/phrase)
- Recommended metadata: document type, issuing unit, receiving unit (if any), abbreviation/number, title, year.

A.6 Intended Use

This codebook undergirds reliable human annotation, supports transparency for external researchers and ML collaborators, and facilitates aggregation for downstream modeling.

B Segmentation Method

This appendix provides a comprehensive technical description of the two-stage computational pipeline developed to segment the corpus of Chinese policy documents. The primary objective of this methodology is to partition each document into discrete, semantically coherent, and atomic units of policy prescription. This process is foundational to our research, as it ensures that subsequent quantitative analyses are based on consistent and replicable data, free from the non-deterministic artifacts of purely generative approaches.

The core challenge this pipeline solves is the inherent unreliability of using an LLM alone for segmentation. A standalone LLM exhibits non-deterministic behavior on hierarchical text, making it unsuitable for research requiring replicable data processing. Our hybrid approach overcomes this by using the LLM solely for structural identification, while a deterministic script performs the final segmentation.

B.1 Corpus Description

The corpus to which this pipeline was applied consists of 337,056 Chinese policy documents issued between 1949 and 2023. The documents fall into three main categories:

- Administrative Regulations: 11,622 documents
- Ministerial Rules: 322,297 documents
- National Laws: 3,137 documents

B.2 Cost Optimization

A central design goal was to mitigate API costs. This was achieved by architecting the pipeline to minimize the expensive data transfer in the LLM's output. To fully appreciate the innovation, we contrast our implemented pipeline with a standard, un-optimized approach.

B.2.1 Hypothetical un-optimized pipeline

A standard approach would involve sending the raw document text to the LLM and instructing it to return the full text with inserted segmentation tags. The process is as follows:

1. **API Call**: Send the entire raw text of document.txt to the LLM.
2. **LLM Task**: The LLM internally processes the text and embeds XML tags (e.g., <L1>...) directly into the content.
3. **LLM Output**: The LLM returns the complete, modified text as its response.

Cost analysis At the project's pricing of \$0.20/1M input and \$0.80/1M output tokens, this model is prohibitively expensive. The output token count is nearly identical to the input count, and output tokens are four times more costly. The total cost for a single document would be approximately: Cost un-optimized $\approx$ (Tokens input $\times$ \$0.20) + (Tokens output $\times$ \$0.80)

Since Tokens output $\approx$ Tokens input, the cost is dominated by the output, making the process five times more expensive than just reading the input.

B.2.2 Implemented optimized pipeline

Our implemented pipeline fundamentally changes the LLM's task from "text editor" to "structural analyst," shifting the workload locally to reduce API costs. The process is as follows:

1. **Local Pre-processing**: A Python script first adds <line #> markers to the document text. This step happens locally and incurs no API cost.
2. **API Call**: Send the numbered text to the GPT-4.1-mini-2025-04-14 model via the OpenAI Batch API.
3. **LLM Task**: The LLM's only task is to identify structural elements and output a compact JSON file listing labels and their corresponding line numbers.
4. **Local Reconstruction**: A local script merges the LLM's JSON output with the original text file to create the XML-tagged document.

Cost analysis This architecture transforms the cost equation. The output is now a very small JSON object, typically only 5-10% of the input token size. The total cost is: Cost optimized $\approx$ (Tokens input $\times$ \$0.20) + (Tokens JSON $\times$ \$0.80)

This workflow nearly eliminates the expensive output token cost, making the entire process vastly more economical.

B.2.3 Results

The cost-optimization strategy was exceptionally effective. The total expenditure for processing the corpus was $294.99. In contrast, the un-optimized pipeline, with its 1:1 input-to-output token ratio, would have incurred an estimated cost of $1,061.30. Our implemented pipeline therefore achieved a verifiable cost reduction of 72.20%. The pipeline was validated through an iterative process of expert review.

B.3 Pipeline Workflow

Figure 8 illustrates the end-to-end process, from raw text file to final segmented output.

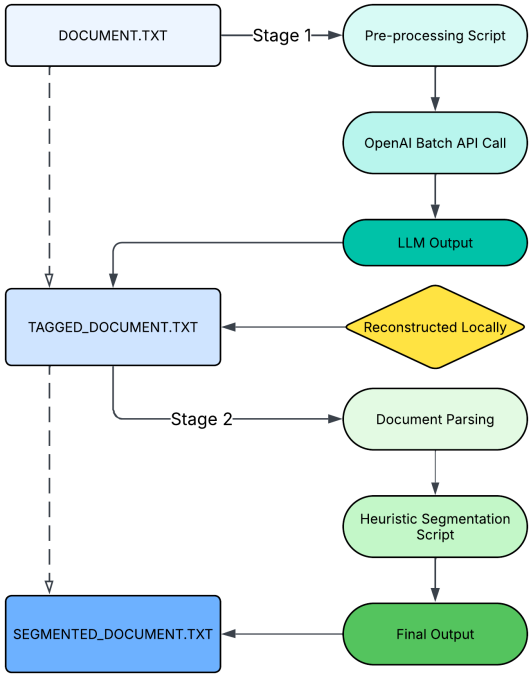

Figure 8: Segmentation Workflow

B.3.1 Stage 1: LLM-Based Structural Tagging

This stage is orchestrated by a master Python script leveraging the OpenAI and tiktoken libraries. It prepares files, submits them to the Batch API, and monitors their progress. The core instruction is contained within the system prompt provided to the GPT-4.1-mini-2025-04-14 model:

您是专门用于对中文法律文件进行结构化分析的AI助理。您的任务是识别文本中的特定段落，并根据其内容和结构进行分类。输入文本的每一行都以<line#>作为前缀。核心目标：隔离层级列表项 (L#标签) 和无意义的样板文本 (NEUTRAL_CONTENT 标签)，同时保留有意义的散文段落不进行标记，以供后续分析。

输出要求：您的输出必须是一个JSON对象。此对象包含一个名为"segments"的键，其值为一个数组。数组中的每个对象都应包含："label": (string) 必须是"L1", "L2", "L3", ..., 或 "NEUTRAL_CONTENT"。"line_numbers": (array of integers) 与该标签对应的原始行号列表。

标记规则：层级标签 (L#): 使用 L1、L2 等标签来标记结构化的、有编号或有字母的列表项。以 第#条 开头的列表项是最高层级，必须标记为L1。

中性内容 (NEUTRAL_CONTENT): 仅用于标记客观的、程序性的、或不具有任何分析意义或情感色彩的样板文本。包括但不限于：文件标题、文首、目录、人名、引用编号、称呼、结尾落款和日期、简单的项目列举。

不标记的内容: 提供解释、推理、条件或背景信息且不属于列表结构的散文段落，被视为“有意义的”。这些段落必须保持不标记。最终指令：严格按照上述规则分析提供的文本，并仅返回一个JSON对象。不要包含任何额外的解释或评论。

B.3.2 Stage 2: Rule-Based Analysis and Segmentation

The second stage uses a local Python script that relies on the BeautifulSoup library with the lxml-xml parser. This stage performs two distinct, sequential tasks: first, it analyzes the document's content to generate metadata, and second, it applies a deterministic ruleset to perform the final segmentation.

1. **Metadata Generation: Heuristic Content Analysis**

   Before segmentation, the script first analyzes the tagged document to determine its most substantively important structural layer. This is achieved by scoring the content within each hierarchical layer (e.g., all L1 tags, all L2 tags) based on a set of heuristic parameters. The goal of this step is to produce a single piece of metadata—the "optimal segmentation layer"—which is saved in the header of the output files for potential use in later analysis.

   This scoring process is governed by tunable parameters in Table 7.

2. **Final Segmentation: Deterministic Rule Application**

Parameter	Value	Rationale
COLOR_DIVERSITY_WEIGHT	3.0	Emphasizes the presence of action-oriented keywords.
LENGTH_WEIGHT	0.5	Favors longer, more substantive text blocks.
LAYER_PENALTY_FACTOR	1.5	Penalizes deeper layers (L3, L4), creating a bias for higher-level structure.
MIN_LENGTH_THRESHOLD	15.0	Minimum character count for a block to be considered score-worthy.
CHINESE_ACTION_WORDS	12.0	Set of keywords indicating policy instructions or legal mandates: "必须", "应", "须", "可以", "可", "建议", "宜", "不得", "禁止", "严禁", "立即", "紧急"

Table 7: Parameters and Weights Description

After the metadata has been generated, a separate function performs the final, deterministic segmentation. This process does not use the heuristic scoring parameters above. Instead, it inserts segment breaks based on the following purely structural rules:

Structural Change A break is inserted when transitioning between a content tag (e.g., <L1>) and a <NEUTRAL_CONTENT> tag. This isolates substantive policy from boilerplate.

Hierarchical Ascent A break is inserted if the hierarchical level of the current tag is the same as or higher than the previous tag (e.g., <L2> to <L2>, or <L3> to <L2>). This rule correctly isolates self-contained articles or major sections.

Punctuation Exception A break is suppressed if the preceding text block ends with a joining punctuation mark (',', '；', ',', ';') and the next tag is at the same level. This sophisticated rule prevents the fragmentation of compound sentences and related clauses.

B.4 Validation and Performance

The segmentation pipeline was validated through a detailed manual review process. We randomly selected a set of documents and compared the results of machine-based segmentation and manual segmentation by two Research Assistants with advanced background knowledge in Chinese politics.

The primary criterion for evaluation was semantic coherence, specifically assessing whether each segment served as a complete and intelligible unit of policy information. Each paragraph was marked as "Yes" for accurate or "No" for inaccurate, allowing for the calculation of accuracy rates.

The results demonstrated impressive accuracy, with 96.7% for Ministerial Rules and 98.25% for National Laws, providing strong evidence of the validation and reliability of our segmentation methodology.

C Prompts for Annotation

C.1 Level-1 (W/R/N) Prompt

角色：你是一名“政策信号一级标注员”。

任务：对给定中文政策文档段落分类标注：

W = 段落中存在对政府/公共部门的肯定性行动指令（包含任何“要/应/负责/建立/加强/推进/组织/指导/监督/落实/制定/完善/开展/实施/安排”等对政府主体布置的任务或授权）。

R = 段落中存在对政府/公共部门的禁止性指令（如“严禁/不得/禁止/停止……（由政府或其部门）”）或对政府及公职人员的违法行为进行处罚。

N = 无指令（仅背景介绍、原则口号、定义释义、程序性条款、日期/效力条款、对非国家主体的要求而未同时对政府布置任务或责任的内容）。

受众原则（极其重要）：只统计指向政府及其附属公共机构（包括国务院各部委、地方政府、监管机构、事业单位等）的行动/禁止为“指令”。

仅限制或要求企业/个人/社会组织，而没有同时给政府安排任务或责任→ 记为 N

若同时出现：例如“企业不得……；主管部门要加强监管……”，则本段视为存在对政府的指令，进入 W 或 R 的判断。

政府/公共部门的定义为中央或地方政府、公职人员或承担了国家分配的任务的附属机构

不要只看关键词：是否出现“必须/可以/鼓

励/不得”等词不等于对应标签，必须结合主语是否为政府/公共机构与句子功能来判断。

例如：“企业必须符合××条件方可办理” → 若仅针对企业且未布置政府任务 → N

输出要求：只输出最终的结果W/N/R

待判断政策内容如下：

«待判断政策段落»

C.2 Level-2 (B/Y/C/G) Prompt

角色：你是一名“政策信号一级标注员”。

任务：对给定的中文政策文档段落进行分类标注, 该段落已经被判定为含有肯定性政策指令，请你进一步判断它属于以下哪一类信号：

B（Black，你可以）：明确同意、批准或中性部署（如“建立/推进/构建/实现”），没有强制性、没有因地制宜的灵活性、也不含糊。

Y（Yellow，你必须）：带有明显压力或紧迫性，要求全面落实或严格执行，常出现“抓紧/尽快/及时/必须/严格执行/全面落实”等；纳入考核、问责、审计也属于 Y。注意：如果“必须”只是作为条件或资格要求（例如“必须符合条件方可申请”），则不算 Y。

C（Charcoal，你应当灵活地做）：指令明确，但强调“因地制宜、多种形式、试点示范、先行先试、借鉴与反馈”等灵活路径或差异化方式。

G（Grey，你可能做也可能不做）：含糊、模糊或保留条件的指令，常见表达包括“在…前提下”“适度”“具备条件的”“可”“鼓励探索”“逐步推进”“协商解决”等，主体或强度不清晰。

输出要求：只输出最终的结果B/Y/C/G

待判断政策内容如下：

«待判断政策段落»