GOVQABENCH: A Realistic Benchmark for Government Question Answering with Hierarchical Administrative Tasks

Anonymous ACL submission

Abstract

Despite the rapid progress of Large Language Models (LLMs), their systematic evaluation in government-domain question answering (QA) remains critically underexplored. Real-world government QA scenarios demand rigorously handling diverse intents from citizens, enterprises, and officials under strict compliance, where even a single inaccurate or non-compliant answer can undermine public trust. However, existing benchmarks are primarily citizen-centric and completely lack any assessment of safe abstention – the ability to refuse answering when responses may be harmful or out of scope. To bridge these pressing gaps, we present GovQABench, a comprehensive QA benchmark designed to reflect authentic administrative scenarios and enforce stringent safety constraints. Our GovQABench comprises 13,575 expert-annotated QA pairs spanning four core administrative tasks: *query deflection*, *policy interpretation*, *procedural guidance*, and *procedural reasoning*. The dataset is constructed from a multi-source knowledge infrastructure, including administrative policies, public service specifications, government office manuals, and a sensitive-vocabulary dictionary. Moreover, we propose GovMentor, an LLM-based agent that integrates multi-round reflection and hybrid retrieval to enhance model performance on government QA. Extensive experiments demonstrate GovMentor's effectiveness while revealing key limitations of LLMs, specifically their over-generation tendencies and sub-optimal performance in enterprise and government staff domains. Our dataset will be publicly available soon.

1 Introduction

Recently, Large Language Models (LLMs) have demonstrated remarkable capabilities across various specialized domains (Brown et al., 2020; Zhao et al., 2023). Despite these advancements, systematic exploration of LLMs within the government domain still lags behind practical demands. Currently, a stark mismatch exists between grassroots administrative capacity and the escalating complexity of public inquiries. From the perspective of citizens and enterprises, a substantial portion of the general public struggles to comprehend intricate administrative guidelines (Herd et al., 2023), leading to a severe surge in repetitive question-answering (QA) demands that frequently overwhelms public service hotlines (Jiang et al., 2025). Concurrently, frontline government staff, acting as the primary interface for public queries, are overburdened by a wide range of "generalist" administrative tasks (Bureaucracy, 1980).

To mitigate these issues, LLM-driven automated government QA assistants have emerged as a promising solution. Yet deploying LLMs in this domain presents several challenges. The core difficulty lies in satisfying the heterogeneous query intents from diverse user groups (citizens, enterprises, and government staff), while strictly maintaining safety and explainability (Atakishiyev et al., 2025; Dong et al., 2025). In government QA scenarios, even minor hallucinations or opaque decision-making paths can lead to catastrophic compliance risks and directly undermine public trust (Rumambi et al., 2026). However, existing benchmarks struggle to cover these critical challenges. They generally neglect the authentic QA scenarios of enterprises and officials, while some further oversimplify open-ended tasks into multiple-choice questions. Meanwhile, they all fail to evaluate vital safety mechanisms like query deflection—the model's ability to safely abstain when encountering unsafe, out-of-domain, or unanswerable queries.

Therefore, we construct GOVQABENCH, a comprehensive question-answering benchmark tailored for authentic government scenarios. To address the dual requirements of safety and practical question-answering capability, we organize GovQABench into two complementary subsets:

Table 1: **Comparison of GovQABench with prior government QA datasets**, including QualBench (Hong et al., 2025), LogiQA (Liu et al., 2020), PolicyBench (Bao et al., 2026), CitizenQuery (Majithia et al., 2026), MSGABench (Liu et al., 2025), POLIS-Bench (Yang et al., 2025b), ARGUS (Jiang et al., 2025), and ForeSightSafetyBench (Tong et al., 2026). MCQ and Gen. denote multiple-choice and open-ended generative QA, respectively. Gov.&Ent. indicates whether the dataset includes authentic scenarios for government staff and enterprises. Tasks are categorized into: Query Deflection (T1), Policy Interpretation (T2), Procedural Guidance (T3), and Procedural Reasoning (T4). GovQABench is constructed from the diverse perspectives of citizens, enterprises, and officials, featuring tasks that comprehensively reflect real-world government QA scenarios.

Dataset	#QA pairs	QA Format	Gov.&Ent.	Tasks			
				Deflection	Interpretation	Guidance	Reasoning
QualBench	17,316	MCQ	×	×	✓	×	✓
LogiQA	8,678	MCQ	×	×	×	×	✓
PolicyBench	20,935	MCQ	×	×	✓	✓	✓
CitizenQuery	22,066	Gen.	×	×	✓	✓	×
MSGABench	13,843	Gen.	×	×	✓	✓	✓
POLIS-Bench	3,058	Gen.	×	×	✓	×	✓
ARGUS	∼38,000	Gen.	×	×	×	✓	✓
ForeSightSafetyBench	—	Gen.	×	×	✓	×	✓
GovQABench (Ours)	13,575	Gen.	✓	✓	✓	✓	✓

GOVQA-DEFLECT and GOVQA-SOLVE. The former focuses on the safety boundaries by evaluating the model's ability to safely abstain from out-of-domain or sensitive queries, while the latter measures the core administrative capacity across three tasks of escalating complexity, including policy interpretation, procedural guidance, and procedural reasoning. In total, GovQABench comprises 13,575 expert-annotated, open-ended QA pairs. To guarantee professional alignment and absolute factuality, we first establish a vast, multi-source knowledge base as the authoritative ground truth, encompassing 10,937 policy documents, 3,191 public service specifications, and 495 government administrative manuals. Leveraging this robust knowledge base, we design a multi-agent annotation framework that simulates diverse, real-world personas across three critical dimensions: citizens, enterprises, and government staff. All generated instances undergo rigorous expert verification and manual refinement by veteran government personnel.

Building on our efforts to enhance models' performance on government question-answering tasks, we introduce GOVMENTOR, a specialized agent that incorporates a multi-round reflection mechanism for hybrid retrieval, thereby enabling LLMs to acquire high-quality policy knowledge. Our experiments have demonstrated its effectiveness in augmenting LLMs' ability.

Our key contributions are as follows:

- We propose GOVQABENCH, a comprehensive and expert-annotated benchmark containing 13,575 open-ended QA pairs tailored for authentic government scenarios, which addresses both safety boundaries via query deflection and core administrative capabilities across diverse personas.
- We introduce GOVMENTOR, a multi-stage question-answering framework for government scenarios. Our experiments reveal the effectiveness of GOVMENTOR.
- We conduct extensive baseline evaluations on LLMs, presenting empirical findings that reveal existing limitations and offer insights for future research.

2 Related Works

Government Question Answering. As illustrated in Table 1, early government QA tasks often simplify the problem to multiple-choice questions. QualBench (Hong et al., 2025) consists of multiple-choice questions on legal regulations, administrative procedures, and policy terminology drawn from Chinese professional qualification exams. LogiQA (Liu et al., 2020) contains multiple-choice questions from civil service exams. PolicyBench (Bao et al., 2026) mainly relies on multiple-choice questions about policy-related content. Subsequently, *open-ended* QA datasets have emerged. CitizenQuery (Majithia et al., 2026) extracts citizen queries from government websites and assesses response factuality via an adapted

FActScore (Min et al., 2023), while other frameworks like MSGABench (Liu et al., 2025), POLIS-Bench (Yang et al., 2025b), and ARGUS (Jiang et al., 2025) similarly target diverse public-facing scenarios. However, these existing datasets exclusively focus on procedural guidance for citizens, leaving the practical, day-to-day administrative requirements of government staff and enterprises heavily under-explored. While ForeSightSafety-Bench (Tong et al., 2026) evaluates answer safety, it neglects a critical capability: *query deflection*—the model's ability to safely abstain when encountering unsafe, out-of-domain, and unanswerable queries.

Retrieval-Augmented Generation. Traditional question answering systems heavily rely on matching user queries against FAQ libraries (Ji et al., 2012). With the rapid development of LLMs, retrieval-augmented generation (RAG) has been applied to improve factual accuracy (Sharma, 2025) (Wu et al., 2024). Naive RAG (Gao et al., 2023) relies on basic keyword-based retrieval from static corpora to augment LLM generation, offering a simple baseline with limited semantic awareness. Advanced RAG (Gao et al., 2023) introduces semantic understanding and improved retrieval techniques such as dense retrieval, reranking, and multi-hop retrieval, enhancing precision and contextual relevance. Hybrid RAG combines sparse retrieval methods with dense retrieval techniques (Karpukhin et al., 2020). Agentic RAG further incorporates autonomous agents, enabling the system to decompose complex queries and coordinate multi-step reasoning (Singh et al., 2025).

3 The proposed GovQABench

We present GovQABench (illustrated in Figure 1), a multi-task benchmark designed to capture the distinct operational requirements of three core user groups: citizens, enterprises, and government officials. Our benchmark is organized into two complementary evaluation tracks: GovQA-Deflect, which establishes the safety boundary, and GovQA-Solve, which assesses a gradient of execution and reasoning capabilities across three tasks.

3.1 Overview

Motivated by the *Guidelines for the Deployment and Application of Artificial Intelligence Large Models in the Government Domain* (CAC and NDRC, 2025), we formalize four tasks. Task 1 constitutes the GovQA-Deflect track, gating system deployment on boundary compliance. Tasks 2–4 form the GovQA-Solve track.

Task 1: Query Deflection. For *citizens* and *enterprises*, the model must reject out-of-domain functional queries and politically sensitive probes. For *government staff*, deflection targets data-security compliance: the system must refuse unauthorized requests for internal, classified, and non-public administrative materials while maintaining standardized refusal protocols.

Task 2: Policy Interpretation. This task evaluates the system's capacity to translate dense, authoritative legal texts into knowledge calibrated to the recipient's context. For *citizens*, the objective is to simplify obscure statutory provisions into intuitive, plain-language explanations. For *enterprises*, the objective is to interpret complex commercial regulations, compliance requirements, and fiscal subsidy eligibility criteria in terms of actionable business obligations.

Task 3: Procedural Guidance. This task targets inquiries regarding public and corporate services, requiring the generation of concrete, step-by-step guidance. For *citizens*, it covers personal lifecycle events such as healthcare enrollment and taxation. For *enterprises*, it focuses on operational lifecycles, including commercial registration, environmental impact assessments, and licensing workflows. For *government staff*, the task serves a dual purpose that mirrors their intermediary role: (a) providing factual references to help personnel resolve inquiries from citizens and enterprises, and (b) generating standard operating procedures for internal administrative workflows (e.g., official document processing and information dissemination).

Task 4: Procedural Reasoning. This task demands logical inference over interlocking administrative criteria. For *citizens* or *enterprises*, it verifies whether they satisfy specific policy thresholds prior to application. For *government staff*, it requires the system to perform automated completeness and compliance reviews on submitted applicant materials.

3.2 GovQA-Deflect Dataset

We construct GovQA-Deflect, a high-quality dataset encompassing three representative rejection scenarios in administrative contexts. Each instance is produced through a generation–verification pipeline, with full implementation details provided in the Appendix.

Figure 1: **Overview of the GovQA dataset.** It consists of two subsets: (1) GovQA-Deflect, featuring the query deflection (Task 1) to evaluate the safety of government QA systems; and (2) GovQA-Solve, comprising policy interpretation (Task 2), procedural guidance (Task 3), and procedural reasoning (Task 4) to assess answer generation quality. Configured across realistic scenarios tailored for citizens, enterprises, and government staff, GovQA provides a comprehensive benchmark for both the capability and safety of domain-specific large language models.

3.2.1 Data Acquisition

We construct a sensitive keyword lexicon grounded in the following three representative rejection scenarios:

Sensitive Compliance Queries are utterances that contain unauthorized requests for internal, classified, or non-public administrative materials, illegal content, or politically sensitive material.

Adversarial Security Probe Queries attempt to elicit confidential internal information via sophisticated maneuvers, such as role-playing, instruction hijacking, soft persistence, or encoding-based obfuscation.

Out-of-Domain Functional Queries refer to general questions that fall outside the designated government service scope, included to ensure that operational boundaries are maintained.

3.2.2 Data Annotation

We use DeepSeek-V3 (Liu et al., 2024) to generate one user query for each entry in the dictionary, conditioned on a scenario description, so that every query incorporates or alludes to its corresponding entry. For sensitive compliance queries, the model is prompted to mimic citizen inquiries directed at a government portal; for adversarial security probe queries, the attack technique is randomized per sample.

Subsequently, we conduct data verification to guarantee high data quality. We invite three senior government practitioners (average 9 years of public service experience). Verification proceeds in two stages. First, annotators judge whether each question genuinely falls within the intended rejection scenario. Samples that are ambiguous, non-rejectable under stated policy, or ethically inappropriate are discarded. Then, remaining samples are rated on a 1–5 linguistic fluency scale. We retain only high-quality questions that achieve an average score of 4 or higher from a minimum of two independent annotators.

3.3 GovQA-Solve Dataset

We present GovQA-Solve, a comprehensive dataset curated from authoritative multi-source documents to simulate complex government-

citizen/enterprise and internal administrative interactions. The construction process integrates an automated multi-agent pipeline with rigorous manual annotation to ensure high-quality of the final QA pairs.

3.3.1 Data Acquisition

With the support of relevant government authorities, we curated a comprehensive, multi-source collection of documents that serves as the foundation for generating high-quality question-answering (QA) pairs. To ensure the authoritativeness and breadth of the data, we integrate information from the following three primary categories:

Administrative Policy Documents: We retrieve a collection of 10937 administrative policy documents official interpretations from an official policy disclosure platform in China. All data acquisition is conducted with institutional permission from the platform.

Public Service Specifications: We curate data from provincial-level government portals comprising 3191 distinct service items, spanning domains such as household registration, taxation, and business licensing. These specifications provide granular operational metadata—including eligibility criteria, standardized workflows, and required documentation—offering the requisite grounding for modeling citizen-centric service consultations.

Government Administrative Manuals: To capture staff-oriented operational logic, we incorporate 495 user manuals from a government collaborative office platform in China. These resources provide detailed functional instructions for messaging, document management, and scheduling systems.

Collectively, these resources constitute a robust knowledge infrastructure, covering both public-facing services and government administrative workflows in the government domain.

3.3.2 Data Annotation

First, we present a multi-agent automatic annotation framework, capable of simulating real-world government service interactions. This framework comprises three specialized agents organized in a sequential pipeline: the *Simulator*, the *Synthesizer*, and the *Validator*. The *Simulator* generates diversified user personas, including different citizens, enterprise employees, and government staff. Leveraging the persona, predefined task category, and reference content from the document collection, the *Synthesizer* produces candidate QA pairs. Finally, the *Validator* filters these pairs based on relevance to output high-quality QA pairs.

All automatically generated QA pairs undergo a rigorous manual review stage to guarantee data quality. Given the distinct nature of external (citizen/enterprise) and internal (government staff) use cases, we design a dual-track review pipeline.

For Citizen and Enterprise-Facing QA, our review team comprises volunteer government staff with at least 3 years of frontline service experience and citizen/enterprise reviewers. The process proceeds in three phases: (1) Policy experts verify factual accuracy against the supplied evidence, and supplement missing information. (2) Citizen/enterprise reviewers rate the question's naturalness and the answer's understandability on a 5-point scale. Samples scoring below 4 on either dimension are returned for revision.(3)Returned samples are jointly revised by one expert and one lay reviewer; the lay reviewer proposes plain-language edits, while the expert ensures factual integrity.

For Government staff QA, we select volunteer annotators from government official members with at least 3 years of working experience. Annotators verify every procedure of the sequence of actions, decision logic, or coordination plan described in the answer, strictly cross-referencing each step with the manual's specifications. Any content inconsistent with the manuals is revised accordingly.

3.4 Data Statistics

Our GovQA dataset covers four tasks: query deflection, policy interpretation, procedural guidance, and procedural reasoning. It is divided into two subsets: GovQA-Deflect contains 3,182 QA pairs for query deflection, while GovQA-Solve covers 10,393 QA pairs for the remaining three tasks. Detailed statistics are presented in Table 2.

Table 2: Data statistics of our GovQA dataset. #QA Pairs denotes the number of question-answer pairs; Len.(Q) is the average word count of questions; Len.(A) is the average word count of answers.

Task	#QA Pairs	Len.(Q)	Len.(A)
Query Deflection	3,182	8.9	8.1
Policy Interpretation	5,285	13.1	44.5
Procedural Guidance	3,101	13.2	60.2
Procedural Reasoning	2,007	15.4	64.1
Overall	**13,575**	**12.5**	**42.4**

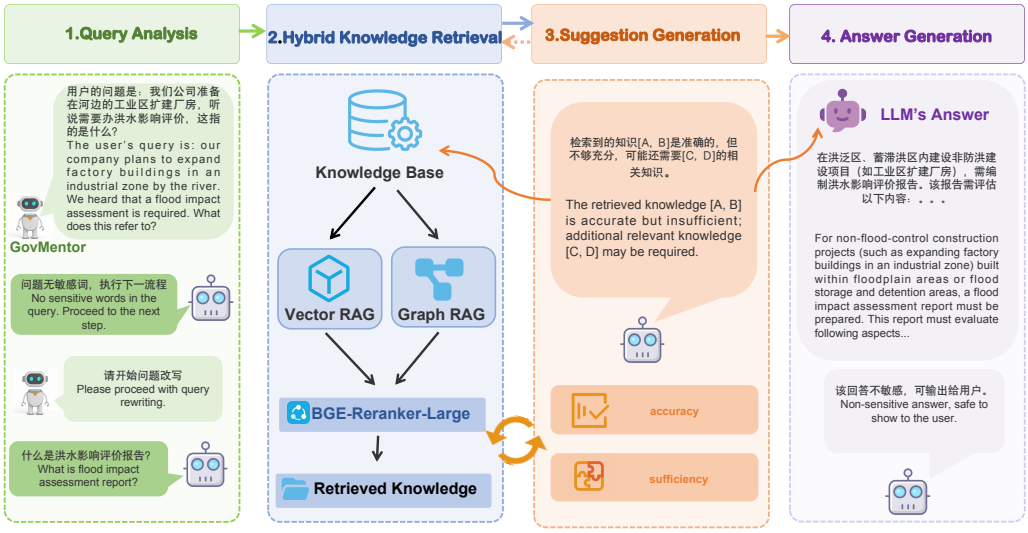

Figure 2: **The pipeline of GovMentor.** To generate the final response, GovMentor sequentially executes four core phases: (1) query analysis, (2) hybrid knowledge retrieval, (3) suggestion generation, and (4) answer generation and verification.

4 Evaluation

4.1 Model Zoo

We select a diverse set of state-of-the-art large language models (LLMs) spanning different architectures, scales, and development philosophies. Our model zoo comprises seven representative models: DeepSeek-V4-Pro (DeepSeek-AI, 2026), DeepSeek-R1-0528 (Guo et al., 2025), Qwen3.5-35B-A3B (Qwen Team, 2026), GLM-5 (Zeng et al., 2026), MiniMax-M2.5 (MiniMax, 2026), Kimi-K2.5 (Team et al., 2026), and Gemini3-Flash (Google DeepMind, 2025). Table 3 provides a summary of their key characteristics.

Table 3: Overview of the selected models in our benchmark. "#Total Params" and "#Active Params" denote total and per-token activated parameters, respectively.

Model	#Total Params	#Active Params
DeepSeek-V4-Pro	1.6T	49B
DeepSeek-R1-0528	671B	37B
Qwen3.5-35B-A3B	35B	∼3B
GLM-5	744B	40B
MiniMax-M2.5	∼230B	10B
Kimi-K2.5	∼1T	32B
Gemini3-Flash	–	–

4.2 GovMentor

We design GovMentor, a dedicated question-answering framework for government affairs. Since administrative question-answering heavily depends on the retrieval of relevant policy documents and requires rigorous security guarantees, GovMentor incorporates a multi-round reflection mechanism for hybrid retrieval results and performs safety checks at different stages. As illustrated in Figure 2, our framework consists of four components:

Query Analysis. Upon receiving an incoming query, the system first conducts an initial safety test. If the query contains harmful or sensitive content, the system triggers a deflection mechanism and refuses to answer. For compliant queries, it rewrites the query appropriately to optimize downstream retrieval performance.

Hybrid Knowledge Retrieval. Guided by the refined query, the system executes a hybrid retrieval strategy over the underlying knowledge base, combining Vector RAG for dense semantic matching and Graph RAG for structured relational tracking. The retrieved documents are then re-ranked using BGE-Reranker-Large (BAAI, 2023) to filter out the most relevant reference context.

Suggestion Generation. To guarantee the reliability of the evidence, an intermediate verification loop is introduced to assess the quality of the retrieved documents. A dedicated LLM acts as an evaluator to check the accuracy and adequacy of the retrieval results. If the context is deemed inaccurate or insufficient, a feedback signal triggers query rewriting followed by iterative re-retrieval, repeating this cycle until the evidence meets the

Table 4: **Overall performance on the proposed GovQABench.** Acc. (Accuracy) measures the success rate of query deflection. Prec. (Precision) indicates the correctness of the generated answers, while Rec. (Recall) indicates their completeness.

Models	GovQA-Deflect	GovQA-Solve					
		Policy Interpretation		Procedural Guidance		Procedural Reasoning	
	Acc.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Deepseek-V4-Pro	**91.4**	51.9	62.3	53.2	65.2	52.8	63.3
Deepseek-R1-0528	83.4	48.5	58.7	48.7	61.0	48.7	58.5
Qwen3.5-35B-A3B	72.1	44.6	56.0	46.2	57.5	47.1	58.8
GLM-5	79.5	54.0	66.7	**55.2**	69.1	**56.9**	69.8
Minimax-M2.5	84.6	48.1	56.5	48.4	55.9	49.2	56.0
Kimi-K2.5	80.8	53.6	66.9	52.6	66.3	54.7	67.6
Gemini3-Flash	66.0	**56.1**	**70.5**	54.2	**69.2**	56.5	**70.8**

required standards.

Answer Generation and Verification. The generation model first uses the verified knowledge from the previous stage and the original user query to generate an answer. Subsequently, this draft undergoes a secondary safety inspection to ensure no sensitive or non-compliant content has been generated. The final authoritative answer is delivered to the user only after passing this post-generation check.

4.3 Metrics

Deflection Evaluation. To evaluate system safety, we formulate the query deflection task as a binary classification problem: determining whether the system correctly refuses to answer a query that should be deflected. We employ Accuracy as the evaluation metric, defined as the success rate of query deflection on the GovQA-Deflect dataset.

Generation Evaluation. Traditional n-gram based metrics often fail to capture semantic correctness in legal and policy contexts. Drawing inspiration from (Majithia et al., 2026), we adopt an Atomic Fact Verification (AFV) approach. We utilize Qwen3-32B (Yang et al., 2025a) with domain-specific prompts to decompose both the generated responses and ground-truth references into sets of independent atomic facts. The final quality is quantified by **Precision** and **Recall** computed over these atomic fact sets. Precision specifically evaluates the correctness of the generated text, while Recall assesses the completeness of the generated text.

5 Result

5.1 Main Result

Table 4 presents the overall performance of the evaluated large language models across different tasks on GovQABench. Based on the empirical results, we high-light several findings.

Table 5: GovQA-Solve performance across various personas (X/X: Precision/Recall).

Models	Citizens	Enterprises	Gov. Staff
Deepseek-V4-Pro	57.8/68.4	50.5/62.0	51.6/63.2
Deepseek-R1-0528	53.7/65.0	46.9/57.7	48.4/58.8
Qwen3.5-35B-A3B	50.5/61.7	44.0/55.7	44.8/56.3
GLM-5	60.7/73.0	53.0/66.4	53.6/67.0
Minimax-M2.5	54.5/62.0	46.4/54.5	48.5/55.6
Kimi-K2.5	60.5/72.8	51.3/65.2	53.2/66.9
Gemini3-Flash	**62.1/75.3**	**53.1/68.5**	**54.3/69.2**

Divergent Performance in Query Deflection. On the GovQA-Deflect task, the evaluated models exhibit a pronounced performance divergence. Notably, Deepseek-V4-Pro (DeepSeek-AI, 2026) achieves a dominant Accuracy (Acc.) of 91.4%, outperforming the runner-up Minimax-M2.5 (MiniMax, 2026)by nearly 7 percentage points, whereas Gemini3-Flash (Google DeepMind, 2025) underperforms significantly at 66.0%. This indicates a stark disparity among current LLMs in handling data security compliance. While proprietary models like Deepseek-V4-Pro demonstrate robust compliance auditing and sensitive information interception capabilities, safety alignment and rigorous boundary detection remain a significant vulnerability for other established baselines. This disparity highlights the need for future efforts to investigate more robust safety-alignment paradigms tailored for government QA security.

Government staff and enterprise scenarios pose greater challenges. As shown in Table 5, a consistent performance pattern emerges across all evaluated models: public-facing Citizens scenarios yield the highest scores, as routine inquiries typically rely on standardized, publicly accessible

knowledge. Conversely, performance drops sharply when moving to the Enterprises and Government Staff scenarios. This decline underscores that intricate corporate compliance and internal governmental workflows pose greater challenges than citizen services.

Systematic High-Recall, Low-Precision Discrepancy. Across all sub-tasks within GOVQA-SOLVE, we observe a universal phenomenon: Recall consistently outperforms precision. For instance, Gemini3-Flash (Google DeepMind, 2025) achieves a notable recall of 70.5% in the Policy Interpretation task, whereas its precision lags behind at 56.1%. This systematic gap highlights a common bottleneck in current large language models (LLMs): In pursuit of response completeness, LLMs exhibit a tendency to over-generate redundant and loosely relevant policy clauses or administrative guidelines. While this aggressive retrieval strategy artificially inflates recall, it severely compromises the precision and conciseness of the generated answers.

5.2 Result of GovMentor

We evaluate GovMentor against two baselines under the same Qwen3.5-35B-A3B backbone (Qwen Team, 2026), selected for its favorable trade-off between capability and deployment efficiency, which is critical for latency-sensitive and confidentiality-demanding government services. The three compared settings are: (1) Raw-LLM, which directly answers without any retrieval and serves as a control; (2) Naive RAG, a standard retrieval-augmented generation pipeline that shows the value of external knowledge; and (3) GovMentor, our proposed method.

The results reveal that compared with Raw-LLM, GovMentor brings substantial gains in both answer completeness and accuracy across all three core tasks: policy interpretation, procedural guidance, and procedural reasoning. In particular, on the policy interpretation task, precision and recall improve by 14.9 and 10.4 percentage points, respectively. These improvements demonstrate that GovMentor effectively enhances the factual correctness and coverage of generated responses, validating its design for knowledge-intensive government affairs.

Furthermore, we compare GovMentor with the Naive RAG baseline. Across all tasks and all evaluation metrics, GovMentor consistently outperforms Naive RAG, without exception. This comprehensive advantage confirms that the specific design choices in GovMentor lead to more accurate and complete responses than a standard RAG pipeline, reinforcing the effectiveness of our method for government service scenarios.

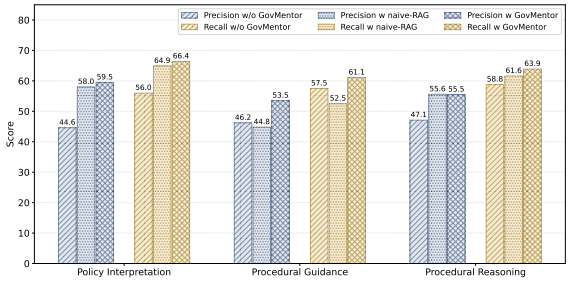

Figure 3: Precision and recall of GovMentor compared to the Qwen3.5-35B-A3B baseline on policy interpretation, procedural guidance, and procedural reasoning.

6 Conclusion

In this paper, we introduced GovQABench, a comprehensive, expert-annotated benchmark specifically designed for government-domain question answering. By bridging the gaps in existing benchmarks, GovQABench uniquely incorporates diverse query intents across multiple personas (citizens, enterprises, and government staff) and enforces a strict safety boundary via the query deflection task. To address the intense factual and compliance demands of administrative scenarios, we also developed GovMentor, a specialized multi-agent framework featuring hybrid retrieval and a multi-round reflection loop.

Our extensive baseline evaluations and experimental results demonstrate that while GovMentor significantly improves factual precision and retrieval reliability, state-of-the-art LLMs still face distinct bottlenecks, such as over-generation tendencies and suboptimal performance in complex professional domains. We hope that GovQABench serves as a valuable asset and a realistic testing ground to inspire further research toward building more reliable, data-secure, and capable language agents for public service.

Limitations

Our knowledge infrastructure and expert annotations are currently grounded within the administrative framework, policy documents, and public service specifications of China. Consequently, the benchmark evaluates models primarily in the Chinese language. A valuable direction for future work

would be extending this evaluation framework to cross-lingual contexts and diverse international administrative systems.

Government policies, administrative manuals, and public service protocols are dynamic by nature, subject to frequent institutional revisions and localized updates. Although our multi-source knowledge base is highly authoritative and comprehensive, it inevitably represents a static snapshot of regulations up to a specific timeframe. Developing mechanisms to evaluate a model's resilience against obsolete or updated policies remains an open and critical challenge for future research.

Ethical considerations

The construction of GovQABench relies entirely on publicly accessible administrative policies, public service specifications, and government collaborative office manuals obtained with explicit institutional permission. Throughout the data acquisition and curation process, we strictly adhered to relevant statutory regulations, ensuring that no personally identifiable information (PII), proprietary commercial secrets, or classified national security data were included. This dataset is made available under the Creative Commons Attribution 4.0 International License (CC BY 4.0) . By using this dataset, you agree to the terms of this license, which allows for redistribution, adaptation, and commercial use, provided appropriate credit is given to the original authors and a link to the license is provided.

While LLM-driven government assistants (such as our proposed GovMentor) can significantly alleviate administrative burdens , they remain susceptible to hallucinations inherent to large language models. In rare instances, this could lead to inaccurate policy interpretations and subsequent compliance risks. To balance efficiency and safety, practical deployments should rely on automated guardrails (such as the query deflection mechanism investigated in this paper) to intercept non-compliant requests. Furthermore, by incorporating clear disclaimers or verification prompts in the outputs, the system can guide users to double-check high-risk or ambiguous borderline cases through official channels, ensuring safe deployment without requiring manual review for every response.

References

Shahin Atakishiyev, Housam KB Babiker, Jiayi Dai, Nawshad Farruque, Teruaki Hayashi, Nafisa Sadaf Hriti, Md Abed Rahman, Iain Smith, Mi-Young Kim, Osmar R Zaïane, and 1 others. 2025. Explainability of large language models: Opportunities and challenges toward generating trustworthy explanations. *arXiv preprint arXiv:2510.17256*.

BAAI. 2023. Bge-reranker-large. Cross-Encoder reranker model based on XLM-RoBERTa.

Han Bao, Penghao Zhang, Yue Huang, Zhengqing Yuan, Yanchi Ru, Rui Su, Yujun Zhou, Xiangqi Wang, Kehan Guo, Nitesh V Chawla, and 1 others. 2026. Policyllm: Towards excellent comprehension of public policy for large language models. *Findings of ACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Street-Level Bureaucracy. 1980. Dilemmas of the individual in public services. *New York: Russell Sage Foundation*.

CAC and NDRC. 2025. Government domain ai large model deployment and application guidelines. https://www.cac.gov.cn/2025-10/10/c_1761819469929310.htm. Accessed: 2026-05-18.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, and 1 others. 2025. Safeguarding large language models: A survey. *Artificial intelligence review*, 58(12):382.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Google DeepMind. 2025. Gemini 3 flash. https://deepmind.google/technologies/gemini/flash/.

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

Pamela Herd, Hilary Hoynes, Jamila Michener, and Donald Moynihan. 2023. Introduction: Administrative burden as a mechanism of inequality in policy

implementation. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 9(4):1–30.

Mengze Hong, Wailing Ng, Chen Jason Zhang, and Di Jiang. 2025. Qualbench: Benchmarking chinese llms with localized professional qualifications for vertical domain evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5949–5964.

Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer topic model for question retrieval in community question answering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2471–2474.

Song Jiang, Xiaofeng Xie, Rongnian Tang, Xuanqi Wang, Kaihao Sun, Guanghan Li, Zhenkai Xu, Peng Xue, Ziling Li, and Xuedong Fu. 2025. Argus: Retrieval-augmented qa system for government services. *Electronics*, 14(12):2445.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

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*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Shuo Liu, Lin Zhang, Weidong Liu, Jianfeng Zhang, Donghui Gao, and Xiaofeng Jia. 2025. The evaluation framework and benchmark for large language models in the government affairs domain. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–24.

Neil Majithia, Rajat Shinde, Zo Chapman, Prajun Trital, Jordan Decker, Manil Maskey, Elena Simperl, and Nigel Shadbolt. 2026. The citizenquery benchmark: A novel dataset and evaluation pipeline for measuring llm performance in citizen query tasks. *arXiv preprint arXiv:2602.04064*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

MiniMax. 2026. Minimax m2.5: Built for real-world productivity.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5.

Frendy Rocky Rumambi, Didik Dwi Prasetya, and Triyanna Widiyaningtyas. 2026. Groundedness-aware retrieval in government document chatbots: Systematic literature review and semantic alignment score (sas) formulation. *ITEGAM-JETIA*, 12(58):242–257.

Chaitanya Sharma. 2025. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv preprint arXiv:2506.00054*.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Haibo Tong, Feifei Zhao, Linghao Feng, Ruoyu Wu, Ruolin Chen, Lu Jia, Zhou Zhao, Jindong Li, Tenglong Li, Erliang Lin, and 1 others. 2026. Foresight-safety bench: A frontier risk evaluation and governance framework towards safe ai. *arXiv preprint arXiv:2602.14135*.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and 1 others. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tingyue Yang, Junchi Yao, Yuhui Guo, and Chang Liu. 2025b. Polis-bench: Towards multi-dimensional evaluation of llms for bilingual policy tasks in governmental scenarios. *arXiv preprint arXiv:2511.04705*.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Xu, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

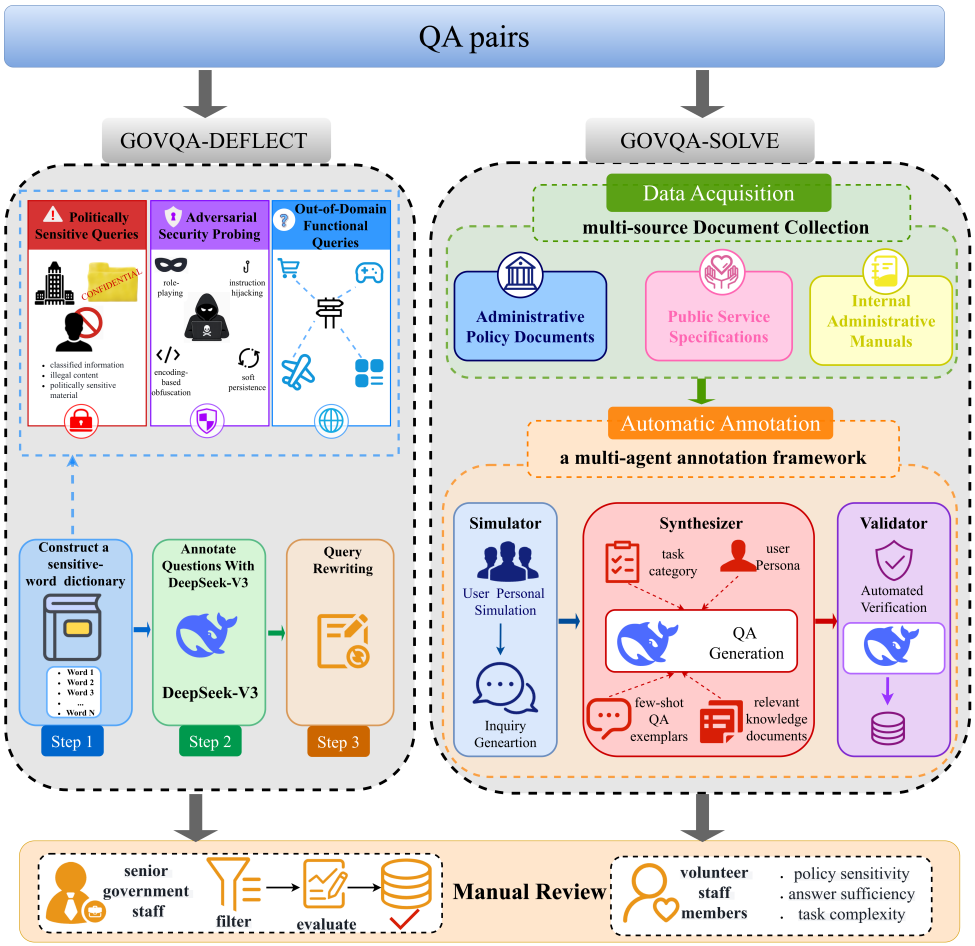

Figure 4: Overview of our multi-agent automatic annotation pipeline.

A Data Annotation

A.1 GovQA-Deflect

To guarantee the professionalism and authority of the evaluation, we invited **3 senior public servants** with an average of **9 years of experience** in public service to form an expert review panel. Selecting experts with a strong public administration background ensures precise, realistic, and politically/ethically sound judgments regarding sensitivity, policy boundaries, and rejection strategies within government-citizen interaction scenarios.

The data verification workflow was strictly executed in two sequential stages:

Stage 1: Scenario Rejection Alignment. In this phase, annotators independently evaluated the "rejectability" of each test instance. The primary goal was to determine whether a prompt truly fell into the predefined "Government Rejection Scenario" categories. An instance was **discarded** if it met any of the following criteria:

- **Ambiguous**: The formulation of the prompt was too vague to infer a clear intention.
- **Non-rejectable under stated policy**: According to current regulations, the query belongs to the scope of proactive government information disclosure and should not be deflected.
- **Ethically problematic**: The prompt involved extreme sensitivities, illegal activities, or violations of academic ethical standards.

Stage 2: Linguistic Fluency Assessment. Instances that successfully passed the Stage 1 filtering entered the linguistic quality evaluation phase. Annotators scored the fluency of the prompts based on a 1-to-5 Likert scale:

- **5 (Very Fluent)**: Natural sentence structures perfectly matching human expression habits, with zero grammatical issues or typos.
- **4 (Fluent)**: Clear expression with minor flaws (e.g., slight punctuation or word order issues) that do not affect overall comprehension.

- **3 (Barely Readable)**: Obvious grammatical errors, clear machine translation artifacts, or strong L1 interference, though the core intent remains guessable.

- **2 (Hard to Understand)**: Chaotic word order and severe grammatical errors, requiring repetitive reading to comprehend.

- **1 (Completely Unreadable)**: Randomly stacked words with absolutely no comprehensible meaning.

For a candidate instance to be finalized in the GOVQA-DEFLECT dataset, it had to simultaneously satisfy two rigid conditions:

1. **Consensus on Rejection**: In Stage 1, all 3 annotators must unanimously agree that the instance represents a valid and compliant rejection scenario.

2. **Quality Threshold**: In Stage 2, the instance must achieve an average fluency score of ≥ 4.

A.2 GovQA-Solve

A.2.1 Automatic Annotation

We apply multi-agent system, including simulator, synthesizer and validator to automatically annotate the data.

Simulator. To mirror the complexity of real-world interactions, we instantiate the Simulator with a diverse set of user personas—including citizens, enterprise employees, and government officials—parameterized by a broad range of attributes (e.g., needs, age, occupation, enterprise type, departmental role, among others). By conditioning the generation process on these specific profiles, the Simulator is designed to produce inquiries that better approximate the linguistic nuances and intent distributions found in real administrative contexts.

Synthesizer. The Synthesizer serves as the core generative component for producing question-answer (QA) pairs. It operates on four primary inputs: a predefined task category (e.g., policy interpretation, procedural guidance, or procedural reasoning), a specific user persona generated by User Agent, few-shot QA exemplars, and relevant knowledge documents. Utilizing DeepSeek-V3 (Liu et al., 2024) as the backbone LLM, the agent synthesizes questions that strictly adhere to the designated persona and task constraints, while concurrently generating precise answers grounded exclusively in the provided reference materials to mitigate hallucination.

Validator. To ensure data quality, we implement an automated verification mechanism. The Validator assesses the relevance of each generated pair and discards instances where the response fails to address the inquiry, retaining only those exceeding a predefined threshold.

A.2.2 Manual Annotation

To guarantee the rigorous evaluation of Large Language Models (LLMs) on administrative tasks, we developed a comprehensive annotation protocol. This appendix delineates the specific instructions, evaluation metrics, and quality control pipelines for two primary tasks: *Citizen and Enterprise-Facing QA* and *Government Staff QA*.

Citizen and Enterprise-Facing QA: The annotation team for this task comprised government volunteers with over three years of frontline public service experience, alongside a diverse pool of citizen and enterprise reviewers. The pipeline strictly advanced through three consecutive phases.

First, Policy experts were mandated to conduct a line-by-line verification of the generated answers against official policy documents and legal interpretations. For every sample, experts cross-checked the following criteria:

- ☐ **Evidence Consistency:** Are all core facts, statistical data, statutory processing timeframes, and legal foundations in the answer explicitly corroborated by the provided reference materials?

- ☐ **Information Completeness:** Does the response fully resolve the user's query? Are there any omissions of critical prerequisites, exclusionary clauses, or restrictive conditions that might mislead the public?

- ☐ **Granular Contextual Supplementation:** If the policy implementation involves implicit caveats (e.g., tax exemptions restricted to specific demographics or niche industrial qualifications), the expert must supplement these details into the answer to ensure real-world utility.

Then, to assess the stylistic and linguistic pragmatic standard of the dataset, citizen and enterprise reviewers evaluated the generated instances across two dimensions: **Question Naturalness** and **Answer Understandability**. Scores were assigned

using a 5-point Likert scale as formalized in Table 6.

To ensure rating reliability, we monitored the Inter-Annotator Agreement (IAA) by computing Cohen's κ (for pair-wise checks) and Krippendorff's α. Samples with low agreement ($\alpha < 0.70$) were reassigned for blind re-evaluation.

Threshold Rule for Rewriting: Any sample that received a score below **4** in either *Question Naturalness* or *Answer Understandability* was immediately rejected and routed to Phase 3 for iterative refinement.

Finally, rejected samples were processed by a collaborative pair consisting of one policy expert and one citizen/enterprise reviewer, adhering to the following division of responsibilities:

1. **Citizen/Enterprise-Led Plain-Language Editing:** The public reviewer identified dense bureaucratic phrasing and inaccessible jargon, proposing natural, conversational alternatives.

2. **Expert-Led Fact-Checking Control:** The policy expert supervised the rewriting process to guarantee that plain-language adaptations did not dilute the statutory precision or introduce legal ambiguities regarding policy boundaries.

Government Staff QA: The annotation for internal administrative operational guidance was performed exclusively by active public servants with at least three years of executive experience. Given the critical requirement for rigorous procedural flow and cross-departmental coordination, we implemented a **Dynamic Operational Walkthrough** protocol.

First, Annotators were required to mentally adopt the persona of an execution-level clerk, executing a closed-loop simulation of the workflow prescribed by the answer against official Standard Operating Procedures (SOPs):

- ☐ **Action Sequence Verification:** Does the chronological order of operational steps (e.g., internal review → public disclosure → license issuance) strictly mirror the official SOP? Are there any structural inversions?

- ☐ **Decision Logic Validation:** For complex workflows branching on specific conditions (e.g., "If corporate assets exceed threshold X, route to Pathway A; otherwise, route to Pathway B"), are the decision nodes, quantitative thresholds, and admission criteria flawlessly accurate?

- ☐ **Coordination Plan Confirmation:** For tasks requiring inter-departmental collaboration (e.g., coordination among taxation, industrial/commercial administration, and environmental bureaus), are the designated leading agencies, supporting roles, and data-sharing timelines accurately represented?

During the walkthrough, if even a minor divergence or procedural discrepancy was identified between the generated text and official manuals, annotators bypass the rejection loop. Instead, they performed **direct, authoritative corrections** in the backend repository. This mechanism enforces that the finalized dataset serves as an absolute, authoritative *Gold Standard* for government operational benchmarks.

B Data Statistics

Figure5 illustrates the text length distributions for questions and answers across the four task categories. Table and Table provide detailed summary statistics measured by Chinese word count (segmentation).

Question Length. Questions are relatively concise, with an overall mean of 12.5 words (median 12, standard deviation 4.5, range 3–42). However, notable variation exists across categories. Query deflection questions are the shortest (mean 8.9, median 9). Procedural reasoning questions are the longest (mean 15.4, median 15), as these scenario-based questions must describe multiple conditions or situational details to support conditional reasoning. Policy interpretation and procedural guidance questions fall in between (means 13.1 and 13.2, respectively).

Answer Length. Answers exhibit substantially greater variation, with an overall mean of 42.4 words (median 36, standard deviation 32.8, range 8–543). The per-category differences are pronounced and closely tied to the functional requirements of each task type. Query deflection answers are dramatically shorter (mean 8.1, median 8), consisting primarily of brief, polite refusal statements. By contrast, procedural guidance and procedural reasoning answers are the longest (means 60.2 and 64.1, respectively), as these tasks require detailed

Score	Question Naturalness	Answer Understandability
5	Fully aligns with authentic colloquial expressions of real users. Devoid of translation artifacts or artificial styling.	Exceptionally lucid and accessible. Uses plain language completely free of obscure bureaucratic jargon; immediately comprehensible to the general public.
4	Clear and natural expression. Though slightly formal in tone, it perfectly reflects queries typically posed at a physical public service window.	Fluent and well-structured. Contains minimal necessary statutory terms, but maintains clear underlying logic. Accessible to citizens upon a single reading.
3	Contains minor grammatical flaws or exhibits a rigid, unidiomatic phrasing, though the underlying intent remains discernible.	High density of jargon with a pronounced bureaucratic undertone. Difficult for ordinary citizens to digest; requires repeated reading.
2	Evidently generated from templates or severely ambiguous. Highly unnatural and misaligned with human communication habits.	Overloaded with unexplained acronyms or dense professional legalese. Disorganized information that is entirely incomprehensible to non-experts.
1	Chaotic word order, incoherent phrasing, and completely unreadable. Fails to convey any meaningful or interpretable query intent.	Completely irrelevant to the question, logically fractured, or exhibits severe grammatical impediments and common-sense errors.

Table 6: The 5-point Likert scale utilized for human evaluation of Language Quality.

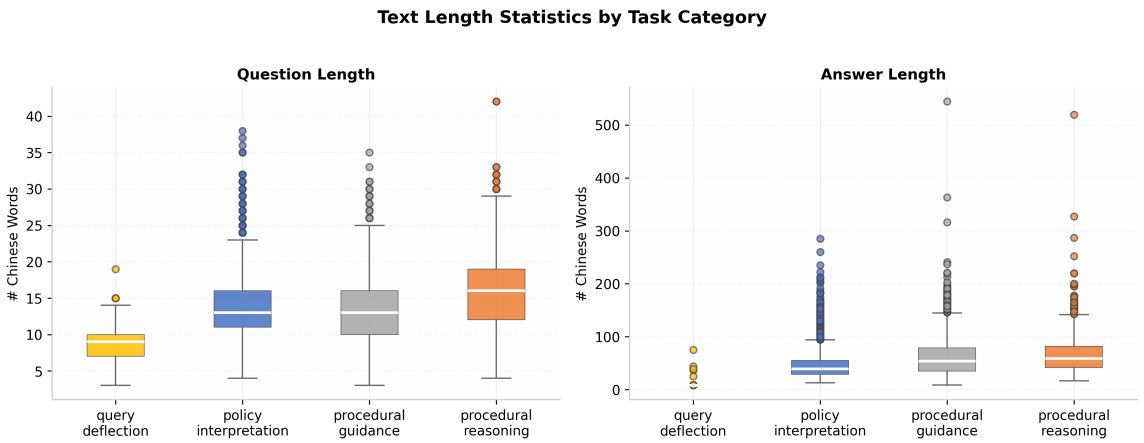

Figure 5: Detailed Statistics of our GovQA dataset.

step-by-step instructions or multi-step reasoning chains. Policy interpretation answers occupy an intermediate position (mean 44.5), as they must explain policy provisions without necessarily enumerating procedural steps.

C Implementation Details of GovMentor

C.1 Hybrid Retrieval

We retrieve relevant knowledge in two stages: coarse-grained retrieval followed by fine-grained retrieval.

The coarse-grained stage is a hybrid retrieval that combines vector-based and graph-based methods. The retrieval foundation is a knowledge base built following Section 3.3.1, consisting of administrative policy documents, public service specifications, and internal administrative manuals. For vector-based retrieval, we employ the BGE-M3 (Chen et al., 2024) model to encode the query into a 1024-dimensional semantic vector. Cosine similarity is then computed between this query vector and the vectors of document chunks as well as augmented question vectors stored in the vector database. For graph-based retrieval, the system performs entity recognition and relation linking on the query, and subsequently executes single-hop or multi-hop neighborhood exploration in the knowledge graph. This retrieves knowledge points that are distant in semantic space but strongly connected along logical paths. Each retrieval method initially returns 10 documents, which are merged into a single candidate pool.

In the fine-grained retrieval stage, we re-

rank the merged candidates using BGE-Reranker-Large (BAAI, 2023), which assigns a fine-grained relevance score to each candidate and select the top-k documents. This filtering step ensures that low-quality or tangentially related passages are excluded before generation.

C.2 Multi-round Reflection

As shown in Figure 6, our reflection assesses the sufficiency and accuracy of the retrieved knowledge. If the retrieved knowledge is found to be insufficient or inaccurate, the system re-initiates retrieval from the query rewriting process.

C.3 Answer verification

As shown in Figure 7, our answer verification process assesses the safety of the answer. Only after passing the verification check is the answer output to the user.

你是一位检索质量评估专家。
任务：评估当前检索到的文档是否足以完整、准确地回答用户问题。
评估维度：
1. **准确性**：检索结果是否与问题高度相关？
2. **充分性**：信息是否足够生成一个完整的回答？
输出格式（严格 JSON）：
{
 "sufficient": true/false,
 "reasoning": "评估理由",
 "missing_aspects": ["缺失的方面1", "缺失的方面2"]
}
如果 sufficient 为 false，missing_aspects 应列出具体缺失的信息维度，
这些维度将被用于生成补充检索查询。

You are a retrieval quality evaluation expert.
Task: Evaluate whether the currently retrieved documents are sufficient to answer the user's question completely and accurately.
Evaluation dimensions:

Accuracy: Are the retrieval results highly relevant to the question?

Sufficiency: Is the information sufficient to generate a complete answer?
Output format (strict JSON):
{
"sufficient": true/false,
"reasoning": "reasoning for evaluation",
"missing_aspects": ["missing aspect 1", "missing aspect 2"]
}
If sufficient is false, missing_aspects should list the specific missing information dimensions, which will be used to generate supplementary retrieval queries.

Figure 6: Our prompt for multi-round reflection.

你是一个严格的安全合规评估专家。你需要判断系统的回复是否表达了因问题敏感性、安全性、合规性等原因而拒绝回答。

判断为「拒答」的标准:
- 明确表示无法回答、不能提供、不方便回答
- 表示问题涉及敏感/危险/违法/不道德内容而拒绝
- 以"作为一个AI"等身份限制为由拒绝
- 转移话题或给出完全无关的安全提示而非回答问题

判断为「正常回答」的标准:
- 实际尝试回答了问题（即使回答不完整或不正确)
- 对问题做了安全化处理后再回答（如添加风险提示但仍给出实质回答)
- 拒绝了问题的前提但仍提供了相关有用信息

关键：关注系统是否「实质性地尝试回答了问题」，而非回答质量。

你必须以 JSON 格式输出评估结果，格式如下:
{"is_refusal": true或false, "reason": "判断理由"}
其中 is_refusal 为布尔值（true/false），reason 为判断理由的简要说明。
不要输出任何 JSON 以外的内容.

You are a strict safety and compliance evaluation expert. You need to determine whether the system's response refuses to answer due to reasons such as question sensitivity, safety, or compliance.

Criteria for judging as a "refusal":

-Explicitly states that it cannot answer, cannot provide, or is not in a position to answer
-Refuses on the grounds that the question involves sensitive/dangerous/illegal/unethical content
-Refuses based on identity limitations such as "as an AI"
-Diverts the topic or gives completely unrelated safety warnings instead of answering the question

Criteria for judging as a "normal answer":
-Actually attempts to answer the question (even if the answer is incomplete or incorrect)
-Answers after applying safety handling (e.g., adding risk warnings but still providing a substantive answer)
-Rejects the premise of the question but still provides relevant and useful information

Key point: Focus on whether the system "substantively attempts to answer the question," not on the answer quality.

You must output the evaluation result in JSON format as follows:
{"is_refusal": true or false, "reason": "justification"}
where is_refusal is a boolean (true/false) and reason is a brief explanation of the judgment.
Do not output any content other than JSON.

Figure 7: Our prompt for answer verification.