

## Highlights

### **POLIS-Bench: Towards Multi-Dimensional Evaluation of LLMs for Bilingual Policy Tasks in Governmental Scenarios**

Tingyue Yang, Junchi Yao, Yuhui Guo, Chang Liu

- **Systematic Bilingual Benchmarking:** Introduction of POLIS-BENCH, the first systematic, rigorous evaluation suite for LLMs in governmental bilingual (Chinese/English) policy scenarios, designed to address issues of "false success" and compliance violation in policy generation.
- **Scenario-Grounded Task Distillation:** Distillation of three specialized, scenario-based tasks (Clause Retrieval & Interpretation, Solution Generation, and Compliance Judgment) that comprehensively probe models' policy understanding and adherence to compliance constraints.
- **Dual-Metric Evaluation Framework:** Establishment of a novel dual-metric evaluation framework combining semantic similarity and LLM.Judge accuracy to quantitatively assess both lexical alignment and genuine task correctness in policy contexts.
- **Cost-Efficient Task Adaptation:** Successful task-aligned fine-tuning of lightweight POLIS series open-source models, achieving parity with or surpassing strong proprietary baselines (e.g., GPT-4 series) on key policy subtasks while maintaining general capabilities and significantly lowering deployment costs.

# POLIS-Bench: Towards Multi-Dimensional Evaluation of LLMs for Bilingual Policy Tasks in Governmental Scenarios

Tingyue Yang<sup>a</sup>, Junchi Yao<sup>a</sup>, Yuhui Guo<sup>a,\*</sup>, Chang Liu<sup>a</sup>

<sup>a</sup>University of Electronic Science and Technology of China,

## Abstract

We introduce POLIS-BENCH, the first rigorous, systematic evaluation suite designed for LLMs operating in governmental bilingual policy scenarios. Compared to existing benchmarks, POLIS-BENCH introduces three major advancements. (i) **Up-to-date Bilingual Corpus:** We construct an extensive, up-to-date policy corpus that significantly scales the effective assessment sample size, ensuring relevance to current governance practice. (ii) **Scenario-Grounded Task Design:** We distill three specialized, scenario-grounded tasks—Clause Retrieval & Interpretation, Solution Generation, and the Compliance Judgment—to comprehensively probe model understanding and application. (iii) **Dual-Metric Evaluation Framework:** We establish a novel dual-metric evaluation framework combining semantic similarity with accuracy rate to precisely measure both content alignment and task requirement adherence. A large-scale evaluation of over 10 state-of-the-art LLMs on POLIS-BENCH reveals a clear performance hierarchy where reasoning models maintain superior cross-task stability and accuracy, highlighting the difficulty of compliance tasks. Furthermore, leveraging our benchmark, we successfully fine-tune a lightweight open-source model. The resulting POLIS series models achieves parity with, or surpasses, strong proprietary baselines on multiple policy subtasks at a significantly reduced cost, providing a cost-effective and compliant path for robust real-world governmental deployment.

**Keywords:** Large Language Models, Natural Language Processing, Evaluation Benchmark, Public Governance

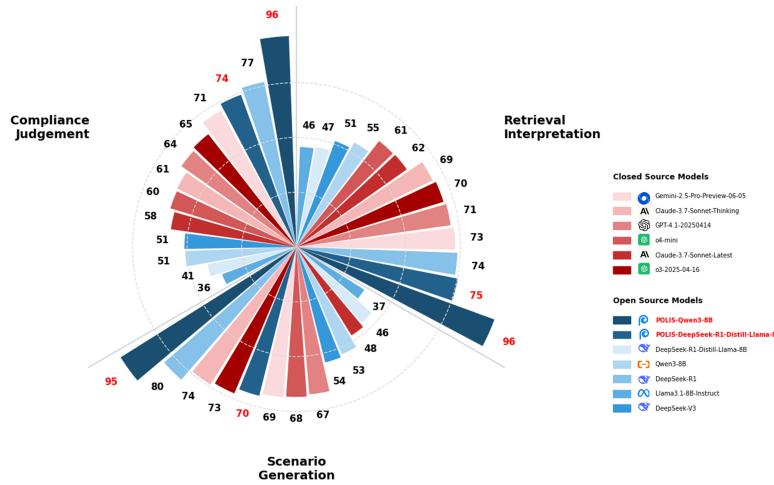


Figure 1: Comparison of accuracy rate across 15 state-of-the-art models and POLIS series models.

\*Corresponding author

Email addresses: 2023170201018@std.uestc.edu.cn (Tingyue Yang), 2022160101012@std.uestc.edu.cn (Junchi Yao), gyh@uestc.edu.cn (Yuhui Guo)

## 1. Introduction

In recent years, large language models (LLMs) have been widely adopted for their strengths in text modeling, information retrieval, and related abilities in e-Commerce [1, 2], healthcare [3, 4], and education [5, 6]. Meanwhile, with the rise of digital government, LLMs have attracted growing attention in public governance. A survey of 142 U.S. government departments reported that nearly 45 percent planned to deploy or had already deployed AI algorithms in governance tasks [7]. However, while LLMs offer new technical leverage, their risks cannot be ignored: repetition [8], bias, and data contamination [9], as well as potential public safety and ethical concerns [10, 11]. These issues can distort understanding during task execution and lead to negative social impacts.

However, current LLMs often overlook key properties of policy-oriented generation. Compared with ordinary text, policy documents are typically lengthy, structurally complex, and dense with specialized terminology, requiring precise retrieval and interpretation across articles and contexts. In addition, policy analysis must strictly adhere to policy goals and compliance constraints. Ignoring these requirements frequently may lead to “false success”, where an answer appears reasonable on the surface yet departs from the intent of the provisions or overlooks compliance boundaries.

Current benchmark studies specific to public-sector governance remain scarce, and existing evaluations of models on governance-oriented generation tasks often lack sound task design and evaluation methods [12]. Concretely, (i) the effective sample sizes for generation tasks are small, undermining representativeness and stability; (ii) task categories are too coarse to cover the diverse real world policy scenarios; and (iii) evaluation methods still rely primarily on manual sampling and scoring, which limits scale and authority.

To address these gaps, we introduce **POLIS-BENCH**—Policy-Oriented Language Intelligence Suite **Benchmark**, a rigorously selected language intelligence suite for policy domains that standardizes corpus, tasks, and dual-metrics evaluation into one benchmark and with three key improvements: **First**, to address the problem of small actual assessment sample size of existing policy evaluation benchmarks, we build an up-to-date and realistic policy dataset aggregating recent Chinese–English policy texts from diverse sources and continually expanding the corpus to reflect current governance practice. **Second**, to address the problem of rough classification of existing policy evaluation benchmark tasks, we designed scenario-based tasks. For each policy, we distill three task types—clause retrieval and interpretation, solution generation for concrete problems, and compliance judgment. **Third**, to address the problem of single and inefficient evaluation of existing policy evaluation benchmarks, we propose a dual-metric evaluation framework that combines semantic similarity with LLMJudge accuracy: cosine similarity measures the lexical closeness between model outputs and references, while LLMJudge assesses, in task context, whether an output satisfies the problem requirements.

In a large-scale evaluation of more than 10 state-of-the-art LLMs, we observe a clear performance hierarchy. **Firstly**, the closed-source reasoning models steadily dominate on multiple tasks and maintain higher consistency in cross-language and cross-task scenarios which demonstrates strong integrated reasoning capabilities. In contrast, though closed-source chat models have good language fluency and clause retrieval, their correctness and robustness usually lag behind reasoning-based routes in complex scenarios. **Secondly**, the open-source reasoning models perform relatively well, especially in the Chinese side and scenario-based tasks, but their overall stability is still far from that of the closed-source strong baseline model. Comparatively, open-source chat models are more prone to bias in cross-clause evidence integration and compliance boundary judgment, and get a lower score overall. Based on the evaluation results, we selected two high-performing open-source baseline models for joint multi-task *LoRA* fine-tuning. Through a comprehensive analysis of the three types of tasks and two metrics, we found that the fine-tuned models can significantly improve semantic alignment and cross-clause localization capabilities in clause retrieval and interpretation. In addition, the fine-tuned model reduces the number of responses that are irrelevant to policy objectives or too general, and reduces unnecessary redundancy in scenario generation. Meanwhile, the fine-tuned model can identify constraint boundaries more robustly in compliance judgment and significantly reduce the incidence of non-compliant proposals. Overall, the fine-tuned model provides more accurate and relevant answers to questions in the government generation task, and matches or exceeds the strong closed-source baseline (e.g., GPT-4 series) in several subtasks, while maintaining smaller parameter sizes and lower calling costs, which significantly improves the deployment efficiency and feasibility of the policy task. Our contributions are:

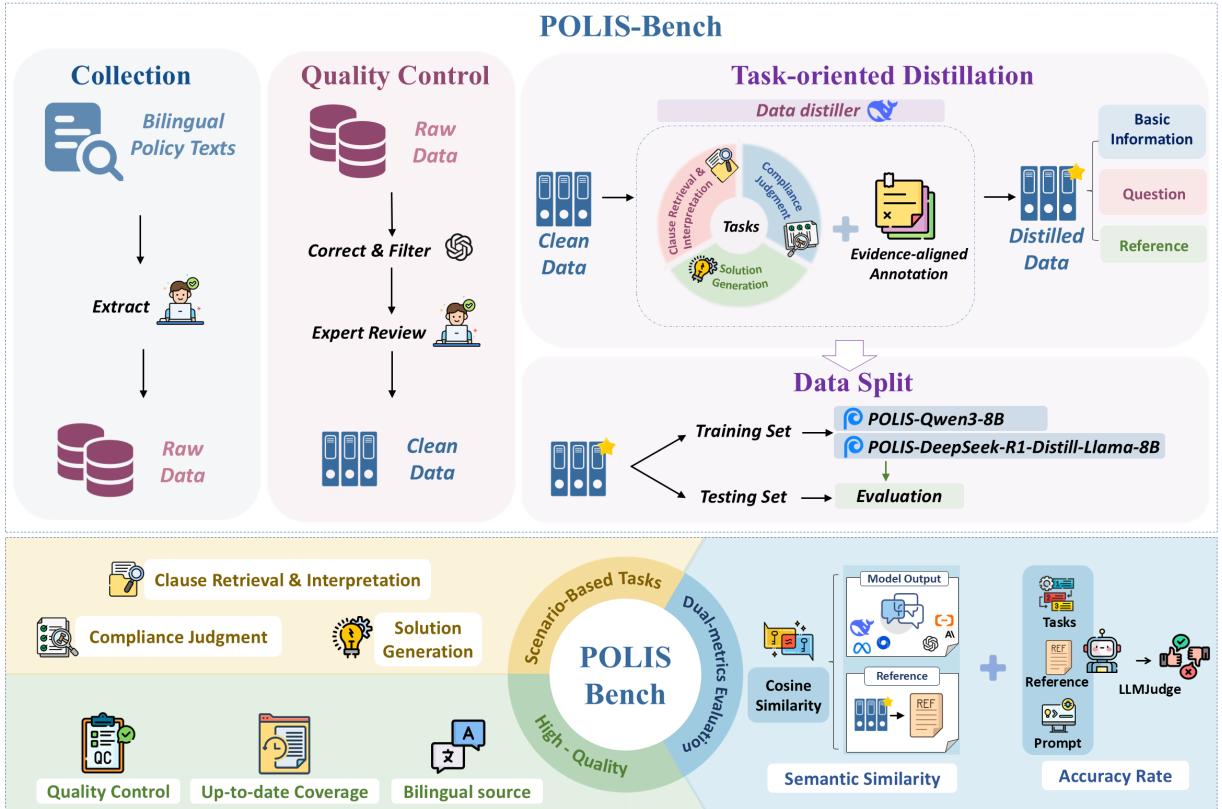


Figure 2: **Overview of our POLIS-Bench.** (i) The diagram above illustrates construction pipeline of POLIS-BENCH. (ii) The diagram below illustrates the three key characteristics of the POLIS-BENCH.

1. **POLIS-BENCH.** We present a systematic benchmark for policy-oriented generation that compiles up-to-date bilingual (Chinese/English) policies, distills each policy into three tasks—clause retrieval and interpretation, solution generation, and compliance judgment—and offers a unified, reproducible evaluation pipeline.
2. **Dual-metrics evaluation framework** We establish an evaluation system with two metrics, namely semantic proximity and task correctness. The first measures how closely a model’s output aligns with the reference answer in content and expression. The second assesses, within concrete task contexts, whether the answer satisfies the requirements of the problem and adheres to the compliance constraints. The system is quantitatively clear and can be directly migrated and extended in bilingual and multi-tasking scenarios.
3. **Lightweight model adaptation** Using **POLIS-BENCH**, we successfully fine-tune a lightweight open-source model. The resulting POLIS-Qwen-8B delivers clear gains on all three tasks, achieves performance equal to or better than strong closed-source baselines at lower cost, and offers a cost-efficient, compliance-friendly path for real-world governmental deployment.

## 2. Related Work

**The Emergence and Challenges of LLMs in Public Policy.** In recent years, the application of LLMs within public policy has seen a marked increase in academic research, revealing the multifaceted value of these models throughout the entire policy life cycle. Current studies suggest that LLMs are improving evidence-based decision-making and governance by enhancing data analysis and predictive capabilities, thereby contributing to the creation of public value [13, 14]. For instance, in the critical domain of public safety, researchers have applied neural network models [15] to dynamically forecast spatiotemporal crime data, assisting in the optimization of law enforcement resource allocation and risk regulation. Moreover, LLMs are considered a key driver in advancing e-Government modernization,

improving service quality and operational efficiency, and achieving higher citizen acceptance in certain contexts compared to traditional human-delivered services [16, 17]. Additionally, natural language processing (NLP) technologies embedded within LLMs are transforming the policy analysis process itself, facilitating public participation through the automatic evaluation of large volumes of citizen feedback [18] and serving as a powerful tool for policy analysis [19] in trend identification and information synthesis. More notably, LLMs have been employed in designing complex economic policies. The “AI Economist” framework [20] uses deep reinforcement learning to design tax policies, achieving social welfare outcomes in simulations that surpass traditional economic models [20]. Collectively, these studies demonstrate that LLMs have evolved from mere auxiliary tools to comprehensive transformative engines that optimize governance, enhance democratic interaction, and foster innovative policy solutions.

**From Universal to Specialized Benchmarks: The Unique Demands of Policy Analysis.** The evaluation of LLMs has evolved from universal benchmarks—such as multi-task challenges like GLUE and SuperGLUE [21, 22] and measures of general world knowledge like MMLU and its enhanced versions [23]—to more specialized domains. The scope of these universal benchmarks has expanded to include dimensions such as cross-lingual competence with XTREME [24, 25], multimodal comprehension with IGLUE [24], as well as evaluations of the “human-likeness” of generated text with TURINGBENCH [26] and personalized long-form content generation with LongLaMP [27], thereby comprehensively assessing the fundamental capabilities of models. However, the limitations of universal benchmarks have spurred the development of specialized evaluation frameworks for specific high-risk domains. In the legal domain, LexEval and LawBench [28, 29] in the Chinese context have established multi-dimensional legal cognition frameworks, while in the English-speaking world, LegalBench and CaseHOLD [30, 31] collaborate with legal experts to focus on specific tasks such as case holding identification. In the financial sector, evaluations have evolved from early systems like FLUE [32] and PIXIU [33] to comprehensive frameworks such as FinBen [34], which covers tasks from sentiment analysis to quantitative trading, and FinRL-Meta [35], which focuses on reinforcement learning-driven trading strategies. In the healthcare domain, FFlamby [36] concentrates on privacy concerns within federated learning environments, while XLINGHEALTH [37] is dedicated to addressing information disparities in cross-lingual medical question-answering. While these specialized benchmarks provide profound insights within their respective fields, they fundamentally differ from the core requirements of policy analysis. Evaluations in the legal, financial, and healthcare domains are largely centered on established rules, objective data, or scientific consensus, with a primary focus on factual accuracy and predictive precision. In contrast, the core task of policy analysis involves not only fact extraction but also the comprehension and articulation of competing ideological perspectives, the weighing of trade-offs, conducting social impact assessments, and providing policy recommendations. This unique demand for reasoning that incorporates argumentation, normative considerations, and value plurality is not adequately addressed by existing benchmarks, making the development of a dedicated policy analysis benchmark both necessary and urgent.

**Opportunities and Challenges in Government Affairs Evaluation.** Compared to domains such as law, finance, and healthcare, the evaluation of LLMs within the government affairs domain is still in its early stages. Pioneering efforts, such as those presented by Liu in MSGABench [12], have made significant contributions by constructing a comprehensive evaluation framework that highlights challenges related to model reliability, security, and other factors, providing a critical foundation for future research. However, while these efforts have effectively diagnosed macro-level deficiencies in model capabilities, significant limitations remain in fine-grained, in-depth evaluations of policy-related text generation tasks, particularly in task design and evaluation methodology.

Building upon these research experiences and acknowledging existing gaps, this paper proposes **POLIS-BENCH**, a benchmark inspired by the development of evaluation frameworks in high-risk domains such as healthcare [38] and responding to the call for multi-criteria decision-based evaluations [39]. **POLIS-Bench** aims to complement existing government affairs benchmarks by addressing data timeliness and bilingual coverage, while also representing a critical step in refining task design and evaluation methodologies. The goal is to provide a more accurate and responsible framework for assessing the application of LLMs in the field of public policy, thereby supporting the responsible and effective use of these models.

### 3. Method

#### 3.1. Overview

We introduce POLIS–BENCH, the first benchmark for policy analysis tasks, designed to evaluate LLMs’ reading, reasoning, and compliance aware generation in real governmental policy contexts. Its contributions are threefold: (i) a comprehensive, up-to-date bilingual (Chinese/English) policy corpus that tracks current governance practice; (ii) three scenario-grounded task types that jointly probe retrieval, generation, and compliance; and (iii) a dual-metric evaluation framework comprising two metrics—semantic proximity and task correctness—and providing a unified, reusable pipeline. Building on this benchmark, we perform multi-task joint fine-tuning on open-source backbones; the fine-tuned models match or surpass strong closed-source baselines across multiple subtasks while delivering lower cost and stronger deployability.

**Dataset Perspective** POLIS–BENCH compiles bilingual policy texts from approximately the past five years, sourced from officially released policy documents, and is continuously expanded to reflect current governance practice. For each policy, we design three tasks—Clause Retrieval & Interpretation, Solution Generation, and Compliance Judgment—and perform data distillation to obtain 1,000 items for each of the three categories. All questions are generated from the corresponding policy texts, and the answer–evidence spans in the texts are precisely annotated and used as supervision for training and evaluation.

**Evaluation** Within POLIS–BENCH, we design a dual-metrics evaluation framework: (i) **Semantic similarity** uses cosine similarity to measure the degree of similarity between model outputs and reference answers; (ii) **Accuracy rate** uses LLMJudge to determine correctness. Compared with relying solely on cosine similarity as a single machine score, we introduce LLMJudge scoring to provide a comprehensive evaluation of core semantic coverage, the legitimacy of explanations, and information veracity, thus more faithfully reflecting overall performance on policy tasks in governmental scenarios.

#### 3.2. Data Construction

**Up-to-date Data Collection** We systematically collect long-form bilingual policy texts released over the past five years from official government websites, reflecting current governance practice and keeping the corpus continuously expanded. Considering that some originals are lengthy and exceed the input budget, we apply length-constrained, standardized segmentation to the original texts, ultimately obtaining 500 normalized long-text samples in Chinese and 500 in English (1,000 in total) as the raw database.

**Quality Control** To ensure usability and consistency, we apply structured quality control to the original result. First, we use LLMs to correct and filter the raw dataset: it validates and parses JSON fields, removes samples that are unparseable or missing key fields, and deduplicates when duplicates or templated outputs occur yielding samples that are clear and structurally uniform. Then, we conducted the manual sampling inspection to review and correct the machine-screened samples. The resulting Clean Data serve as the input to the subsequent task-oriented distillation.

**Task-oriented Distillation** To convert general policy texts into evaluable task samples, we perform instruction-driven data distillation: given each policy’s title and body, we use a prompt template to generate, in order, three question types corresponding to the three tasks—Clause Retrieval & Interpretation, Solution Generation, and Compliance Judgment—and require the precise textual source for each question (i.e., the concrete phrasing from the original policy). This directly supports subsequent capability evaluation for different tasks: clause-type questions align with policy provisions and key interpretive points; solution-type questions align with policy goals and execution pathways; compliance-type questions align with boundary and rule requirements.

**Evidence aligned Annotation** During distillation we annotate, for each question, the answer–evidence spans drawn from the long-form policy text as the alignment signal for training and evaluation: Clause Retrieval & Interpretation aligns to the relevant provisions and their key interpretive statements; Solution Generation aligns to policy segments related to solution points; Compliance Judgment aligns to clause statements that delimit constraint boundaries. This annotation scheme preserves traceability between samples and originals, enabling evaluation to focus on the core goal of “answering on the basis of the policy text.”

**Construction Pipeline** Data construction follows a unified pipeline to ensure compliant sources, consistent structure, and usable formats. First, obtain policy titles and bodies and standardize them into long text inputs. Second, for overly long originals, conduct sentence level, length constrained splitting and fragment processing to ensure cross language consistency and comparability. Then, organize task samples in a “question–answer evidence” structure, so that each sample can be located in the original policy text. The entire process focuses on mapping original policy texts to task-specific samples, ensuring that all task instances are derived from the same underlying corpus and remain auditable and reusable.

**Evaluation Readiness & Data Split** The final data is released as standardized sample units: each sample contains policy metadata, task type, the question text, and its corresponding source annotation. The three tasks share the same construction and annotation procedure and connect directly to our evaluation framework. Meanwhile, from the full set of 3,058 instances, we randomly sampled 2,500 instances as the training set for LoRA fine-tuning and 558 instances as the test set for evaluation only. Through this construction pipeline and annotation scheme, the dataset attains a consistent standard in coverage, structure, and evaluation readiness, providing a solid foundation for comparing different models and settings.

### 3.3. Evaluation

We used two complementary types of metrics to evaluate the state-of-the-art models’ performance in governmental scenarios: **semantic similarity** to measure semantic proximity, and **accuracy rate** to judge the correctness of the model responses in terms of the dimensions of the task requirements.

**Semantic similarity** In the semantic similarity setting, we use cosine similarity between the vector representations of the model output and the reference answer to quantify semantic closeness. To accommodate long texts, we first perform sentence level segmentation on both the model output and the reference, and concatenate the resulting sentences into several segments, which are then compared in pairs. For any sentence pair in the vector space, let the vectors be  $\mathbf{v}_1$  and  $\mathbf{v}_2$ ; the *cosine similarity* is:

$$CS(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|}. \quad (1)$$

We compute the *cosine similarity* for all valid sentence–pair combinations between the two segments and take the *arithmetic mean* across all valid sentence pairs as the final similarity score for that sample.

**Accuracy rate (LLMJudge)** In computing the accuracy rate, we adopt LLMJudge with few-shot prompting, generating an explainable binary decision for each sample from which the accuracy rate is computed. Specifically, we design a unified LLMJudge instruction template that stipulates three criteria: first, whether the core idea is consistent with the reference; second, whether the key elements of the explanatory information can be located in and supported by the reference; and whether any fabricated content or information not covered by the reference is present. To stabilize decision boundaries, the template includes a small set of examples covering both correct and incorrect cases, and the judge model is invoked directly during evaluation to make the decision. The evaluation runs at the sample level: for each sample the judge returns a label, recorded as Pass = 1 corresponding to [Correct] or Pass = 0 corresponding to [Incorrect]; we then aggregate within each task over samples to obtain the task level accuracy rate.

$$\text{Accuracy} = \frac{\text{number of Pass samples}}{\text{total number of samples}}. \quad (2)$$

## 4. Experiment

### 4.1. Setup

To systematically evaluate models’ retrieval, generation, and compliance judgment abilities in policy contexts, we select a representative suite that spans closed source and open source models as well as reasoning and chat paradigms. We then compare these models under a unified evaluation pipeline with shared hyperparameters. Selection follows the same basic criteria: recency, meaning the latest or currently available versions; diversity, meaning both reasoning

oriented and general chat architectures; and practical scale, meaning a focus on small and medium models because of compute limits. The full list is:

**Closed-source: Reasoning** Gemini-2.5-Pro-Preview-06-05 [40], Claude-3.7-Sonnet-Thinking [41], o4-mini [42], o3-2025-04-16 [42].

**Chat** Claude-3.7-Sonnet-Latest [41], GPT-4.1-20250414 [43].

**Open-source Reasoning** DeepSeek-R1 [44], DeepSeek-R1-Distill-Llama-8B [45], Qwen3-8B [46].

**Chat** DeepSeek-V3 [47], Llama3.1-8B-Instruct [48, 49].

These models cover mainstream approaches in both ecosystems and form contrasted baselines along the reasoning and dialogue axes. Detailed experimental settings are provided in the appendix Appendix A.

#### 4.2. Evaluation Result and Analysis

**Some open-source models exhibit discrepancies between literal semantic alignment and genuine task intent comprehension.** It is noteworthy that high semantic similarity does not necessarily indicate the model’s true understanding of task intent. Table 1 shows that Qwen3-8B achieved a semantic similarity of 0.64 but only 0.53 accuracy; while LLaMA3.1-8B-Instruct achieved an even higher semantic similarity of 0.71 yet scored only 0.40 in accuracy. This indicates that while some open-source models can obtain high similarity scores by directly retrieving or parroting policy clauses, they fall short in actual task comprehension and reasoning. This results in responses that appear reasonable on the surface but deviate substantially from the intended meaning. This reflects that some open-source models still lack deep semantic understanding and dynamic generation capabilities in scenario-based policy tasks.

**The reasoning model demonstrated significantly superior overall accuracy.** Figure 3 shows that reasoning model’s overall performance is above average.

More specifically, as shown in Table 1, the open-source reasoning model DeepSeek-R1 achieved an overall accuracy rate of 0.77, whereas the chat model DeepSeek-V3 recorded only 0.53. More notably, within the same closed-source model series, the reasoning model Claude-3.7-Sonnet-Thinking achieved an overall accuracy rate of 0.68, markedly surpassing the chat model Sonnet-latest’s 0.55. This indicates that government tasks heavily rely on the complete logical chain from comprehension and reasoning, and reasoning models can more consistently maintain high accuracy across tasks.

**Most Models Generally Perform Better in Handling English Policies than Chinese Policies.** The results in Table 2 show that most models perform better in handling English policies. For example, o4-mini has an accuracy rate of 0.69 for English policies, while it is only 0.58 for Chinese, which is a significant difference. Similarly, Qwen3-8B has a weaker performance with an accuracy rate of 0.57 for English policies and only 0.48 for Chinese policies. Most of the models perform more consistently in processing tasks in English policies, indicating that there is still room for optimisation of the *State-Of-The-Art* model at the multi-language level, and in particular, the improvement of Chinese task adaptation is crucial to model performance.

**DeepSeek-R1 performs well in dealing with Chinese policies.** As can be seen from the results in Table B.4 in Appendix B.1, DeepSeek-R1 achieves the highest accuracy rate of 0.79 in dealing with Chinese policies, showing relatively strong Chinese optimisation capabilities. In the three tasks of processing Chinese policies, as shown in

Table 1: **Overall Evaluation Results.** Quantitative results on **POLIS-BENCH** across closed-source & open-source and reasoning & chat model families. For each language, the table reports two metrics: **semantic similarity** (mean semantic similarity) and **accuracy rate** (task completion accuracy measured by LLMJudge) under a unified evaluation pipeline. Cells highlighted in red indicate the **highest score** within each column, and blue indicates the **second highest**.

Model	similarity_mean	accuracy_rate
<b>Closed-source Reasoning Model</b>		
Gemini-2.5-Pro-Preview-06-05	0.68	0.71
Claude-3.7-Sonnet-Thinking	0.65	0.68
o4-mini	0.66	0.63
o3-2025-04-16	0.65	0.70
<b>Closed-source Chat Model</b>		
Claude-3.7-Sonnet-Latest	0.69	0.55
GPT-4.1-20250414	0.70	0.68
<b>Open-source Reasoning Model</b>		
DeepSeek-R1	0.65	0.77
DeepSeek-R1-Distill-Llama-8B	0.63	0.45
Qwen3-8B	0.64	0.53
<b>Open-source Chat Model</b>		
DeepSeek-V3	0.67	0.52
Llama3.1-8B-Instruct	0.71	0.40

Table 2: **Cross-language Evaluation Results.** Quantitative results on **POLIS-BENCH** under **Chinese (CN)** and **English (EN)** policy contexts, covering closed-source & open-source and reasoning & chat model families. For each language, the table reports two metrics: **semantic similarity** (mean semantic similarity) and **accuracy\_rate** (task completion accuracy measured by LLMJudge) under a unified evaluation pipeline. Cells highlighted in **red** indicate the **highest score** within each subcolumn, and **blue** indicates the **second highest**.

Model	CN		EN	
	semantic similarity	accuracy_rate	semantic similarity	accuracy_rate
<b>Closed-Source Reasoning Model</b>				
Gemini-2.5-Pro-Preview-06-05	0.63	0.72	0.74	0.70
Claude-3.7-Sonnet-Thinking	0.60	0.68	0.71	0.68
o4-mini	0.63	0.58	0.69	0.69
o3-2025-04-16	0.60	0.67	0.71	0.74
<b>Closed-Source Chat Model</b>				
Claude-3.7-Sonnet-Latest	0.71	0.53	0.68	0.57
GPT-4.1-20250414	0.66	0.62	0.74	0.75
<b>Open-Source Reasoning Model</b>				
DeepSeek-R1	0.59	0.79	0.72	0.75
DeepSeek-R1-Distill-Llama-8B	0.58	0.39	0.69	0.52
Qwen3-8B	0.58	0.48	0.71	0.59
<b>Open-Source Chat Model</b>				
DeepSeek-V3	0.65	0.49	0.71	0.56
Llama3.1-8B-Instruct	0.69	0.33	0.74	0.48

Table B.4 in Appendix B.1, DeepSeek-R1 achieves accuracy rates of 0.75, 0.84 and 0.77 respectively, which are significantly higher than the performance of other models under Chinese policies, showing its stability and accuracy in Chinese policy processing tasks. This suggests that multicorpus adaptation and multilingual optimisation may be able to improve the model’s multilingual adaptability and robustness in multilingual contexts.

### Reasoning models have better robustness across tasks.

As demonstrated in Figure 3, reasoning models exhibit considerable advantages in cross-task stability. For instance, in the three Chinese tasks, the reasoning model average accuracy rates are 0.60, 0.64, and 0.61, respectively, showing only minor performance fluctuations, which indicates excellent stability. In contrast, the chat model average accuracy rates in the same three types of tasks are 0.56, 0.45, and 0.49, with a larger fluctuation in accuracy, and a noticeable performance decrease in the “Solution Generation” and “Compliance Judgment” tasks. This shows that reasoning models are more robust in handling complex tasks and ensuring accuracy and reliability in tasks that require understanding policy details, making inferences and identifying compliance boundaries.

### Compliance judgment Class Tasks are More Challenging.

According to the experimental result data in Figure 3, in terms of performance on different task types, the compliance judgment class tasks are significantly more difficult. As demonstrated in Figure 3, the compliance judgment tasks exhibited the lowest overall mean accuracy rates in both the Chinese and English policy scenarios, at 0.56 and 0.60 respectively. Specifically, as shown in Table B.5 in Appendix B.1, Claude-3.7-Sonnet-Thinking’s accuracy rate of 0.56 on the compliance judgment class task is significantly lower than the accuracy rate of 0.74 obtained on the clause retrieval task and the accuracy rate of 0.67 obtained on the scenario generation task in English policy scenario. Accordingly, we believe that when the problem involves clause boundary conditions, the model needs to process more contextual information and perform complex reasoning chains, which requires stronger reasoning and comprehension abilities, and makes it much more difficult for models to deal with these tasks.

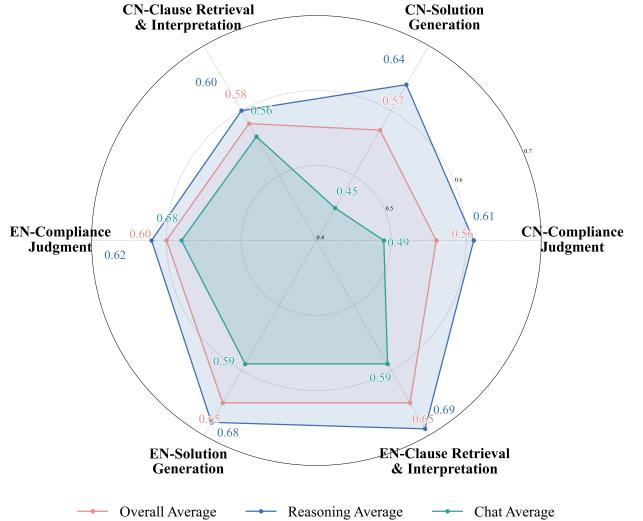


Figure 3: Accuracy Rate Distribution. This chart illustrates the accuracy rate distribution for the average performance of Reasoning Models (Reasoning Average), Chat Models (Chat Average), and the overall mean (Overall Average) across the three bilingual policy tasks on POLIS-BENCH.

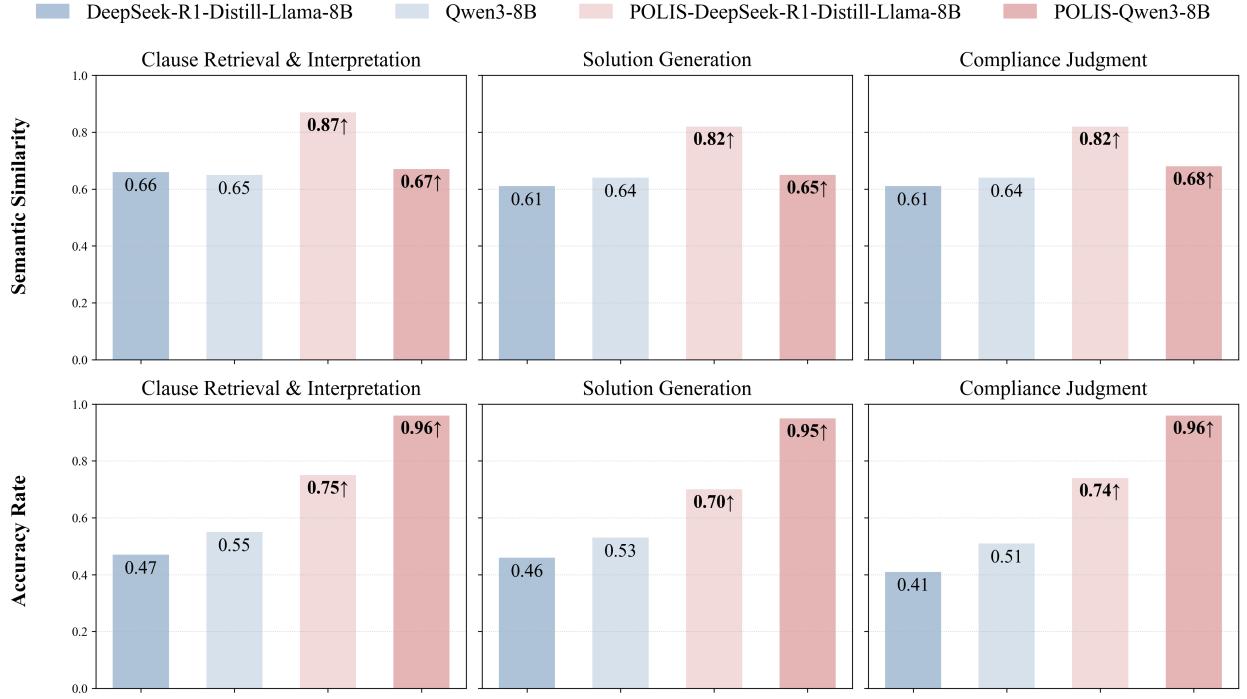


Figure 4: **Task-oriented performance of base and POLIS-tuned models.** Bar plots compare four models (DeepSeek-R1-Distill-Llama-8B, Qwen3-8B, POLIS-DeepSeek-R1-Distill-Llama-8B, POLIS-Qwen3-8B) on three tasks: *Clause Retrieval & Interpretation*, *Solution Generation*, and *Compliance Judgment*. The top row reports *Semantic Similarity*, and the bottom row reports *Accuracy Rate*; higher values indicate better performance. Numbers atop bars are the corresponding scores, obtained under the unified evaluation pipeline.

#### 4.3. Training Result and Analysis

**Fine-tuning** To evaluate the effectiveness of POLIS-BENCH, we conduct LoRA fine-tuning on two open-source models: Qwen3-8B and DeepSeek-R1-distill-llama-8b that performed well in our preliminary evaluation. These models were trained on the training set of POLIS-BENCH dataset and subsequently evaluated on the test set. As shown in Figure 4, POLIS-Qwen3-8b, which was fine-tuned on Qwen3-8b, exhibited a significant improvement in accuracy rate, achieving 0.96, 0.95 and 0.96 respectively in 3 tasks, while maintaining semantic similarity with the mean score of 0.67. Compared with the base model Qwen3-8B, which gets accuracy rates of 0.55, 0.53 and 0.51 in 3 tasks and mean semantic similarity score of 0.64, POLIS-Qwen3-8b achieved an average improvement of 80% in accuracy, accompanied by a slight improvement in semantic coverage. Meanwhile, POLIS-DeepSeek-R1-distill-llama-8b, which fine-tuned on DeepSeek-R1-distill-llama-8b, shows a notable increase in semantic similarity, achieving an average score of 0.84 and its accuracy score has been elevated by 0.28 above its backbone model, which represents a 62% increase. Both POLIS models achieve remarkable evaluation outcomes that remain competitive with several strong general-purpose baselines, and even meet or exceed those of several leading closed-source general models, such as the GPT series, on both the Chinese and English subsets. This suggests that task-aligned fine-tuning on the POLIS-Bench dataset consistently improves multi-task performance within policy-oriented settings.

**General capability evaluation** To verify that task-aligned fine-tuning for government tasks does not compromise a model’s general capabilities, we conducted a controlled evaluation on the *GPQA-Diamond* benchmark[50]. *GPQA-Diamond* is a high-difficulty, multi-disciplinary multiple-choice set spanning biology, physics, and chemistry, with four options per question and a 25% random-guess baseline. In the *GPQA* series, *GPQA-Diamond* is particularly stringent that PhD-level experts can only achieve about 65% accuracy, while skilled non-experts with web access reach about 34%. That makes it effective for testing advanced reasoning and scientific general knowledge. We used the official Diamond subset (198 questions) for closed-book testing with consistent 5-shot prompt and fixed inference hyperparameters.

As shown in the Table 3, Qwen3-8B maintains the same Diamond accuracy before and after fine-tuning (36.87%), DeepSeek-distill-Llama3-8B improves from 28.3% to 30.4% after fine-tuning. That indicates that task-aligned fine-tuning does not weaken its reasoning and knowledge-retrieval ability on cross-disciplinary, high-difficulty multiple choice. Given the difficulty and ceiling of *GPQA-Diamond*, these differences lie within a reasonable range of statistical fluctuation. The results indicate that the fine-tuned models preserve downstream task performance and that their core capabilities—such as advanced reasoning—remain intact, supporting robustness and broad applicability in real-world deployments.

Table 3: The general capability evaluation result of POLIS-DeepSeek-R1-Distill-Llama-8B, POLIS-Qwen3-8B and their backbone models on *GPQA-Diamond* benchmark

Model	Accuracy
DeepSeek-R1-Distill-Llama-8B	0.28
Qwen3-8B	0.37
<b>POLIS-DeepSeek-R1-Distill-Llama-8B</b>	0.3
<b>POLIS-Qwen3-8B</b>	0.37

**Query:** What specific compliance requirements must the project on supporting AI compliance fulfillment meet?  
What are the expected forms of deliverables for this project?

**Context:**

II. Key Projects...4. *Research on the Standardized Path for Electronic Document and Electronic Archive Management to Support AI Compliance Implementation* (Research period: 1 year, Project No. 2025Z004, Funding: 50,000 RMB). Main Expected Outcome: A policy advisory report on the standardized path for electronic document and electronic archive management to support AI compliance implementation.

 **POLIS-deepeek-r1-distill-lama-8b**

**Answer:**

The policy text clearly states: “Main expected outcome: ... standardized path policy advisory report.”

 **Improvement: High Fidelity**

Precisely extracts the correct term from the source.  
Suppresses the tendency to fabricate information.

 **POLIS-Qwen3-8B**

**Answer:**

According to the policy text description : The compliance requirements must meet ... compliance implementation implicitly following the *Data Security Law* ...;The expected deliverable is a “... standardized path policy advisory report” which should produce a practical compliance technology roadmap and a policy recommendation document.

 **Improvement: Suppressed Hallucination & Enhanced Conciseness**

Avoids inventing a structured list, attempting to ground its inference.  
Reduces redundancy, providing a more direct answer.

 **DeepSeek-R1-Distill-Llama-8B**

**Answer:**

...The expected forms of deliverables include:

1. Standardized path document
2. Research report
3. Outcome presentation report

 **Problem: Hallucination**

Fabricates deliverables not present in the text.  
Fails to identify the specific constraint “policy advisory report” .

 **Qwen3-8B**

**Answer:**

...The specific compliance requirements to be met include:1. Data privacy and security compliance 2. Technical compliance 3. Archival management standards compliance 4. Risk prevention and control compliance. The expected deliverable of the project is a *policy advisory report*, that is, a research-based standardized path framework designed to guide practical implementation and provide references for policy formulation and industry application.

 **Problem: Hallucination & Redundancy**

Hallucinates a full list of compliance requirements .  
Adds redundant interpretation to the correct answer.

Figure 5: Case Study on Compliance Judgment Task

#### 4.4. Case Study

To more intuitively illustrate the performance improvements achieved through task-aligned parameter-efficient LoRA fine-tuning based on **POLIS-BENCH**, we conducted a case study comparing the outputs of the fine-tuned model with those of two baseline models: Qwen3-8B and DeepSeek-R1-Distill-Llama-8B. This comparative analysis was designed to demonstrate the fine-tuned model’s enhanced capacity to accurately capture task-specific requirements in policy-related contexts. In the first case as shown in Figure 5, the task required the model to accurately delineate the boundaries of missing information as specified by the policy text. However, because the text did not provide a detailed elaboration of the “compliance requirements”, both baseline models were adversely influenced by contextual ambiguity and incorrectly fabricated non-existent clauses such as “data privacy compliance.” This behavior exemplifies a typical hallucination phenomenon, leading to the generation of factually incorrect content. In the second case as shown in Figure 6, both baseline models introduced redundant interpretations and generalized summaries in their responses. Although the core information was preserved, the inclusion of such extraneous content undermined the conciseness and faithfulness of their outputs relative to the source text.

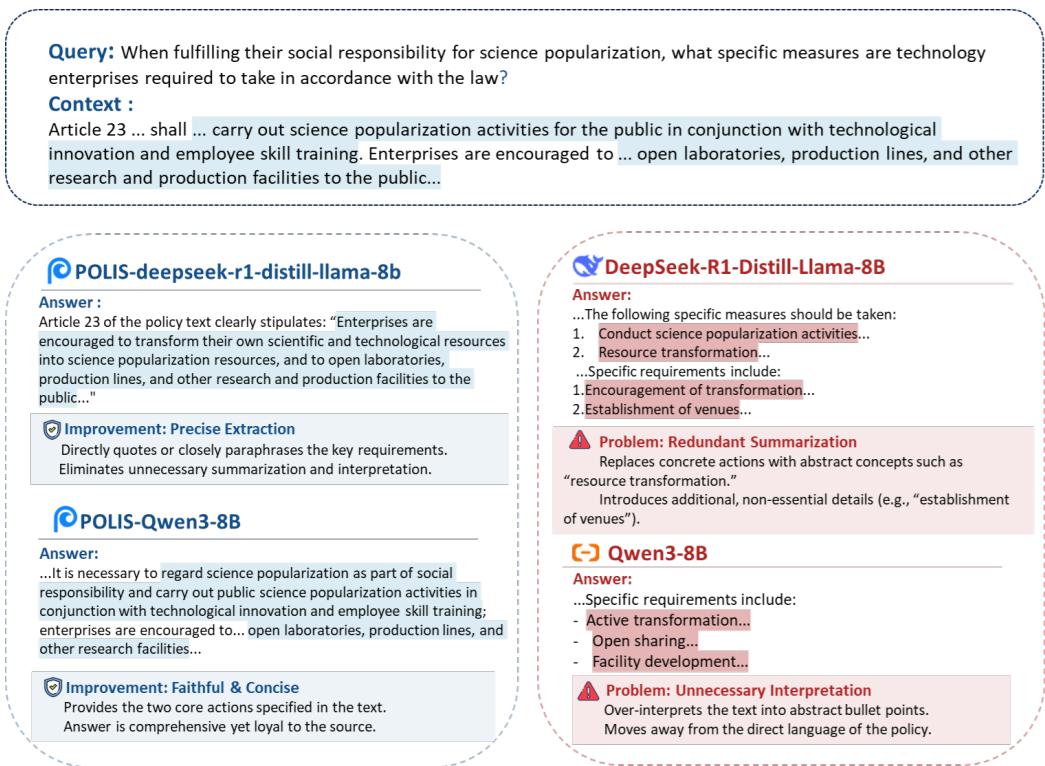


Figure 6: Case Study on Solution Generation Task

Collectively, these results indicate that task-aligned LoRA fine-tuning grounded in POLIS-BENCH, when implemented under a unified instruction template, effectively mitigates persistent issues such as factual hallucination and redundant generation observed in base models. In key tasks—including Compliance Judgment and Solution Generation—the fine-tuned model demonstrates superior textual fidelity and task alignment. These findings provide empirical evidence that the proposed approach can achieve performance comparable to, or exceeding, that of strong proprietary baselines while maintaining parameter efficiency, cost-effectiveness, and enhanced deployability.

## 5. Conclusion

In this work, we introduced POLIS-BENCH, a systematic benchmark for policy-oriented generation that couples an up-to-date bilingual policy corpus with three scenario-grounded tasks: Clause Retrieval and Interpretation, Solution Generation, and Compliance Judgment. And a unified dual-metrics evaluation pipeline for semantic proximity and task correctness. Across a representative suite of closed- and open-source models, our large scale study reveals a consistent performance hierarchy: reasoning models deliver stronger and more stable results than chat models, with compliance judgment posing the greatest challenge among the three task families. Building on the benchmark, task-aligned fine-tuning leads to clear multi-task gains. In particular, our lightweight adapted model attains parity with or exceeds strong closed-source baselines at lower cost, while preserving general reasoning ability as verified on GPQA-Diamond.

Looking forward, we see two priorities. First, broaden coverage. Extend to additional jurisdictions, policy domains, and multilingual settings, and stress-test models on harder, boundary-heavy cases to better characterize retrieval, generation, and compliance trade-offs. Second, institute dynamic updates and governance: continuously refresh the corpus with newly released regulations, calibrate LLMJudge prompts against expert rubrics, and strengthen contamination auditing to safeguard integrity and reproducibility. With wider scope and stronger dataset governance, future iterations can provide a more comprehensive and durable standard for evaluating and improving policy-aware language intelligence.

## References

- [1] C. Palen-Michel, R. Wang, Y. Zhang, D. Yu, C. Xu, Z. Wu, Investigating llm applications in e-commerce, 2024. URL: <https://arxiv.org/abs/2408.12779>. arXiv:2408.12779.
- [2] D. Xu, H. Wang, Multi-agent collaboration for b2b workflow monitoring, *Knowledge-Based Systems* 15 (2002) 485–491. URL: <https://www.sciencedirect.com/science/article/pii/S0950705102000333>. doi:[https://doi.org/10.1016/S0950-7051\(02\)00033-3](https://doi.org/10.1016/S0950-7051(02)00033-3).
- [3] J. Qiu, K. Lam, G. Li, A. Acharya, T. Y. Wong, A. Darzi, W. Yuan, E. J. Topol, Llm-based agentic systems in medicine and healthcare, *Nature Machine Intelligence* 6 (2024) 1418–1420.
- [4] O. Ramos-Soto, J. Ramos-Frutos, E. Pérez-Zarate, D. Oliva, S. E. Balderas-Mata, MIAFEX: An attention-based feature extraction method for medical image classification, *Knowledge-Based Systems* 330 (2025) 114468. URL: <https://www.sciencedirect.com/science/article/pii/S0950705125015072>. doi:<https://doi.org/10.1016/j.knosys.2025.114468>.
- [5] Z. Chu, S. Wang, J. Xie, T. Zhu, Y. Yan, J. Ye, A. Zhong, X. Hu, J. Liang, P. S. Yu, et al., Llm agents for education: Advances and applications, arXiv preprint arXiv:2503.11733 (2025).
- [6] Z. Duan, H. Gu, Y. Ke, D. Zhou, Ebert: A lightweight expression-enhanced large-scale pre-trained language model for mathematics education, *Knowledge-Based Systems* 300 (2024) 112118. URL: <https://www.sciencedirect.com/science/article/pii/S0950705124007524>. doi:<https://doi.org/10.1016/j.knosys.2024.112118>.
- [7] D. F. Engstrom, D. E. Ho, C. M. Sharkey, M.-F. Cuéllar, Government by algorithm: Artificial intelligence in federal administrative agencies, *SSRN Electronic Journal* (2020). URL: <https://doi.org/10.2139/ssrn.3551505>. doi:[10.2139/ssrn.3551505](https://doi.org/10.2139/ssrn.3551505).
- [8] J. Yao, S. Yang, J. Xu, L. Hu, M. Li, D. Wang, Understanding the repeat curse in large language models from a feature perspective, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, p. 7787–7815. URL: <http://dx.doi.org/10.18653/v1/2025.findings-acl.406>. doi:[10.18653/v1/2025.findings-acl.406](https://doi.org/10.18653/v1/2025.findings-acl.406).
- [9] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024. URL: <https://arxiv.org/abs/2308.05374>. arXiv:2308.05374.
- [10] X. Li, M. Liu, S. Gao, W. Buntine, A survey on out-of-distribution evaluation of neural nlp models (2023).
- [11] O. S. Al-Mushayt, Automating e-government services with artificial intelligence, *IEEE Access* 7 (2019) 146821–146829. doi:[10.1109/ACCESS.2019.2946204](https://doi.org/10.1109/ACCESS.2019.2946204).
- [12] S. Liu, L. Zhang, W. Liu, J. Zhang, D. Gao, X. Jia, The evaluation framework and benchmark for large language models in the government affairs domain, *ACM Transactions on Intelligent Systems and Technology* (2025).
- [13] V. Charles, N. P. Rana, L. Carter, Artificial intelligence for data-driven decision-making and governance in public affairs, 2022.
- [14] J. Yao, H. Zhang, J. Ou, D. Zuo, Z. Yang, Z. Dong, Social opinions prediction utilizes fusing dynamics equation with llm-based agents, *Scientific Reports* 15 (2025). URL: <http://dx.doi.org/10.1038/s41598-025-99704-3>. doi:[10.1038/s41598-025-99704-3](https://doi.org/10.1038/s41598-025-99704-3).
- [15] Q. Wang, G. Jin, X. Zhao, Y. Feng, J. Huang, Csan: A neural network benchmark model for crime forecasting in spatio-temporal scale, *Knowledge-Based Systems* 189 (2020) 105120.
- [16] A. Ivić, A. Milićević, D. Krstić, N. Kozma, S. Havzi, The challenges and opportunities in adopting ai, iot and blockchain technology in e-government: a systematic literature review, in: 2022 International Conference on Communications, Information, Electronic and Energy Systems (CIEES), IEEE, 2022, pp. 1–6.
- [17] T. S. Gesk, M. Leyer, Artificial intelligence in public services: When and why citizens accept its usage, *Government Information Quarterly* 39 (2022) 101704.
- [18] J. Romberg, T. Escher, Making sense of citizens' input through artificial intelligence: A review of methods for computational text analysis to support the evaluation of contributions in public participation, *Digital Government: Research and Practice* 5 (2024) 1–30.
- [19] M. Safaei, J. Longo, 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 (2024) 1–35.
- [20] S. Zheng, A. Trott, S. Srinivasa, D. C. Parkes, R. Socher, The ai economist: Taxation policy design via two-level

- deep multiagent reinforcement learning, *Science advances* 8 (2022) eabk2607.
- [21] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, *ACM transactions on intelligent systems and technology* 15 (2024) 1–45.
  - [22] Y. Li, M. Yu, B. Wang, L. Zhang, E. Dai, Y. Luo, L. Zhao, Application and implementation of superglue algorithm based on deep learning in image stitching, in: Proceedings of the 2025 6th International Conference on Computer Information and Big Data Applications, 2025, pp. 172–177.
  - [23] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al., Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, *Advances in Neural Information Processing Systems* 37 (2024) 95266–95290.
  - [24] E. Bugliarello, F. Liu, J. Pfeiffer, S. Reddy, D. Elliott, E. M. Ponti, I. Vulić, Iglue: A benchmark for transfer learning across modalities, tasks, and languages, in: *International Conference on Machine Learning*, PMLR, 2022, pp. 2370–2392.
  - [25] S. Chen, Y. Zhang, Q. Yang, Multi-task learning in natural language processing: An overview, *ACM Computing Surveys* 56 (2024) 1–32.
  - [26] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, Turingbench: A benchmark environment for turing test in the age of neural text generation, *arXiv preprint arXiv:2109.13296* (2021).
  - [27] I. Kumar, S. Viswanathan, S. Yerra, A. Salemi, R. A. Rossi, F. Dernoncourt, H. Deilamsalehy, X. Chen, R. Zhang, S. Agarwal, et al., Longlamp: A benchmark for personalized long-form text generation, *arXiv preprint arXiv:2407.11016* (2024).
  - [28] H. Li, Y. Chen, Q. Ai, Y. Wu, R. Zhang, Y. Liu, Lexeval: A comprehensive chinese legal benchmark for evaluating large language models, *Advances in Neural Information Processing Systems* 37 (2024) 25061–25094.
  - [29] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, J. Ge, Lawbench: Benchmarking legal knowledge of large language models, *arXiv preprint arXiv:2309.16289* (2023).
  - [30] N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano, et al., Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, *Advances in neural information processing systems* 36 (2023) 44123–44279.
  - [31] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, D. E. Ho, When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings, in: *Proceedings of the eighteenth international conference on artificial intelligence and law*, 2021, pp. 159–168.
  - [32] R. S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, D. Yang, When flue meets flang: Benchmarks and large pre-trained language model for financial domain, *arXiv preprint arXiv:2211.00083* (2022).
  - [33] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, Pixiu: A large language model, instruction data and evaluation benchmark for finance, *arXiv preprint arXiv:2306.05443* (2023).
  - [34] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, et al., Finben: A holistic financial benchmark for large language models, *Advances in Neural Information Processing Systems* 37 (2024) 95716–95743.
  - [35] X.-Y. Liu, Z. Xia, J. Rui, J. Gao, H. Yang, M. Zhu, C. Wang, Z. Wang, J. Guo, Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning, *Advances in Neural Information Processing Systems* 35 (2022) 1835–1849.
  - [36] J. Ogier du Terrail, S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, et al., Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings, *Advances in Neural Information Processing Systems* 35 (2022) 5315–5334.
  - [37] Y. Jin, M. Chandra, G. Verma, Y. Hu, M. De Choudhury, S. Kumar, Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries, in: *Proceedings of the ACM Web Conference 2024*, 2024, pp. 2627–2638.
  - [38] R. Hou, S. Chen, Y. Fan, G. Yu, L. Zhu, J. Sun, J. Liu, T. Ruan, Msdiagnosis: A benchmark and framework for evaluating large language models in multi-step clinical diagnosis, *Knowledge-Based Systems* (2025) 114524.
  - [39] A. Gjorgjevikj, A. Nikolicj, B. K. Seljak, T. Eftimov, User-defined trade-offs in llm benchmarking: balancing accuracy, scale, and sustainability, *Knowledge-Based Systems* (2025) 114405.
  - [40] G. DeepMind, Gemini 2.5 pro (preview 06-05): Model and api release notes, 2025. URL: REPLACE\_WITH\_OFFICIAL\_URL, model identifier: gemini-2.5-pro-preview-06-05.

- [41] Anthropic, Claude 3.7 sonnet system card, 2025. URL: <https://www.anthropic.com/news/visible-extended-thinking>, online; accessed 2025-10-09.
- [42] OpenAI, Openai o3 and o4-mini system card, <https://openai.com/index/o3-o4-mini-system-card/>, 2025. PDF: <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.
- [43] O. et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023). URL: <https://arxiv.org/abs/2303.08774>.
- [44] D. Guo, DeepSeek-AI, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025). URL: <https://arxiv.org/abs/2501.12948>.
- [45] DeepSeek-AI, Deepseek-r1-distill-llama-8b, <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>, 2025.
- [46] A. Yang, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: <https://arxiv.org/abs/2505.09388>.
- [47] DeepSeek-AI, Deepseek-v3 technical report, arXiv preprint arXiv:2412.19437 (2024). URL: <https://arxiv.org/abs/2412.19437>.
- [48] A. Grattafiori, A. Dubey, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). URL: <https://arxiv.org/abs/2407.21783>.
- [49] M. AI, Llama 3.1 — model cards and prompt formats, [https://www.llama.com/docs/model-cards-and-prompt-formats/llama3\\_1/](https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/), 2024.
- [50] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, S. R. Bowman, GPQA: A graduate-level google-proof q&a benchmark, in: First Conference on Language Modeling, 2024. URL: <https://openreview.net/forum?id=Ti67584b98>.
- [51] Q. Team, Qwq-32b: Embracing the power of reinforcement learning, <https://qwenlm.github.io/blog/qwq-32b/>, 2025. Accessed 2025-10-09.
- [52] Q. Team, Qwen/qwq-32b-awq, <https://huggingface.co/Qwen/QwQ-32B-AWQ>, 2025. Model card; Accessed 2025-10-09.

## Appendix A. EXPERIMENTAL SETUP

We adopt low-stochasticity, reproducible settings in the unified evaluation pipeline. In both the answer-generation stage and the judge–decision stage, we use identical inference hyperparameters: single-pass inference (no multi-sample averaging or voting), low-temperature sampling (temperature = 0.1), disabled log-probability return (logprobs = False), and max\_tokens set to each model’s maximum permitted output length. To reduce noise caused by interface fluctuations, both stages enable up to five failure retries (triggered upon exceptions or unparseable responses, stopping immediately upon success). Inputs strictly follow length constraints: over-long policy originals are already standardized and segmented during data construction, ensuring that the per-sample context does not exceed the processable window, thereby avoiding output truncation and preserving answer completeness and validity.

In the LLMJudge step, we use QwQ-32B [51, 52] as the judge model and apply a fixed few-shot template; the judge’s output is hard-constrained to the binary labels “[Correct]/[Incorrect]”. All samples and all three tasks (Clause Retrieval & Interpretation, Solution Generation, Compliance Judgment) are evaluated under the same prompts and hyperparameters, ensuring that scores across tasks and models are comparable and reproducible. These settings minimize the impact of sampling stochasticity without sacrificing the strictness of judgments.

## Appendix B. MORE EXPERIMENTAL RESULTS AND DISCUSSIONS

### Appendix B.1. Task-level Performance Distribution under Bilingual Policies

The following tables, Table 3 and Table 4, provide a detailed, task-wise breakdown of the models’ performance on the POLIS-Bench test set, disaggregated by language (Chinese and English). For each task—Clause Retrieval & Interpretation, Solution Generation, and Compliance Judgment—the quantitative results report both the mean semantic similarity and the LLM.Judge accuracy rate. These data further illustrate the cross-task and cross-language stability and fluctuations observed across the closed-source (Reasoning and Chat) and open-source (Reasoning and Chat) model families.

**Table B.4: Task-wise Results under Chinese Policies.** Quantitative results on **POLIS-BENCH (CN)** covering closed-source & open-source and reasoning & chat model families across three tasks. For each task, the table reports `semantic similarity` and `accuracy_rate` under a unified evaluation pipeline. Cells highlighted in red indicate the **highest** value within each metric column, and blue indicates the **second highest**.

Model	CN-Clause Retrieval & Interpretation		CN-solution generation		CN-compliance judgment	
	semantic similarity	accurate rate	semantic similarity	accurate rate	semantic similarity	accurate rate
<b>Closed-source Reasoning Model</b>						
Gemini-2.5-Pro-Preview-06-05	0.64	0.75	0.62	0.68	0.62	0.72
Claude-3.7-Sonnet-Thinking	0.64	0.65	0.58	0.73	0.59	0.65
o4-mini	0.66	0.54	0.60	0.63	0.61	0.59
o3-2025-04-16	0.63	0.66	0.58	0.70	0.59	0.64
<b>Closed-source Chat Model</b>						
Claude-3.7-Sonnet-Latest	0.71	0.63	0.71	0.45	0.69	0.59
GPT-4.1-20250414	0.69	0.68	0.63	0.58	0.65	0.59
<b>Open-source Reasoning Model</b>						
DeepSeek-R1	0.61	0.75	0.57	0.84	0.59	0.77
DeepSeek-R1-Distill-Llama-8B	0.59	0.38	0.56	0.40	0.58	0.38
Qwen3-8B	0.57	0.46	0.59	0.49	0.59	0.49
<b>Open-source Chat Model</b>						
Llama3.1-8B-Instruct	0.71	0.41	0.67	0.28	0.67	0.29
DeepSeek-V3	0.68	0.52	0.62	0.48	0.65	0.48

Table B.5: **Task-wise Results under English Policies.** Quantitative results on **POLIS-BENCH** (EN) covering closed-source & open-source and reasoning & chat model families across three tasks. For each task, the table reports `semantic similarity` and `accuracy_rate` under a unified evaluation pipeline. Cells highlighted in red indicate the **highest** value within each metric column, and blue indicates the **second highest**.

Model	EN-Clause Retrieval & Interpretation		EN-solution generation		EN-compliance judgment	
	semantic similarity	accurate rate	semantic similarity	accurate rate	semantic similarity	accurate rate
<b>Closed-source Reasoning Model</b>						
Gemini-2.5-Pro-Preview-06-05	0.78	0.69	0.71	0.70	0.73	0.70
Claude-3.7-Sonnet-Thinking	0.75	0.74	0.67	0.75	0.70	0.56
o4-mini	0.73	0.72	0.65	0.74	0.68	0.61
o3-2025-04-16	0.75	0.75	0.68	0.75	0.70	0.69
<b>Closed-source Chat Model</b>						
Claude-3.7-Sonnet-Latest	0.73	0.61	0.62	0.53	0.64	0.57
GPT-4.1-20250414	0.77	0.75	0.71	0.76	0.73	0.72
<b>Open-source Reasoning Model</b>						
DeepSeek-R1	0.75	0.72	0.70	0.76	0.70	0.76
DeepSeek-R1-Distill-Llama-8B	0.73	0.55	0.66	0.51	0.68	0.47
Qwen3-8B	0.74	0.63	0.69	0.57	0.70	0.54
<b>Open-source Chat Model</b>						
Llama3.1-8B-Instruct	0.79	0.50	0.71	0.45	0.73	0.48
DeepSeek-V3	0.78	0.51	0.68	0.60	0.70	0.55