PolicyLLM: Towards Excellent Comprehension of Public Policy for Large Language Models

Han Bao[1], Penghao Zhang[2], Yue Huang[1], Zhengqing Yuan[1], Yanchi Ru[1], Rui Su[1], Yujun Zhou[1], Xiangqi Wang [1], Kehan Guo[1], Nitesh V Chawla[1], Yanfang Ye[1], Xiangliang Zhang[1],
[1]University of Notre Dame, [2]Independent Researcher
Correspondence: hbao@nd.edu

Abstract

Large Language Models (LLMs) are increasingly integrated into real-world decision-making, including in the domain of public policy. Yet, their ability to comprehend and reason about policy-related content remains underexplored. To fill this gap, we present ***PolicyBench***, the first large-scale cross-system benchmark (US-China) evaluating policy comprehension, comprising 21K cases across a broad spectrum of policy areas, capturing the diversity and complexity of real-world governance. Following Bloom's taxonomy, the benchmark assesses three core capabilities: (1) **Memorization**: factual recall of policy knowledge, (2) **Understanding**: conceptual and contextual reasoning, and (3) **Application**: problem-solving in real-life policy scenarios. Building on this benchmark, we further propose ***PolicyMoE***, a domain-specialized Mixture-of-Experts (MoE) model with expert modules aligned to each cognitive level. The proposed models demonstrate stronger performance on application-oriented policy tasks than on memorization or conceptual understanding, and yields the highest accuracy on structured reasoning tasks. Our results reveal key limitations of current LLMs in policy understanding and suggest paths toward more reliable, policy-focused models[1].

1 Introduction

In recent years, Large Language Models (LLMs) (Vaswani et al., 2017; Touvron et al., 2023; Achiam et al., 2023) have achieved remarkable progress, demonstrating intelligent and superior performance in a wide range of natural language processing (NLP) tasks, including machine translation (Zhu et al., 2024), code generation (Svyatkovskiy et al., 2020; Chen et al., 2021), and article writing (Yuan et al., 2022). In parallel with these advancements, a growing body of research has focused on systematically benchmarking LLM capabilities across multiple cognitive dimensions, including language understanding (Wang et al., 2024), reasoning (Xiang, 2023; Cobbe et al., 2021; Joshi et al., 2017), and knowledge acquisition (Yang et al., 2015).

Beyond traditional NLP tasks, LLMs are increasingly being deployed in high-stakes real-world decision-making contexts, such as education (Xiao et al., 2023), law (Fei et al., 2024; Guha et al., 2023; Zhou et al., 2024), healthcare (Tang et al., 2023) and public administration (Pesch, 2025). Among these, public policy stands out as particularly consequential: supporting policy analysis and generation requires not only factual knowledge, but also contextual reasoning and value-sensitive judgment (Hou et al., 2025). Missteps can have tangible social consequences—for example, a model that miscalculates rural funding allocations by relying on the wrong fiscal base may cause substantial under-allocation of resources. Ensuring that LLMs develop a reliable and nuanced understanding of policy content is therefore both a technical necessity and an ethical imperative.

Understanding and applying public policy presents unique challenges for LLMs. While the field's interdisciplinary nature, contextual dependence, and linguistic complexity are well recognized, the central obstacles to advancing policy-aware AI can be framed as a three-tiered problem. 1) ***The Evaluation Challenge: Lack of Rigorous Benchmarks.*** There is currently no comprehensive benchmark to systematically assess the policy comprehension capabilities of LLMs. Without standardized evaluation frameworks, it is difficult to measure performance across skills ranging from factual recall to conceptual reasoning and practical application, hindering objective comparison and targeted improvement. 2) ***The Diagnostic Challenge: Identifying Strengths and Weaknesses.*** Aggregate metrics obscure where models succeed and fail. It remains unclear which cognitive abilities, policy domains, or linguistic contexts pose

[1]Dataset has been released at https://github.com/wad3birch/PolicyLLM.

Findings of the Association for Computational Linguistics: ACL 2026, pages 2249–2274
July 2-7, 2026

the greatest difficulties. Fine-grained diagnostic analysis is therefore essential to pinpoint strengths and weaknesses. 3) ***The Adaptation Challenge: Developing Specialized Models.*** General-purpose LLMs often struggle with the distinct demands of policy tasks. A key challenge is how to adapt existing architectures to better handle the multifaceted requirements of policy analysis, thereby closing the gaps revealed through rigorous evaluation and diagnosis.

To rigorously evaluate the gap, we present *PolicyBench*, a cross-system benchmark (US-China) specifically designed to assess LLMs' understanding of public policy in both China and the United States. *PolicyBench* encompasses a broad spectrum of policy domains and features meticulously crafted questions targeting three cognitive levels: memorization, understanding, and application. Through extensive experiments, we find that model performance improves steadily from memorization to application tasks, with LLMs showing particular strength in structured reasoning scenarios such as numerical calculation and scenario-based decision-making, while still facing challenges in abstract or ambiguous policy contexts and in handling Chinese policy texts. To further enhance LLMs' policy-related reasoning, we propose *PolicyMoE*—a MoE model (Jacobs et al., 1991; Jordan and Jacobs, 1994) trained on policy-focused data (Kang et al., 2024). *PolicyMoE* integrates three specialized expert models, each excelling in distinct capabilities. Experimental results demonstrate that *PolicyMoE* significantly outperforms general-purpose LLMs on policy tasks.

Overall, our main contributions are as follows:

- ▷ We construct *PolicyBench*, a comprehensive cross-system benchmark for evaluating LLMs' policy understanding across diverse domains—in both Chinese and US contexts.
- ▷ Through extensive experiments and human evaluation on *PolicyBench*, we uncover key findings on the strengths and limitations of LLMs in cross-system policy understanding.
- ▷ We propose *PolicyMoE*, an MoE model fine-tuned on *PolicyBench*, which achieves superior performance over strong baselines and underscores the potential of domain-adaptive pretraining for governance-related tasks.

2 The *PolicyBench*

In this section, we provide a detailed introduction to the design and construction principles of *PolicyBench*. To construct our benchmark, **we focused on two of the world's most significant yet distinct policy environments: mainland China (CN) and the United States federal government (US)**. This deliberate selection provides a high-contrast, cross-system testbed for evaluating an LLM's core policy comprehension capabilities across different governance systems. While we acknowledge this does not encompass the full global policy frameworks, it establishes a critical and challenging baseline for this foundational area.

2.1 Policy Acquisition

In the process of policy collection, we initially gathered a broad set of Chinese and US policy documents and related materials. To ensure relevance and timeliness of the content, we applied a filtering process that removed outdated policies, duplicate entries, and documents not related to substantive policy content (*e.g., purely procedural notices or administrative logistics*), details in Appendix A. After this curation step, we retained **721** Chinese policies and **1,890** supplementary Chinese policy materials (*e.g., official commentaries, media news, expert interviews, and public consultations*), as well as **603** US policies and **1,082** supplementary US materials:

- For Chinese policies, all documents were sourced exclusively from the Policy Document Repository of the State Council of China.[2] To categorize the policies, we first followed the organizational structure of the State Council, then refined it to better reflect the content distribution within the corpus. Based on this adapted structure, we grouped the policies into eight domains (*e.g., Public safety*). To retrieve relevant materials, we selected representative search terms—such as "Belt and Road Initiative" and "Double Reduction Policy"—based on "Hot Words" highlighted by major official media and platforms, see Figure 6 for details. In addition, we collected supplementary materials including official interpretations, policy outcomes, and expert interviews from extended social media sources.

[2] https://www.gov.cn/zhengce/zhengcewenjianku/

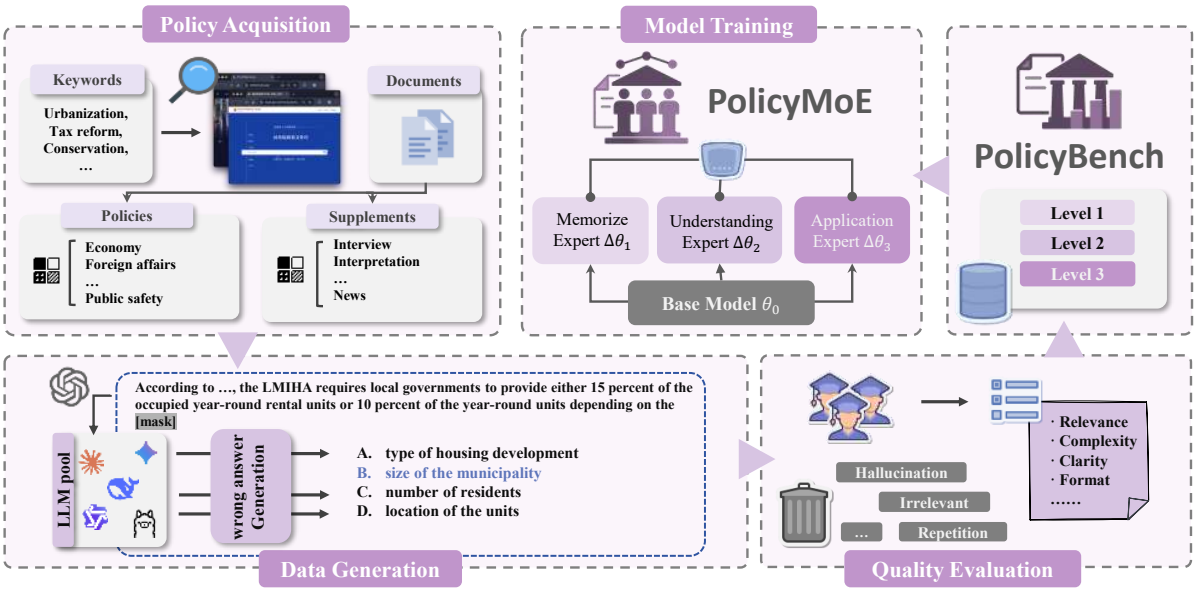

Figure 1: Three levels of evaluating LLM in *PolicyBench*.

- For US policies: As there is no centralized repository for federal policies in the US, we collected policy documents from the official websites of 12 US federal departments (Details in Table 5). Supplementary materials were gathered from authoritative news outlets such as *Reuters*, *Fox News*, etc.

All collected policies fall within the timeframe of 2000 to January 2025, with a primary focus on the most recent decade. All operations fully complied with ethical standards, and all data were lawfully sourced from open-access public databases.

Table 1: Task list of *PolicyBench* and respective numbers ("Mem" = Memorization, "Und" = Understanding, "App" = Application).

Level	ID	Task	Number	
			CN	US
Mem.	1-1	Article/Date Memorization	4,498	3,351
	1-2	Terminology Recognition	237	126
	1-3	Organization Identification	268	228
Und.	2-1	Idea Understanding	1,219	1,013
	2-2	Interest Understanding	1,145	996
	2-3	Institution Understanding	1,620	1,484
App.	3-1	Policy-Based Numerical Reasoning	918	452
	3-2	Scenario-Based Decision-Making	77	179
	3-3	Procedural/Institutional Implementation	320	391
	3-4	Policy Logic and Value Explanation	1,199	1,214

2.2 Dataset Curation

To facilitate fine-grained analysis of model capabilities, we categorize the benchmark into **10 task types**, each targeting a distinct subskill relevant to public policy comprehension. This taxonomy enables a comprehensive assessment of LLMs across a wide range of cognitive and policy domains.

The categorization is structured around a three-level hierarchy informed by Bloom's Taxonomy of Educational Objectives (Krathwohl, 2002), a widely recognized framework in cognitive and educational psychology. Bloom's taxonomy organizes cognitive skills from basic factual recall to deeper understanding and the application of knowledge in real-world contexts. Drawing on this structure, our benchmark defines three assessment levels:

- **Level 1: Memorization.** This level focuses on factual recall, such as memorizing publication dates, institutional actors, specific provisions, or technical terminology. Tasks at this level require minimal inference and primarily test a model's ability to retrieve explicit information from policy documents.
- **Level 2: Understanding.** At this tier, we move beyond literal recall to examine a model's capacity for conceptual understanding and contextualization. Guided by the **3I framework** from policy studies, which emphasizes Ideas, Interests, and Institutions (Hall and Taylor, 1996). Level 2 tasks probe how well a model can interpret underlying motivations, identify key stakeholders, and comprehend institutional logic within policies.
- **Level 3: Application.** The highest level assesses the model's ability to apply policy knowledge in practical scenarios. Tasks at this tier require reasoning about hypothetical or real-world situations, evaluating implications, and suggesting appropriate actions based on policy content.

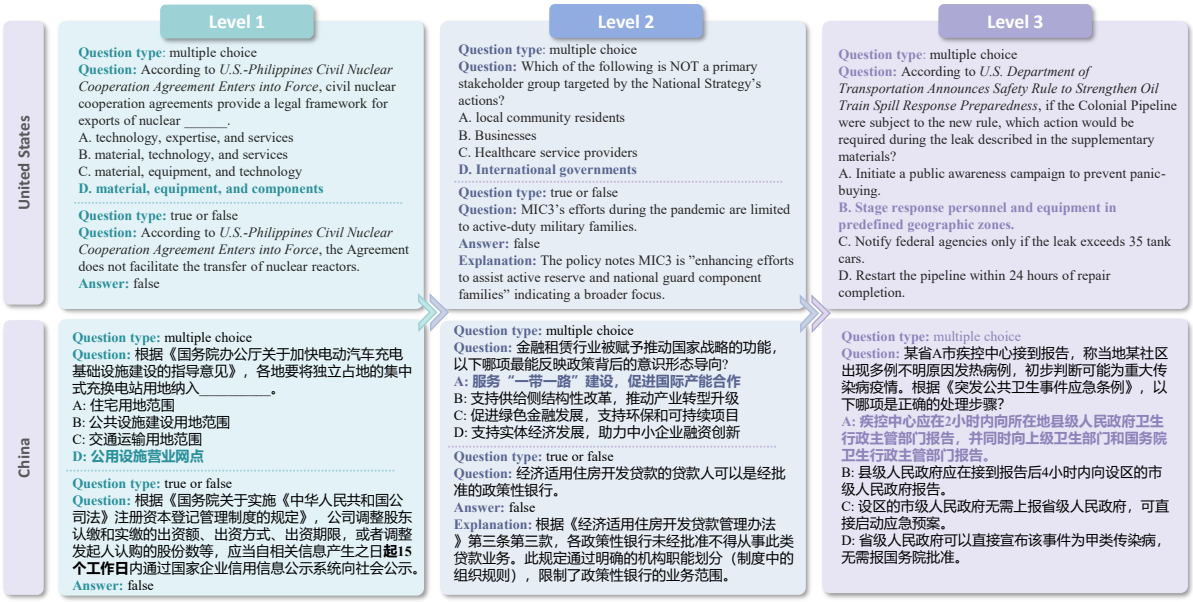

Figure 2: Selected examples from PolicyBench spanning three levels and two languages.

As summarized in Table 1, this hierarchical task design enables a fine-grained evaluation of policy comprehension across multiple cognitive levels. Annotation guidelines and quality control procedures detailed in Appendix E, with formal task definitions are provided in Appendix F.

2.3 Detail of Distractor Generation

To mitigate the bias that can arise when a single LLM generates all distractors, we employ a heterogeneous *model pool* and harvest incorrect options in an iterative fashion. With details mathmatically referred in Algorithm 1 and given a stem–answer pair $\langle q,\ a_{\text{gold}} \rangle$, we initially sample a model m from the pool and prompt it with the original question while *explicitly labelling* a_{gold} as a prior incorrect answer; the model is instructed to propose a new, plausible response that must differ from a_{gold}. If the resulting candidate d is neither redundant nor equal to the gold answer, it is added to the distractor set D. We iteratively resample additional models and repeat the procedure, supplying both the original question and the accumulated distractors, until the total number of distractors satisfies $|D| = k-1$. Finally, we randomly permute $a_{\text{gold}} \cup D$ to form the list of k options.

3 The *PolicyMoE*

Specifically in policy domains, we propose *PolicyMoE*, a method that transforms the overall LLM into a compositional and modular system of experts with different expertise, drawing inspiration from (Kang et al., 2024). In this section, we present the details of *PolicyMoE*. By dividing the policy domain expertise into three specialized expert modules—Memory Policy, Understanding Policy, and Apply Policy—*PolicyMoE* creates a comprehensive system capable of handling diverse policy-related tasks with enhanced precision and efficiency.

3.1 Architecture Overview.

PolicyMoE follows the core principles of the MoE framework while specializing in policy domains:

- Expert Modules: Three dedicated expert models trained on specific policy-related capabilities:
  - Memory Expert: Specializes in recalling policy facts, regulations, historical precedents, and exact policy language
  - Understanding Expert: Focuses on interpreting policy intent, analyzing implications, and explaining policy rationales.
  - Application Expert: Excels at applying policies to specific scenarios, predicting outcomes, and recommending implementation strategies.
- Intelligent Router: A simple linear layer that is shared across all LoRA adapters, which efficiently analyzes input features to determine the most relevant policy domain expertise.

3.2 Constructing Expert Modules

Specialization with LoRA: Each expert module is specialized using Low-Rank Adaptation (LoRA)

Algorithm 1: Distractor Generation for Multiple-Choice Question Construction

Input: Question q, Correct Answer a_{gold}, LLM pool $\mathcal{M}$, Target number of choices $k = 4$
Output: Multiple-choice question with one correct answer and $k-1$ distractors
Initialize distractor set $\mathcal{D} \leftarrow \emptyset$;
while $|\mathcal{D}| < k-1$ **do**
  Sample a model $m \sim \mathcal{M}$;
  Construct prompt:;
    Include the question q;
    Provide a_{gold} as a previous incorrect answer;
    Instruct model m to generate a new plausible answer that is **also incorrect**;
  Generate distractor candidate $d \leftarrow m(q, a_{\text{gold}}$ marked as incorrect);
  if $d \notin \mathcal{D}$ *and* $d \neq a_{gold}$ **then**
    Add d to $\mathcal{D}$;
Randomly shuffle $a_{\text{gold}} \cup \mathcal{D}$ to form final options;
return $\{q, \text{options}, a_{\text{gold}}\}$;

Algorithm 2: Inference Procedure of *PolicyMoE*

Input: Input instruction x
Output: Final model response y
Step 1: Expert Modules Initialization
for *each expert* $i \in \{$*Memory, Understanding, Application*$\}$ **do**
  Load LoRA adapter $\Delta\Theta_i$ into base model Θ_0;
  Construct expert model $\Theta_{\text{spec},i} = \Theta_0 + \Delta\Theta_i$;
Step 2: Routing Decision
Compute routing score vector: $\mathbf{s} = \theta_r x$;
Compute expert weights: $\boldsymbol{\alpha} = \text{softmax}(\mathbf{s})$;
Select top-1 expert index: $i^* = \arg\max(\boldsymbol{\alpha})$;
Step 3: Expert Inference
Use selected expert Θ_{spec,i^*} to generate response:
$y = \texttt{LM}_{\Theta_{\text{spec},i^*}}(x)$;
return y

(Hu et al., 2022), which introduces lightweight, trainable parameters specific to each policy domain while keeping the base LLM intact. The specialized model $\Theta_{\text{spec},i}$ for each domain i is defined as:

$$\Theta_{\text{spec},i} = \Theta_0 + \Delta\Theta_i$$

where $\Delta\Theta_i$ represents the LoRA parameters for domain i. The forward pass for each expert module is:

$$h_i = \theta_0 x + \theta_{B_i}\theta_{A_i} x$$

Here, $\theta_{B_i} \in \mathbb{R}^{d\times\text{rank}}$ and $\theta_{A_i} \in \mathbb{R}^{\text{rank}\times k}$, with $\text{rank} \ll \min(d, k)$.

3.3 Dynamic Integration of Experts

Routing Mechanism: A router module θ_r is introduced to analyze each input token and route it to the most appropriate expert module. The output h for each input x is computed by combining the contributions of the selected expert modules, weighted by their relevance:

$$h = \theta_0 x + \sum_{i=1}^{n} \alpha_i \Delta\theta_i x$$

where α represents the weights computed by the router:

$$\alpha = \text{top-k}(\text{softmax}(\theta_r x))$$

The router is trained using the aggregated synthetic data $D = \{D_i\}_{i=1}^{n}$ to learn optimal module selection for a given task:

$$\mathcal{L}(\theta_r) = -\mathbb{E}_{(x,y)\sim D}\left[\log P_{\Theta_0}(y|x; \theta_r, \{\Delta\Theta_i\}_{i=1}^{n})\right]$$

4 Experiments

4.1 Setup

Models. We select 11 representative state-of-art models: `GPT-4o` (Achiam et al., 2023), `o4-mini`, `Claude-3.7-Sonnet` (Anthropic, 2025), `Claude-3.5-sonnet` (Anthropic, 2024), `Gemini-2.5-Flash` (Google, 2025), `Gemini-2.0-Flash` (Google, 2024), and the open-source models: `Gemma-3-27B` (Team et al., 2025), `Qwen-QwQ-32B` (QwQ, 2024), `Llama-4` (Meta, 2025), `Deepseek-V3` (Liu et al., 2024) and `Deepseek-R1` (Guo et al., 2025), details in Table 6.

Scoring Mechanism. *PolicyBench* adopts a level-aware scoring framework tailored to question type and format. **Levels 1–2** consist exclusively of multiple-choice and true/false questions, which are evaluated using standard accuracy: $\text{Score} = \frac{\text{\# Correct Answers}}{\text{\# Total Questions}}$. **Level 3** includes both objective (multiple-choice, true/false) and subjective

Table 2: Performance (accuracy) of all models on different levels and regions. Gemini-2.5, Gemini-2.0, Claude-3.5 and Claude-3.7 denote Gemini-2.5-Flash, Gemini-2.0-Flash, Claude-3.5-Sonnet and Claude-3.7-sonnet respectively. Red and blue represent the highest and lowest scores in each row respectively.

Level	Region	GPT-4o	o4-mini	Gemini-2.5	Gemini-2.0	Claude-3.7	Claude-3.5	LLaMA-4	Gemma-3-27B	QwQ-32B	Deepseek-V3	Deepseek-R1
Level 1	CN	46.01%	45.93%	54.06%	47.87%	55.29%	53.77%	49.81%	41.75%	55.87%	48.61%	62.02%
	US	52.69%	54.90%	57.73%	53.71%	58.68%	58.76%	52.55%	49.91%	46.40%	50.12%	59.33%
Level 2	CN	56.34%	55.81%	60.57%	56.39%	60.47%	59.74%	56.56%	55.56%	59.79%	55.51%	62.92%
	US	63.40%	64.71%	64.91%	62.25%	68.23%	68.95%	61.17%	62.17%	57.71%	58.62%	65.37%
Level 3	CN	70.24%	79.49%	76.18%	73.80%	73.82%	72.83%	68.54%	71.51%	80.34%	72.33%	73.78%
	US	68.13%	77.00%	69.44%	66.55%	68.28%	68.47%	66.41%	68.37%	69.90%	69.39%	74.60%
AVERAGE		59.47 %	62.97%	63.82%	60.10%	64.13%	63.75%	59.17%	58.21%	61.67%	59.10%	66.34%

Table 3: Average accuracy (%) of all models across Chinese and US (red and blue represent the highest and lowest in each row).

Region	GPT-4o	o4-mini	Gemini-2.5	Gemini-2.0	Claude-3.7	Claude-3.5	LLaMA-4	Gemma-3-27B	QwQ-32B	Deepseek-V3	Deepseek-R1
Chinese	57.53%	60.41%	63.60%	59.35%	63.19%	62.11%	58.30%	56.27%	65.33%	58.82%	62.24%
US	61.41%	65.54%	64.03%	60.84%	65.06%	65.39%	60.04%	60.15%	58.00%	59.38%	66.43%
Average	59.47%	62.98%	63.82%	60.10%	64.13%	63.75%	59.17%	58.21%	61.67%	59.10%	64.34%

(open-ended) questions. Open-ended responses are scored on a 0–5 scale based on alignment with reference answers. To ensure evaluation consistency, we adopt the **LLM-as-a-Judge** (Zheng et al., 2023): for each open-ended question, **two models are randomly sampled** from a pool of four state-of-the-art LLMs: `o4-mini`, `gemini-2.5-flash`, `claude-3.7-sonnet`, and `Deepseek-R1`, to serve as automated graders. Each grader compares the response to a reference answer and assigns a score according to predefined criteria. The final score for that question is computed as the average of the two model scores. We analyze the potential bias in Appendix G.

The overall Level 3 score is calculated as a weighted average across all question types:

$$\text{Score} = \frac{S_{\text{mc}} + S_{\text{tf}} + S_{\text{oe}}}{T_{\text{mc}} + T_{\text{tf}} + 5 \times T_{\text{oe}}},$$

where $S_{\text{mc}}, S_{\text{tf}}, S_{\text{oe}}$ denote the cumulative scores for multiple-choice, true/false, and open-ended questions, respectively, and $T_{\text{mc}}, T_{\text{tf}}, T_{\text{oe}}$ represent the corresponding question counts. The weighting reflects the maximum possible score (5) for open-ended responses. In addition, we conducted some experiments to demonstrate the robustness of LLM-as-a-Judge in Appendix H.

Training Setup. We initialize our MoE architecture using **`Qwen2.5-7B-Instruct`** as the base model, using `bfloat16` precision. Expert modules are fine-tuned using LoRA (Hu et al., 2022) with a rank of 16, scaling factor $\alpha = 32$, and dropout rate of 0.05.

Training is conducted in two stages: expert training for 3 epochs and router training for 4 epochs. We use a batch size of 4 per device with a gradient accumulation step of 4, resulting in an effective batch size of 16. The learning rate is set to 5×10^{-5} throughout. The *PolicyBench* is partitioned into an 80/20 split for training and testing. To prevent data leakage, we perform a grouped split by policy, ensuring all questions from the same source document are kept in the same set. See subsection B.2 for more details.

4.2 Main Results

Performance improves from memorization to application. As shown in Table 2, models exhibit progressively higher accuracy from **Level 1** (memorization) to **Level 3** (application) across both Chinese and US settings. For instance, in the Chinese subset, average model scores range from **41.8%–62.0%** on Level 1 to **70.2%–80.3%** on Level 3. A similar trend is observed in the US subset. We posit that this phenomenon stems from the distinct capabilities emphasized during the different stages of LLM training. Level 1 tasks demand high-fidelity recall of specific facts (e.g., policy dates, exact terminology), which is a function of knowledge acquired during **pre-training**. While vast, this knowledge is stored implicitly, and its precise retrieval can be unreliable. In contrast, Level 2 (Understanding) and Level 3 (Application) tasks heavily rely on structured reasoning, contextual

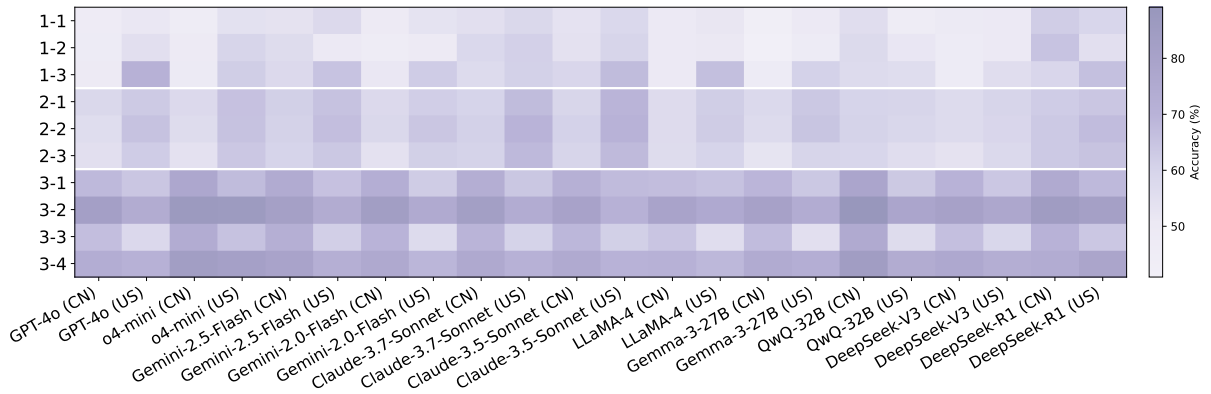

Figure 3: Model performance in 10 subtasks (ID and the specific task are shown in Table 1).

interpretation, and problem-solving—skills that are explicitly and extensively honed during **post-training** (*e.g., instruction tuning, RLHF (Ouyang et al., 2022)*). Therefore, the models' superior performance on these more complex tasks likely reflects a stronger alignment with the generalizable reasoning abilities optimized during fine-tuning, rather than a deeper mastery of the policy domain itself. *Among all evaluated models,* `DeepSeek-R1` achieves the highest overall accuracy at **66.34%**, making it a strong candidate for practitioners seeking a general-purpose model for policy-related applications.

Models excel at structured reasoning tasks but falter on abstract concepts. As shown in the task-wise heatmap (Figure 3), models consistently achieve higher accuracy on specific application-oriented tasks, particularly **Policy-Based Numerical Reasoning** and **Scenario-Based Decision-Making**. For many models, accuracy in these categories exceeds **75%**, with some surpassing **80%**. These tasks typically involve concrete conditions, rule-based logic, or everyday reasoning scenarios—formats that closely align with the pretraining and instruction-following capabilities of large language models. In contrast, accuracy remains relatively lower on more abstract or ambiguous tasks such as **Ideas** and **Institutions**, which require understanding latent policy concepts or institutional relationships. This suggests that LLMs are better equipped to handle tasks with clear logical structures than those requiring interpretive or conceptual comprehension.

Models consistently perform better on US policy questions than Chinese ones. As shown in Table 3, most models achieve higher accuracy on US policy questions than on their Chinese counterparts. The overall average improves from **61.02%** (CN) to **62.39%** (US), with models like `o4-mini` rising from **60.41%** to **65.54%**, and `Claude-3.5-Sonnet` from **62.11%** to **65.39%**. An exception is `QwQ-32B`, which performs notably better in Chinese (**65.33%**) than in English (**58.00%**). This overall trend may be attributed to the dominance of English in pretraining corpora, as well as the higher syntactic and semantic density of Chinese policy texts. These results highlight the need for more robust cross-lingual policy understanding in LLMs. Further more, we also conduct an error analysis (detailed in Appendix D).

Table 4: Qwen2.5-7B-Instruct performance across levels and regions before and after training (%)

Level	Region	Original	Training	Δ
Level 1	CN	36.85%	41.83%	↑ 13.51%
	US	23.35%	35.43%	↑ 51.73%
Level 2	CN	45.68%	47.02%	↑ 2.93%
	US	42.31%	42.78%	↑ 1.11%
Level 3	CN	64.73%	69.12%	↑ 6.78%
	US	46.65%	57.48%	↑ 23.22%

4.3 Results of *PolicyMoE*

Table 4 reports the performance of our model before and after fine-tuning with the *PolicyMoE* framework. The goal of this experiment is not to pursue state-of-the-art results with a moderately sized base model (Qwen2.5-7B-Instruct), but to demonstrate the efficacy of our expert specialization approach. Accordingly, we emphasize the **relative improvements** achieved. Notably, the fine-tuned 7B model not only shows substantial gains but also outperforms several larger baselines in Table 2, underscoring the effectiveness of domain-specific adaptation.

Across task levels, performance improves consistently after fine-tuning. The largest gain appears

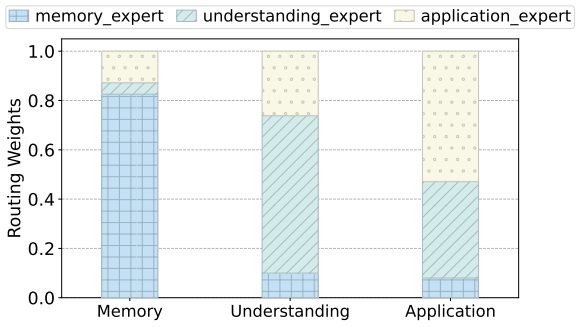

Figure 4: Routing distributions over three experts for each level.

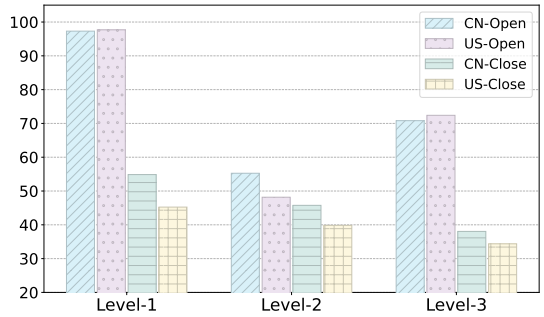

Figure 5: Human accuracy on three levels (%).

in US Level 1 tasks, where accuracy rises from 23.35% to 35.43%—a relative improvement of over 50%. China also records a 13.51% gain at the same level, highlighting the benefit of injecting structured domain knowledge. Improvements on Level 2, which emphasizes policy comprehension, are more modest (2.93% for China and 1.11% for the US), suggesting that higher-level reasoning is less sensitive to task-specific fine-tuning and may require advanced strategies such as chain-of-thought prompting (Wei et al., 2022) or richer supervision. By contrast, Level 3 tasks show more pronounced gains, with the US domain achieving a 23.22% relative improvement, indicating that the Application Expert effectively supports contextual reasoning and scenario-based decision-making.

Overall, *PolicyMoE* delivers clear benefits for both factual recall and applied reasoning. While improvements in abstract comprehension remain limited, the results point to a promising path toward specialized, capable models without relying on prohibitively large architectures. We also compare *PolicyMoE* with standard LoRA (Hu et al., 2022) in Appendix J.

4.4 *PolicyMoE* Router Analysis

To analyze the behavior of the router module, we randomly sample 10 questions from each cognitive level (Level 1–3) and compute the average expert weights assigned by the router. This allows us to observe how the model distributes attention across three experts in response to different types of tasks.

As shown in Figure 4, router weight patterns show clear specialization for factual tasks, with memory weights peaking over 80% when the memory expert is selected. In contrast, understanding and application selections result in more distributed weights, reflecting the multi-dimensional nature of these tasks and the need for shared reasoning across modules.

4.5 Human Performance

To contextualize LLM performance, we established human baselines with 12 university students from the United States and China, distinct from the annotators. Participants, proficient in English or Mandarin but without policy expertise, each answered 100 randomly sampled *PolicyBench* questions under two conditions: **open-book** (with access to policy texts) and **closed-book** (relying on prior knowledge). As shown in Figure 5, the results indicate that:

- **Open- vs. closed-book gaps distinguish memory from reasoning.** The large gap at Level 1 reflects reliance on factual recall, while the minimal difference at Level 2 suggests a shift toward reasoning-intensive challenges.
- **Cross-linguistic consistency supports benchmark validity.** Comparable performance across languages within each setting indicates that task difficulty is not driven by language differences.
- **Non-monotonic accuracy across levels.** Accuracy drops at Level 2 and partially recovers at Level 3, suggesting that abstract reasoning without direct retrieval is most challenging, whereas Level 3 allows participants to combine reasoning with general knowledge.

5 Conclusion

We introduce *PolicyBench*, a cross-system benchmark (US-China) assessing LLM comprehension of public policy, which reveals that models perform well on recall and application tasks, they struggle with understanding questions involving policy intent and institutional reasoning. To bridge this gap, we propose *PolicyMoE*, a domain-specialized MoE model that achieves improved performance. Our findings underscore the need for targeted adaptation to support real-world policy analysis.

Limitations

Geographic and systemic scope. *PolicyBench* currently covers only China and the United States. While these two cases offer a strong contrast across governance systems, they cannot fully represent the diversity of global policy environments. Extending to additional regions would improve generalizability and cross-cultural robustness.

Task diversity. The benchmark mainly relies on multiple-choice and true/false formats, with limited coverage of open-ended tasks. Real-world policy analysis, however, involves greater ambiguity, nuance, and value-sensitive reasoning than what structured formats can capture. Expanding task types would better reflect such challenges.

Model adaptation and architecture. *PolicyMoE* shows clear improvements in factual recall and applied reasoning, but its gains in abstract comprehension remain limited. Moreover, the current router design can only select one expert per query, whereas increasingly complex tasks may require activating multiple experts simultaneously. Future work could explore reasoning-focused models, advanced prompting strategies, and more flexible routing mechanisms that allocate experts based on learned weight distributions.

Ethics Statement

All data used in this study were collected from publicly accessible platforms in compliance with ethical and legal standards. No proprietary or private materials were included. For the human evaluation component, all participants were recruited on a voluntary basis and provided informed consent prior to their involvement. They were clearly informed of the research purpose and their rights, including the right to withdraw at any stage. Finally, all content generated by large language models was carefully reviewed to ensure that it did not contain sensitive, harmful, or inappropriate material. The study adheres to responsible AI research practices and aims to contribute to the safe and transparent development of policy-focused benchmarks and models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Anthropic claude-3.5-sonnet. Https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. 2025. Anthropic claude-3.7-sonnet. Https://www.anthropic.com/claude/sonnet.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. *Preprint*, arXiv:2404.07738.

Han Bao, Yue Huang, Xiaoda Wang, Zheyuan Zhang, Yujun Zhou, Carl Yang, Xiangliang Zhang, and Yanfang Ye. 2026. Position: General alignment has hit a ceiling; edge alignment must be taken seriously. *arXiv preprint arXiv:2602.20042*.

Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuying Chen, Yue Zhao, Tianyi Zhou, Mohamed Elhoseiny, and Xiangliang Zhang. 2024. Autobench-v: Can large vision-language models benchmark themselves? *arXiv preprint arXiv:2410.21259*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962.

Google. 2024. Google gemini-2-flash. Https://deepmind.google/technologies/gemini/flash-lite/.

Google. 2025. Google gemini-2.5-flash. Https://deepmind.google/technologies/gemini/flash/.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Peter A Hall and Rosemary CR Taylor. 1996. Political science and the three new institutionalisms. *Political studies*, 44(5):936–957.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. Pasa: An llm agent for comprehensive academic paper search. *Preprint*, arXiv:2501.10120.

Ce Hou, Fan Zhang, Yong Li, Haifeng Li, Gengchen Mai, Yuhao Kang, Ling Yao, Wenhao Yu, Yao Yao, Song Gao, and 1 others. 2025. Urban sensing in the era of large language models. *The Innovation*, 6(1).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Jiaju Kang, Puyu Han, Tian Zhang, and Luqi Gong. 2025. Policysimeval: A benchmark for evaluating policy outcomes through agent-based simulation. *arXiv preprint arXiv:2502.07853*.

Junmo Kang, Leonid Karlinsky, Hongyin Luo, Zhen Wang, Jacob Hansen, James Glass, David Cox, Rameswar Panda, Rogerio Feris, and Alan Ritter. 2024. Self-moe: Towards compositional large language models with self-specialized experts. *arXiv preprint arXiv:2406.12034*.

Nikos Karacapilidis, Evangelos Kalampokis, Nikolaos Giarelis, and Charalampos Mastrokostas. 2024. Generative ai and public deliberation: A framework for llm-augmented digital democracy. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Yueqing Liang, Liangwei Yang, Chen Wang, Congying Xia, Rui Meng, Xiongxiao Xu, Haoran Wang, Ali Payani, and Kai Shu. 2025. Benchmarking llms for political science: A united nations perspective. *arXiv preprint arXiv:2502.14122*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Meta. 2025. Meta llama-4. Https://www.llama.com/models/llama-4/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Paulina Jo Pesch. 2025. Potentials and challenges of large language models (llms) in the context of administrative decision-making. *European Journal of Risk Regulation*, pages 1–20.

QwQ. 2024. Qwen-qwq-32b. Https://qwenlm.github.io/blog/qwq-32b/.

Mehrdad Safaei and Justin Longo. 2024. The end of the policy analyst? testing the capability of artificial intelligence to generate plausible, persuasive, and useful policy analysis. *Digital Government: Research and Practice*, 5(1):1–35.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL 2023, page 447–451. ACM.

Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. 2020. Intellicode compose: Code generation using transformer. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 1433–1443.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Alice Xiang. 2023. Beyond the imitation game: Collaborative benchmark for measuring and extrapolating the capabilities of language models [co-authored]: What is the tao? *Transactions on Machine Learning Research*.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 610–625.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiaowen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *CoRR*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.

A Details of Data Collection

This section details the curation of our dataset, outlining the collection methodology for both policy documents and their supplementary materials, the tools utilized in the process, and the filtering pipeline applied to ensure data quality.

A.1 Data Collection Tools and Process

Our data was collected primarily between January 2015 and March 2025 using a hybrid approach to navigate the structural inconsistencies of official government websites.

The majority of documents were collected manually from the portals listed in Table 5. This manual approach was essential for bypassing complex site navigation and security mechanisms, and for ensuring the correct retrieval of policy documents with associated attachments, which challenge standard web scrapers. For a small subset of highly structured content, such as news archives, we employed targeted automation scripts using Python's **Selenium** library to assist with batch-downloading.

A.2 Data Filtering and Curation Pipeline

To ensure the quality and relevance of our final dataset, all collected documents underwent a rigorous multi-stage filtering and curation pipeline. The objective was to create a clean, non-redundant, and substantive corpus for constructing **PolicyBench**. The pipeline consisted of the following sequential steps:

1. **Duplicate Removal:** The first step was to eliminate redundant files. We identified and removed duplicates by checking for high similarity in document titles and textual content. Initially, documents with identical or near-identical titles were flagged, after which their content overlap was assessed. Documents with a high degree of textual similarity were considered duplicates, and all but the most complete version were discarded.

2. **Substantive Content Filtering:** Next, we filtered out documents that were not substantive policy texts. A document was classified as "non-substantive" and excluded if it met any of the following criteria:

    - It was purely administrative or procedural (e.g., public meeting announcements, personnel appointment notices, holiday schedules).
    - It was a table of contents, an index, or a cover page without the corresponding full document.
    - The document's title contained keywords from a predefined exclusion list, such as `"Notice of Public Hearing"`, `"Personnel Appointments"`, `"Weekly Agenda"`, or `"Annual Report Summary"`.

3. **Temporal and Relevancy Filtering:** Finally, we applied a filter to remove documents that were considered "outdated" or irrelevant to the contemporary policy landscape. A policy was flagged and removed if:

    - It was explicitly superseded by a more recent version or subsequent legislation from the same issuing authority.
    - It was promulgated before the year 2000, which we established as the historical cutoff for our benchmark to maintain modern relevance.

This structured pipeline allowed us to distill the large volume of collected data into the high-quality, curated set of 721 Chinese policies and 603 US policies that form the foundation of ***PolicyBench***.

B Details of Experiment Setting

B.1 Details of *PolicyBench*

Level-1: Memorization Task. Level-1 tasks are designed to assess factual recall. We begin by using large language models to automatically generate cloze-style and true/false questions. Cloze questions are created by masking factual elements in policy texts—such as *dates*, *legal terms*, *organization names*, and *key definitions*—that reflect domain-specific knowledge and span various policy areas. These cloze items are then transformed into multiple-choice questions. To construct high-quality distractors, we prompt multiple LLMs independently to generate alternative options inspired by (Bao et al., 2024). This multi-model strategy reduces single-model bias and increases the plausibility and diversity of distractors, thereby enhancing the benchmark's robustness.

Level-2: Understanding Task. Level-2 tasks evaluate a model's ability to comprehend the deeper meaning and context of policy content. We prompt LLMs to analyze policies using the 3I framework from policy studies, which highlights three core

dimensions: *Ideas* (underlying beliefs and values), *Interests* (stakeholders involved), and *Institutions* (rules and structures guiding implementation) (Hall and Taylor, 1996). Based on these structured analyses, we generate question-answer pairs that probe a model's understanding of policy motivations, actors, and institutional dynamics. The question generation process parallels that of Level-1: we first construct cloze-style prompts grounded in 3I insights, then convert them into multiple-choice format with distractors generated via multiple LLMs to improve quality and difficulty.

Level-3: Application Task. Level-3 tasks focus on practical reasoning and real-world contextual adaptation. To build these tasks, we draw on supplementary policy materials (*e.g., official commentaries, media coverage, expert interviews, and public consultations*) to develop realistic scenarios where a policy might be applied. Based on these scenarios, we recruit students with relevant academic backgrounds to manually craft questions that require reasoning about a policy's implications, suitability, or potential outcomes in novel contexts. This manual, context-driven approach ensures that Level-3 tasks closely mirror real-world decision-making challenges and reflect authentic policy discourse.

Expert-Led Question Design for Levels 2 and 3. To ensure high cognitive alignment and domain validity, the construction of **Level 2** (Understanding) and **Level 3** (Application) items was strictly led and executed by senior domain experts. The core writing team consisted of five senior Ph.D. candidates specializing in Public Policy, Law, and Computational Social Science. While three undergraduate assistants provided support for data formatting and preliminary cleaning, the substantive generation and validation of reasoning logic were exclusively performed by the senior doctoral researchers to ensure rigor.

Level 2 questions are designed to assess the model's ability to comprehend policy intent, stakeholder interests, and institutional logic, following the 3I framework (Ideas, Interests, Institutions). These questions emphasize abstraction and interpretation rather than factual recall. **Level 3** questions focus on practical reasoning, including scenario-based decision-making, numerical calculations, and value-driven trade-offs, often grounded in real-world policy contexts.

All questions are constructed to be clear, faithful to source policies, and cognitively representative of their designated levels (see Figure 2 for some examples). To ensure quality and consistency, the resulting items undergo a human evaluation process (for details, see Appendix E).

B.2 Details of *PolicyMoE*

Our MoE architecture is initialized using `Qwen2.5-7B-Instruct` as the base model with bfloat16 precision. Expert modules are fine-tuned using LoRA with rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate of 0.05, targeting the attention and MLP projection layers. Expert training is conducted for 3 epochs with a per-device batch size of 4 and gradient accumulation of 4 steps, resulting in an effective batch size of 16. The learning rate is set to 5×10^{-5} with weight decay of 0.01. Router training follows for 2 epochs with batch size 8 and learning rate 1×10^{-4}, using a two-layer MLP architecture with LayerNorm and GELU activation. The dataset is split into 80% training and 20% testing, with maximum sequence lengths of 2048 tokens for expert training and 512 tokens for router training across the three domains: memory, comprehension, and application.

C Related Works

LLM Benchmarks in Social Science. The rapid advancements in LLMs have enabled their application in social science research. The ability to efficiently comprehend long textual inputs and generate human-like responses makes them favorable for text-intensive tasks, like academic paper searching (He et al., 2025; Baek et al., 2025; Asai et al., 2024), document analysis (Wang et al., 2023) and legal research (Colombo et al., 2024) (Savelka, 2023). Such ability underscores the need for tailored evaluation methods that measure their performance in these domains (Bao et al., 2026). Knowledge-intensive benchmarks, such as MMLU (Wang et al., 2024) and TriviaQA (Joshi et al., 2017), test models on a wide range of subjects, from STEM fields to humanities, shedding light on the recalling and reasoning ability of LLMs to handle complex queries in real-world scenarios.

Domain-specific benchmarks now probe specialized knowledge: BioASQ (Tsatsaronis et al., 2015) tests biomedical QA, MedQA (Jin et al., 2021) evaluates clinical reasoning. Legal NLP has advanced through LawBench (Fei et al., 2024) for Chinese statutory analysis and LegalBench (Guha et al., 2023) for Anglo-American jurisprudence

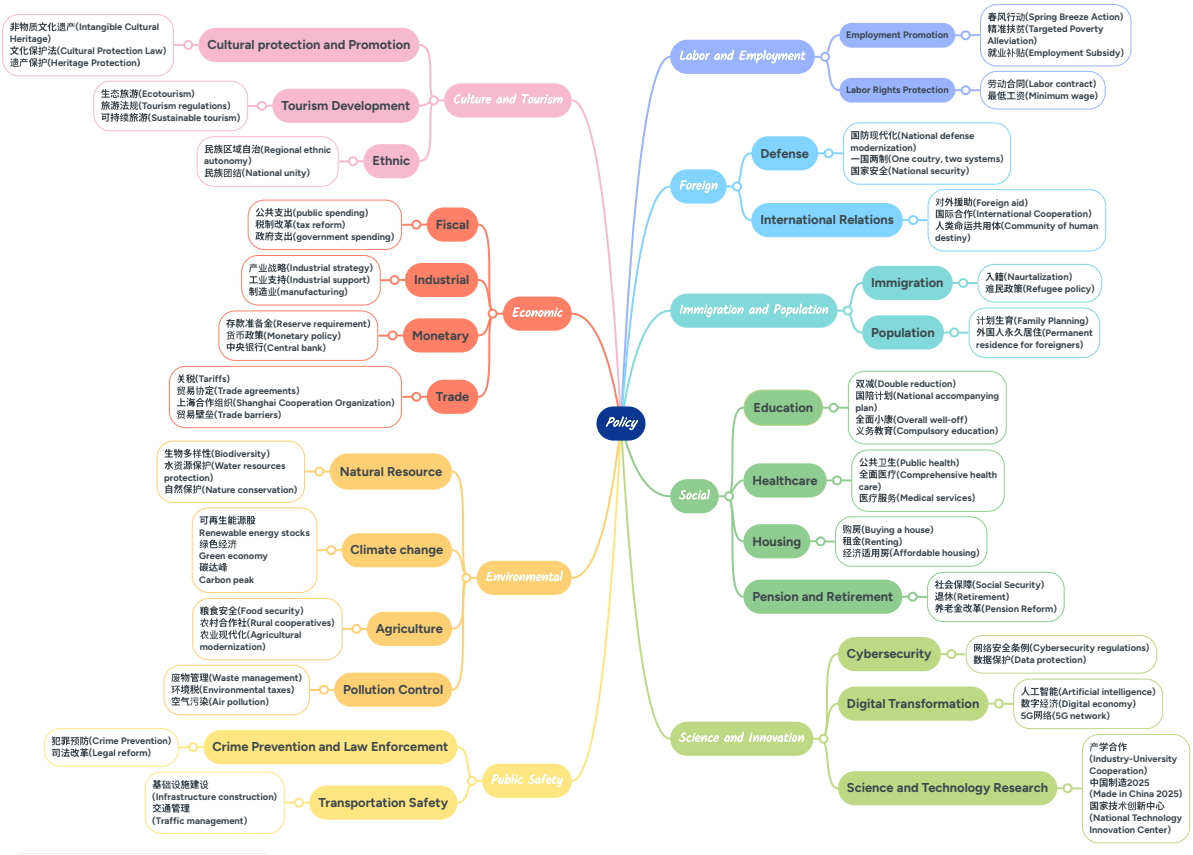

Figure 6: Categories of Chinese Policy Documents and Representative Keywords (Partial List).

tasks. Recent works have begun to address LLM evaluation in the policy domain. For instance, PolicySimEval (Kang et al., 2025) assesses policy outcomes through simulation, while UNBench (Liang et al., 2025) evaluates performance on political science tasks from a UN perspective. These benchmarks differ from *PolicyBench*, which focuses on the fine-grained comprehension of policy texts across broad domestic domains and governmental systems. In parallel, studies by Safaei et al. (Safaei and Longo, 2024) and Karacapilidis et al. (Karacapilidis et al., 2024) explore the application of LLMs for policy generation and deliberation. While these efforts target the 'output' capabilities of LLMs, our work addresses a complementary and foundational gap: evaluating the model's 'input' capability to precisely understand policy language, a crucial prerequisite for any reliable downstream application.

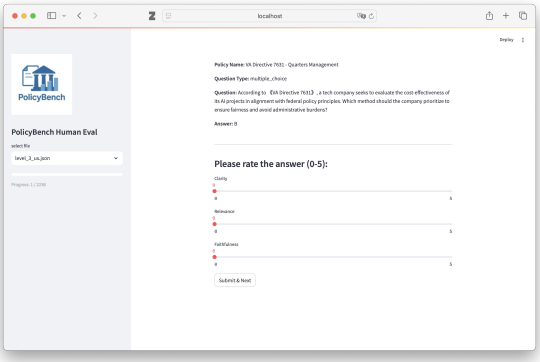

Figure 7: Screenshot of human evaluation interface.

D Error Study

To better understand the limitations of current LLMs on *PolicyBench*, we conducted a qualitative error analysis by sampling representative failure cases from each level, as illustrated in Figure 8.

Table 5: Sources for Policies and Supplementary Materials (By Language)

Language & Content Type	Primary Collection Websites / Platforms
US Sources	
Official Policies	https://www.transportation.gov
	https://www.hhs.gov
	https://www.va.gov
	https://www.commerce.gov
	https://www.usda.gov
	https://www.energy.gov
	https://www.doi.gov
	https://www.ed.gov
	https://www.treasury.gov
	https://www.state.gov
	https://www.dhs.gov
	https://www.hud.gov
Supplementary Materials	https://www.cnn.com/
	https://www.foxnews.com
	https://www.reuters.com
Chinese Sources	
Official Policies	https://www.gov.cn/zhengce/zhengcewenjianku/
Supplementary Materials	https://www.xuexi.cn/
	http://www.people.com.cn/

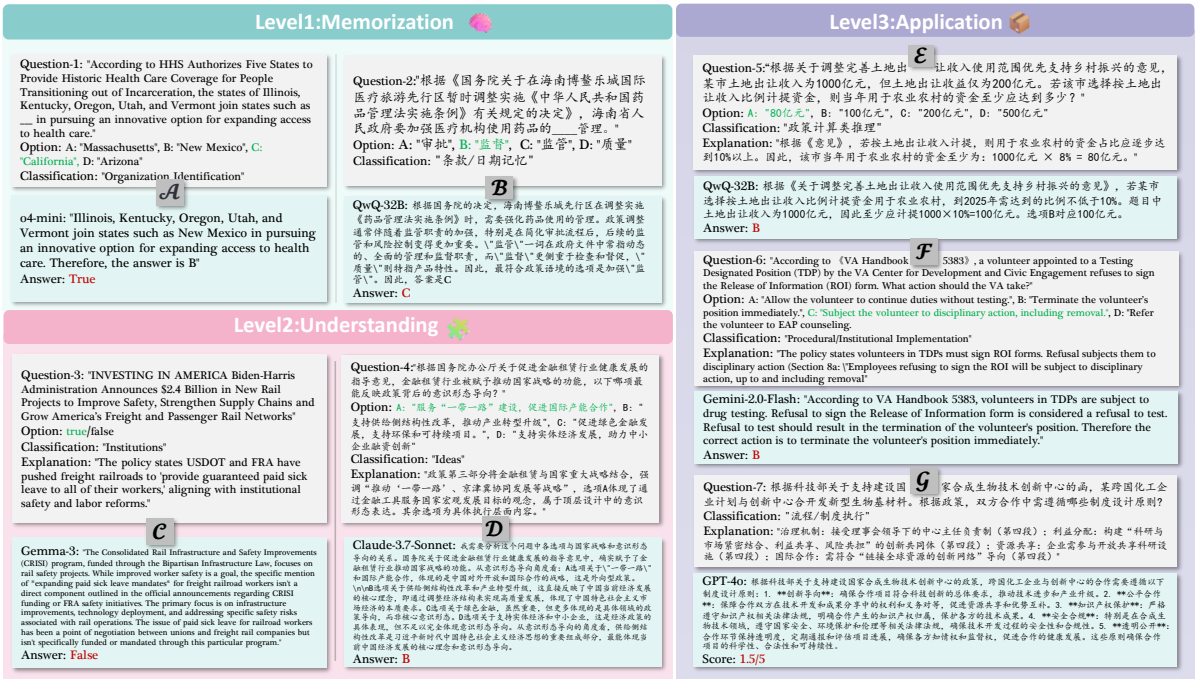

Figure 8: Representative error cases across three cognitive levels in *PolicyBench*.

Table 6: Overview of models evaluated or used as judges in this study.

Model	Developer	Open-source	Version	Role
GPT-4o	OpenAI	✗	/	Evaluated model
o4-mini		✗	2025-04-16	Evaluated model; Judge (Level 3)
Gemini-2.5-Flash	Google DeepMind	✗	preview-04-17	Evaluated model; Judge (Level 3)
Gemini-2.0-Flash		✗	/	Evaluated model
Gemma-3-27B		✓	27B Instruct	Evaluated model
Claude-3.7-Sonnet	Anthropic	✗	20250219	Evaluated model; Judge (Level 3)
Claude-3.5-Sonnet		✗	20241022	Evaluated model
LLaMA-4	Meta	✓	maverick-instruct	Evaluated model
QwQ-32B	Alibaba (Qwen Team)	✓	/	Evaluated model
DeepSeek-V3	DeepSeek AI	✓	/	Evaluated model
DeepSeek-R1		✓	/	Evaluated model; Judge (Level 3)

Level 1 (Memorization): Factual Confusions. The most common errors involve incorrect recall or misidentification, often triggered by the presence of distractors with strong semantic similarity. For example, when asked to identify US states that previously adopted a healthcare policy, the model incorrectly selected "New Mexico" instead of "California"—likely due to co-occurrence bias in training data. In Chinese policy questions, LLMs sometimes confuse regulatory terms with subtle distinctions (e.g., "Jiān Dū" vs. "Jiān Guǎn"), indicating limited sensitivity to nuanced legal language.

Level 2 (Understanding): Misinterpretations. Errors at this stage largely stem from misreading policy intent, ideological framing, or institutional logic. Models tend to misread underlying motivations or over-rely on surface cues. For instance, a model incorrectly identified "Supply-side structural reform" as the core ideological signal of a financial leasing policy, despite explicit references to national strategies like the Belt and Road Initiative. Similarly, when analyzing US labor protection clauses, the model missed the correct interpretation of sick leave provisions, favoring superficial summaries over deeper institutional mandates described in the text.

Level 3 (Application): Reasoning and Procedural Failures. Models frequently struggled with quantitative reasoning, procedural interpretation, and hallucinated conclusions. For example, a model failed to compute the required rural funding quota from land sale revenue, despite having all necessary information. In another case, a model prematurely recommended immediate termination for a volunteer's non-compliance, overlooking the stepwise disciplinary procedures mandated by policy. For open-ended tasks, models sometimes hallucinate plausible-sounding but unsupported principles, rather than extracting the specific cooperation principles explicitly mentioned in the document.

Overall, these error patterns reveal that LLMs struggle across recall, understanding, and real-world reasoning, especially with legal nuance and institutional complexity. Addressing these issues may require targeted supervision or retrieval-augmented approaches.

E Data Quality & Human Evaluation

E.1 Annotation Team & Methodology

To ensure the rigorousness of *PolicyBench*, our annotation team was structured hierarchically. The core annotation, quality control, and validation phases were led by **five senior Ph.D. candidates** specializing in Public Policy, Law, and Computational Social Science. Three undergraduate assistants provided support for preliminary data formatting and cleaning but did not perform substantive labeling or validation tasks.

E.2 General Quality Evaluation

We conducted a comprehensive human evaluation between March 2025 and April 2025. Each generated question was independently evaluated along three key dimensions on a 1–5 Likert scale. We sampled **1,500** (Level 1), **1,000** (Level 2), and **500** (Level 3) questions balanced across languages.

Table 7: The detailed introduction of 10 dimensions.

Dimension	ID	Definition
Article/Date Memorization	1-1	Tests memory of specific articles, dates, numbers, etc. in the policy.
Terminology Recognition	1-2	Tests ability to recognize and understand policy terminology.
Organization Identification	1-3	Tests ability to identify organizations mentioned in the policy.
Idea Understanding	2-1	Examines the ideological foundation and value orientation behind the policy.
Interest Understanding	2-2	Assesses the identification and analysis of key stakeholders affected by or involved in the policy.
Institution Understanding	2-3	Evaluates understanding of the formal and informal rules, organizations, and mechanisms shaping policy implementation.
Policy-Based Numerical Reasoning	3-1	Perform simple mathematical reasoning or calculation based on the numerical provisions in the policy text.
Scenario-Based Decision-Making	3-2	Based on specific scenarios, determine how the parties should make decisions or choose the most appropriate approach based on the policy.
Procedural/Institutional Implementation	3-3	Examine the understanding and memory of the specific operational procedures, implementation steps, and institutional regulations in the policy.
Policy Logic and Value Explanation	3-4	Focus on the background motivation, target value and logical structure of policy making.

As shown in Table 8, the dataset exhibits consistently high quality, with average scores exceeding **96%** across all dimensions.

E.3 Inter-Annotator Agreement (IAA)

To validate the reliability of the human evaluation, we performed a double-blind annotation on a random subset of 600 items (20% of the evaluation set). We utilized ***Cohen's Kappa*** (κ) to measure agreement beyond chance, calculated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the observed agreement and P_e is the expected agreement by chance.

Calculation Method:

- For **Bloom-level classification** (Nominal), we calculated standard unweighted κ.
- For **Quality dimensions** (Ordinal 1–5), we **binarized** the scores into "Accept" (Score ≥ 4) and "Reject" (Score ≤ 3) to strictly measure the consistency of inclusion criteria.

As shown in Table 9, we achieved strong agreement ($\kappa > 0.80$) across all dimensions. Disagreements were resolved via a two-stage adjudication process involving a senior expert.

E.4 Expert Validation Study

To further verify the dataset's validity against a professional standard, we conducted an additional expert validation study with four external experts (***2 Ph.D. candidates in Public Policy, 1 Professor in Computational Social Science, and 1 Senior Sociology Researcher***). They evaluated a stratified sample of 120 questions.

1. Data Validity: Experts assessed the correctness of the Gold Answers and the appropriateness of the content. Agreement for Bloom-level labeling was calculated using ***Krippendorff's*** α, which is robust for multiple raters:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2}$$

where D_o is the observed disagreement and D_e is the expected disagreement by chance. The results in Table 10 confirm the dataset's high quality.

2. Expert Performance Baseline: To establish a human ceiling, the experts answered the questions under an **open-book setting**, simulating realistic policy analysis workflows. As shown in Table 11, experts significantly outperform models in deeper understanding tasks (L2/L3), validating the benchmark's difficulty gradient.

F Formal Definition of the Policy Tasks

To address reviewer concerns regarding the lack of a formal task definition, we provide here a unified formulation of the *Policy Comprehension Task*, followed by fine-grained instantiations corresponding to the three cognitive levels and ten sub-tasks, details in Table 7.

Table 8: Human evaluation results across levels and regions.

Level	Region	Clarity	Relevance	Faithfulness	Avg.
Level 1	**CN**	99.20%	98.01%	98.78%	98.66%
	US	99.15%	98.25%	97.94%	98.45%
Level 2	**CN**	98.86%	98.00%	98.12%	98.32%
	US	98.76%	97.60%	97.22%	97.86%
Level 3	**CN**	99.16%	96.02%	96.14%	97.11%
	US	97.66%	95.64%	96.32%	96.54%

Table 9: Inter-Annotator Agreement statistics. Quality scores were binarized for κ calculation.

Annotation Dimension	Kappa (κ)	Agreement Level
Bloom Classification (L1/L2/L3)	0.856	Strong
Clarity (Pass/Fail)	0.841	Strong
Relevance (Pass/Fail)	0.877	Strong
Faithfulness (Pass/Fail)	0.827	Strong

Table 10: Expert Validation Results ($N = 120$).

Metric	Score
Correctness Verification	96.1%
Content Validity (Rated "Appropriate")	94.5%
Bloom-Label Agreement (Krippendorff's α)	0.86

Table 11: Human Expert Performance (Open-Book) vs. Models.

Level	Expert Accuracy	Comparison vs. SOTA LLM
Level 1 (Memorization)	82.3%	Comparable
Level 2 (Understanding)	**88.4%**	**Significant Gap**
Level 3 (Application)	**90.1%**	**Significant Gap**

F.1 General Formulation

At a high level, we model policy comprehension as a conditional question-answering problem grounded in policy documents. Let C denote a policy context, which may consist of a full policy document, a section, or a clause describing rules, actors, and institutional mechanisms. Let Q_k denote a query designed to probe a specific cognitive level $k \in \{1, 2, 3\}$, corresponding to memorization, understanding, and application, respectively.

Given (C, Q_k), a language model f_θ is expected to produce an answer $\hat{A}$ that is consistent with the policy content and satisfies the cognitive requirement implied by k:

$$\hat{A} = \arg\max_{A \in \mathcal{A}} P(A \mid C, Q_k; \theta), \tag{3}$$

where $\mathcal{A}$ denotes the answer space, which can be either a finite set of options (for multiple-choice questions) or free-form text (for open-ended questions).

F.2 Task Instantiation by Cognitive Level

We instantiate the policy comprehension task into **10 concrete sub-tasks**, organized into three cognitive levels.

Level 1: Memorization (Factual Retrieval) *Objective: Retrieve explicit facts or entities directly stated in the policy text C.*

1-1 Article / Date Memorization Recall specific temporal markers or citation identifiers.

> *Example:* "According to the *U.S.-Philippines Civil Nuclear Cooperation Agreement*, the Agreement entered into force on [Mask]."
> *Answer:* July 2.

1-2 Terminology Recognition Identify the definition of a domain-specific term as stated in the text.

> *Example:* "Civil nuclear cooperation agreements provide a legal framework for exports of [Mask]."
> *Answer:* material, equipment, and components.

1-3 Organization Identification Identify organizational hierarchy or institutional affiliation.

> *Example:* "According to the *National Behavioral Health Workforce Career Navigator*, SAMHSA is an agency within [Mask]."
> *Answer:* U.S. Department of Health and Human Services (HHS).

Level 2: Understanding (Conceptual Interpretation) *Objective: Map explicit policy text to implicit concepts under the "3I" framework (Ideas, Interests, Institutions).*

2-1 Idea Understanding Infer the underlying goal, ideology, or strategic intent.

> *Example:* "The *Big Data Project (BDP)* aims to democratize access to environmental data primarily through collaboration with whom?"
> *Answer:* Cloud Service Providers.

2-2 Interest Understanding Identify stakeholders and their eligibility, benefits, or losses.

> *Example:* "Are Tribal entities eligible to apply for 2501 Program grants according to the USDA announcement?"
> *Answer:* True.

2-3 Institution Understanding Comprehend rules, categories, or allocation mechanisms.

> *Example:* "Which category under the *Clean Energy* program allocates capacity to projects ensuring 50% of benefits go to low-income households?"
> *Answer:* Category 4.

Level 3: Application (Scenario Reasoning) *Objective: Apply policy rules to a novel or hypothetical scenario.*

3-1 Policy-Based Numerical Reasoning Execute numerical calculations derived from policy formulas.

> *Example:* "If a $2 million FEMA project has an 80% federal cost share and a 1.91% cost increase, how much does the local government pay?"
> *Answer:* $7,640.

3-2 Scenario-Based Decision Making Determine the compliant action in a simulated situation.

> *Example:* "If the Colonial Pipeline were subject to the new safety rule during a leak, which action would be required?"
> *Answer:* Stage response personnel in predefined zones.

3-3 Procedural / Institutional Implementation Validate procedural conditions or sequences.

> *Example:* "What is a necessary condition for the transfer of nuclear reactors under the U.S.-Thailand 123 Agreement?"
> *Answer:* Commitment to nonproliferation standards.

3-4 Policy Logic and Value Explanation Explain conflicts or trade-offs between policy objectives.

> *Example:* "Scott Turner's goal to reduce reliance on government aid conflicts with which aspect of DC's housing strategy?"
> *Answer:* Funding for emergency rental assistance.

This structured definition clarifies both the scope and the granularity of policy comprehension capabilities evaluated by *PolicyBench*.

G Multi-Examiner Bias Analysis

To address critical concerns regarding *"Model-Speak"* (stylistic cues) and *"Familiarity Bias"* (models favoring their own generation patterns), we conducted a comprehensive **Multi-Examiner Sensitivity Analysis**. This study empirically validates that our heterogeneous examiner pool serves as a necessary safeguard against evaluation artifacts.

G.1 Experimental Setup

We designed a controlled experiment to decouple the "Examiner" (Question Generator) from the "Examinee" (Evaluated Model).

1. The Examiner Pool: We utilized three distinct top-tier models to generate questions and distractors, ensuring coverage across different model families:

- **GPT Family:** `GPT-4o` (OpenAI)
- **Claude Family:** `Claude-4-Sonnet` (Anthropic)
- **Qwen Family:** `Qwen-3` (Alibaba Cloud)

2. The Examinee Pool: We evaluated 7 models, including the three generator families and external open-weights models, to observe cross-family behaviors:

- *Closed-Source:* `GPT-4o`, `GPT-4o-mini`, `Claude-4-Sonnet`, `Claude-4-Haiku`.
- *Open-Weights:* `Qwen-3`, `Qwen-2.5`, and `Llama-4`.

3. Evaluation Conditions: We tested these models across three distinct settings:

- **Baseline (Consensus):** Questions generated by the full pool. This is the standard *PolicyBench* setting.
- **Single-Examiner:** Questions generated exclusively by one examiner (e.g., *GPT-Only*).

- **LOEO (Leave-One-Examiner-Out):** Questions generated by the remaining two examiners (e.g., *Wo-GPT*).

G.2 Full Leaderboard Sensitivity Results

Table 12 presents the complete performance matrix. The results demonstrate significant score variations under single-examiner conditions, confirming the necessity of our Consensus Baseline.

G.3 Analysis of Biases

1. Self-Scoring Bias (Familiarity). Models consistently perform differently on questions they generated themselves, creating a "Familiarity Bonus" or "Penalty". As shown in Table 13, relying on a single generator creates severe distortions:
Insight: The data reveals divergent biases. GPT-4o benefits from "Familiarity Bias" (+7%), likely exploiting its own stylistic patterns. Conversely, Claude exhibits "Self-Strictness" (-15%), penalizing its own generation logic. **PolicyBench's consensus approach effectively averages out these extremes**, anchoring scores to a neutral ground truth.
2. Mitigating Model-Speak: Leaderboard Stability. To determine if "Model-Speak" (stylistic tells) compromises the validity of the rankings, we analyzed the **Spearman Rank Correlation** (ρ) (Zar, 2005) between the Baseline leaderboard and other conditions.
3. External Model Robustness. For models outside the generator pool, like **Llama-4**, the Baseline provides the most stable evaluation. Llama-4's score varies from 82.0% to 89.0% across single examiners. The Baseline (82.0%) successfully anchors it to a consensus difficulty, filtering out examiner-specific noise.

Conclusion: The low correlation of the GPT-Only set ($\rho = 0.34$) proves that single-source benchmarks are fundamentally biased. The multi-examiner design in *PolicyBench* is a necessary mechanism to ensure that high performance reflects genuine *Policy Comprehension* rather than *Stylistic Alignment*.

H LLM-as-a-Judge Validation

To ensure the reliability and validity of our automated evaluation pipeline, we conducted a two-fold validation process: assessing *stability* (consistency across runs) and *alignment* (accuracy against human experts).

H.1 Evaluation Stability Analysis

Evaluation Stability. To rigorously assess the stability of our LLM-as-a-Judge pipeline and address concerns about the stochastic nature of model outputs, we conducted a multi-run, multi-judge evaluation analysis. We randomly sampled 10 question-answer pairs from the Level 3 open-ended test set. The scoring for each sample was performed as follows:

- **Initial Scoring:** For each evaluation run, two distinct LLM judges were sampled from our pool to score the response on a 0–5 scale (in 0.5 increments). The score for the run was the average of these two ratings.
- **Discrepancy Resolution:** If the scores from the two initial judges differed by more than 1.0 point, a third tie-breaker judge was invoked to provide an additional score. In such cases, the final score for the run was the average of all three judges' ratings. This protocol ensures robustness against outlier judgments.
- **Repetition:** This entire scoring process was conducted three independent times for each of the 10 cases.

We then calculated the mean and standard deviation of the three final scores for each case. The results, summarized in Table 15, demonstrate high consistency, with standard deviations remaining exceptionally low. This indicates that our multi-judge protocol effectively mitigates run-to-run variance and produces reliable evaluations.

H.2 Human-Model Alignment

Stability is a necessary but insufficient condition for validity. To demonstrate that our LLM-as-a-Judge pipeline accurately reflects expert judgment, we conducted a **Human-Model Alignment Study**.
Experimental Setup. We sampled a subset of open-ended responses from Level 3 tasks. These responses were independently scored by our LLM Judge pipeline and by senior human experts (Ph.D. candidates in Public Policy) using the exact same rubric.
Quantitative Results. As shown in Table 16, the LLM Judge demonstrates strong alignment with human experts across three key metrics:

- **Pearson Correlation ($r = 0.87$):** Indicates a strong positive linear relationship between model and human scores.

Table 12: Complete Multi-Examiner Sensitivity Analysis. Scores represent accuracy (%). **"Baseline"** denotes the standard PolicyBench setting (3-Examiner Consensus). **"Wo-X"** denotes the Leave-One-Examiner-Out setting. The data reveals that relying on a single examiner (e.g., GPT-Only or Claude-Only) leads to drastic score fluctuations compared to the stable Baseline.

Model	Baseline (Consensus)	Single-Examiner Setting			LOEO Setting			Avg.
		Claude-Only	GPT-Only	Qwen-Only	Wo-Claude	Wo-GPT	Wo-Qwen	
Qwen-3	89.0	86.0	88.0	**92.0**	**92.0**	**90.0**	88.0	89.29
Llama-4	82.0	84.0	88.0	89.0	88.0	87.0	85.0	86.14
Qwen-2.5	83.0	81.0	85.0	90.0	83.0	86.0	**87.0**	85.00
GPT-4o-mini	66.0	59.0	74.0	85.0	74.0	71.0	67.0	82.67
GPT-4o	75.0	71.0	**82.0**	85.0	85.0	83.0	78.0	79.86
Claude-4-Sonnet	84.0	69.0	49.0	89.0	71.0	52.0	85.0	71.29
Claude-4-Haiku	60.0	47.0	81.0	86.0	59.0	48.0	75.0	91.20

Table 13: Analysis of Self-Scoring Bias. The results show that single-examiner benchmarks suffer from extreme variance (from +7 inflation to -15 deflation).

Model	Baseline Score	Own-Gen Score	Δ	Bias Type
GPT-4o	75.0%	82.0%	**+7.0%**	*Self-Leniency / Pattern Matching*
Qwen-3	89.0%	92.0%	+3.0%	*Moderate Inflation*
Claude-4-Sonnet	84.0%	69.0%	**-15.0%**	*Self-Strictness / Hyper-Critical*

Table 14: Leaderboard Stability Analysis. High correlation in LOEO conditions confirms the robustness of the ranking system, while single-examiner rankings (especially GPT-Only) are highly unstable.

Comparison Pair	Spearman ρ	Kendall τ	Interpretation
Baseline vs. **Wo-Qwen**	**0.901**	0.781	*Highly Stable*
Baseline vs. **Wo-GPT**	0.607	0.524	*Moderately Stable*
Baseline vs. **GPT-Only**	**0.342**	0.293	***Unstable / Biased***

- **Mean Absolute Error (MAE $= 0.42$):** On average, the model's score deviates from the human score by less than 0.5 points (the smallest scoring increment).
- **Agreement Rate** (94%)**:** In 94% of cases, the model's score fell within an acceptable margin (≤ 1.0 point) of the expert score.

Mitigating Subjectivity via Rubric-Based Scoring. The high alignment is primarily attributed to our **Point-Based Rubric** design (detailed in Appendix G). Unlike generic "quality" assessments which can be subjective, our prompt explicitly directs the model to verify the presence of specific *Key Points* derived from the reference answer. The model assigns partial credit based on these matched points rather than an abstract "feeling" of quality. This structured approach significantly reduces hallucination and subjectivity, ensuring the judge acts as an objective verifier.

I Correlation with External Benchmarks

To demonstrate that *PolicyBench* evaluates a distinct capability orthogonal to general reasoning or pure legal knowledge, we analyzed the performance correlation between *PolicyBench* and two representative benchmarks: **MMLU-Pro** (Wang et al., 2024) (General Reasoning) and **LegalBench** (Guha et al., 2023) (Legal Reasoning).

Key Findings:

- **Negative Correlation with General Reasoning** ($r \approx -0.69$)**:** Surprisingly, models with top-tier general reasoning (e.g., `DeepSeek-V3`) do not necessarily excel in policy comprehension. This suggests that policy analysis involves specific logic (e.g., institutional constraints) that general benchmarks fail to capture.
- **No Correlation with Legal Reasoning** ($r \approx -0.07$)**:** The near-zero correlation with LegalBench indicates that *understanding policy* (Ideas, Interests) is fundamentally different from *applying law* (Statutes). *PolicyBench* fills this critical gap in the evaluation landscape.

J Comparing *PolicyMoE* with a Standard LoRA

To validate the architectural contribution of *PolicyMoE* and demonstrate its superiority over standard parameter-efficient fine-tuning, we conducted

Table 15: Stability analysis of the final LLM-as-a-Judge scores. The low standard deviation across three independent runs for each case demonstrates the reliability of our multi-judge evaluation protocol.

Case ID	Final Score (Run 1)	Final Score (Run 2)	Final Score (Run 3)	Mean	Std. Dev.
Case 1	4.00	4.50	4.00	4.17	0.29
Case 2	4.00	4.00	4.25	4.08	0.12
Case 3	3.00	3.25	3.00	3.08	0.12
Case 4	4.33	4.50	4.25	4.36	0.10
Case 5	2.25	2.25	2.50	2.33	0.12
Case 6	5.00	5.00	5.00	5.00	0.00
Case 7	3.67	4.00	3.50	3.72	0.25
Case 8	1.50	1.25	1.00	1.25	0.20
Case 9	4.00	3.75	4.00	3.92	0.12
Case 10	4.67	5.00	4.50	4.72	0.25

Table 16: Human–Model alignment statistics. High correlation, low error, and strong agreement indicate that the LLM judge closely matches expert evaluation.

Metric	Value
Pearson Correlation (r)	**0.87**
Mean Absolute Error (MAE)	**0.42**
Agreement Rate	**94%**

a controlled ablation study. We compared our *PolicyMoE* architecture against a well-tuned **Standard LoRA** baseline.

J.1 Experimental Setup

To ensure a fair comparison, both models were trained on the same stratified subset of the PolicyBench training data using `Qwen2.5-7B-Instruct` as the backbone.

- **Standard LoRA:** Fine-tuned using a single LoRA adapter (Rank=16, Alpha=32) applied to all target modules, treating all levels of tasks as a unified objective.
- ***PolicyMoE* (Ours):** Fine-tuned using our routing architecture with three specialized experts (Memory, Understanding, Application), utilizing the same data and hyperparameters.

J.2 Results & Analysis

As shown in Table 18, *PolicyMoE* outperforms the Standard LoRA baseline across all metrics, with an average accuracy improvement of **+2.68%**.

Mitigating Task Interference. These results suggest that the MoE architecture helps alleviate *task interference* arising from heterogeneous cognitive objectives.

- **Decoupling Memorization and Reasoning:** Standard LoRA must encode both exact factual recall (Level 1) and flexible scenario reasoning (Level 3) within a single low-rank adaptation, which can lead to competing gradient signals.
- **Improved Performance on Level 1:** By routing Level 1 queries to a dedicated **Memory Expert**, *PolicyMoE* achieves a **+3.81%** improvement over Standard LoRA, suggesting that specialized experts help preserve precise factual representations.
- **Implication:** These findings indicate that *PolicyMoE* provides a structurally meaningful extension over standard LoRA when adapting LLMs to multi-level policy comprehension tasks.

K Selected Policy Samples

This section showcases examples of the policy titles from our dataset, from both China and the United States. A partial list is provided in Figure 9.

L Prompt Template

This section shows the 3 key prompts used for data curation. Prompt 1 converts clean policy text into cloze-style questions. Prompt 2 instructs LLMs to generate incorrect answers, which serve as the distractors for multiple-choice questions. Prompt 3 employs an LLM-as-a-judge methodology to evaluate the models' responses to the questions. Prompts used for processing policies from different countries were designed to correspond to the respective national language. Only the English prompts are presented here. Prompts were translated using Google Translate to ensure consistency between the two languages. All translated content was sub-

Figure 9: Collected Policies (Part).

Table 17: Performance comparison and correlation analysis across benchmarks.

Model	**MMLU-Pro**	**LegalBench**	***PolicyBench* (Avg)**	***PolicyBench* (L2)**
DeepSeek-V3	**81.9%**	80.1%	59.10%	57.68%
GPT-4o	80.3%	79.8%	59.47%	56.08%
Claude-3.7-Sonnet	80.3%	78.1%	**64.13%**	58.48%
Gemini-2.5-Flash	77.9%	**81.7%**	63.82%	**64.06%**
Pearson Correlation (r)	**-0.69**	**-0.07**	**1.00**	–

Table 18: Performance comparison between the base model, Standard LoRA, and *PolicyMoE* on a controlled training subset. *PolicyMoE* achieves larger gains on Memorization (L1) and Application (L3) tasks.

Model	**Level 1** (Memorization)	**Level 2** (Understanding)	**Level 3** (Application)	**Average**
Base (Qwen2.5-7B)	30.10%	44.00%	55.69%	43.26%
Standard LoRA	34.82%	44.51%	59.45%	46.26%
***PolicyMoE* (Ours)**	**38.63%**	**44.90%**	**63.30%**	**48.94%**
Δ *(MoE vs. LoRA)*	***+3.81%***	*+0.39%*	***+3.85%***	***+2.68%***

sequently reviewed and tested by human annotators to avoid potential semantic inconsistencies.

Prompt 1: Level-1 cloze-style Generation

You are a policy expert and your task is to generate questions based on the given policy text.
- Strictly follow the given material to generate questions, do not fabricate content.
- Generate 5-10 fill-in-the-blank questions and 3-8 true or false questions based on the length of the policy.
- Answer to the questions should be clear and precise, avoid ambiguous answers.
- Answer to the questions should preferably be a single word or phrase.
- If the answer is a false judgment question, Avoid altering, adding, or deleting the original text when the answer is not an incorrect judgment question.
- Don't generate questions related to non-key information such as file numbers.
- Each question should start with "According toXXX", please provide the full name of the policy.
- Questions should not be limited to a single aspect (such as time), and should be diversified.

policy title: {policy}
policy content: {policy_content}

Prompt 2: Distractor generation

You are an ai assistant tasked with answering policy-related questions.

- Answer the questions based on your knowledge.
- Please note that some incorrect answers are provided below. You must not make the same mistakes,
- Your answer needs to be semantically distinct from the given incorrect answer.
- Don't say you can't see the image, just answer based on your knowledge.
- Don't generate overly lengthy answers, keep them concise and to the point.
- The answer you generate needs to be factually different from the given incorrect answer.
- Try to use straightforward words instead of being too abstract or vague.

question:{question}
wrong answers:{wrong_answer}

Prompt 3: LLM-as-a-Judge

You are an expert evaluator. Your task is to score the following open-ended answer based on a reference answer and scoring criteria. Follow these rules carefully:
1. For calculation or factual questions where the result must be precise (e.g., math, unit conversion, logical problems), if the final answer is incorrect, the score should be 0, regardless of the explanation.
2. For general questions (e.g., reasoning, explanation, analysis), the reference answer includes multiple key points.
- Compare the given answer with the reference key points.
- For each matched key point, assign partial credit proportionally.
- If the answer includes correct but unlisted points (beyond the reference answer), you may award partial credit with explanation.
3. Provide a score from 0 to 5. Generally:
- 5 = Completely correct and well explained
- 4 = Mostly correct, with minor issues
- 3 = Partially correct, some key points missing or wrong
- 2 = Mostly incorrect but with small redeeming aspects
- 1 = Barely relevant or correct
- 0 = Completely wrong or irrelevant
4. In your reasoning, clearly list:
- Which points in the reference answer are matched
- Any extra correct points beyond the reference
- Justify any deductions
5. Be strict but fair. Do not be lenient.
—
Question: {question}
Reference Answer:
{reference_answer_with_point_marks}
User Answer: {user_answer}
—
Now output:
Score: X
Reasoning: ...