*electronics*

*Article*

# ARGUS: Retrieval-Augmented QA System for Government Services

**Song Jiang, Xiaofeng Xie *, Rongnian Tang, Xuanqi Wang, Kaihao Sun, Guanghan Li, Zhenkai Xu, Peng Xue, Ziling Li and Xuedong Fu**

School of Mechanical and Electrical Engineering, Hainan University, No. 58 Renmin Avenue, Haikou 570228, China
* Correspondence: xfxie@hainanu.edu.cn

**Abstract:** The emergence of large language models (LLMs) has introduced new possibilities for government-oriented question-answering (QA) systems. Nonetheless, limitations in retrieval accuracy and response quality assessment remain pressing challenges. This study presents ARGUS (Answer Retrieval and Governance Understanding System), a fine-tuned LLM built on a domain-adapted framework that incorporates hybrid retrieval strategies using LlamaIndex. ARGUS improves factual consistency and contextual relevance in generated answers by incorporating both graph-based entity retrieval and associated text retrieval. A comprehensive evaluation protocol combining classical metrics and RAGAS indicators is employed to assess answer quality. The experimental results show that ARGUS achieved a ROUGE-1 score of 0.68 and a semantic relevance score of 0.81. To validate the effectiveness of individual system components, a chain-of-thought mechanism inspired by human reasoning was employed to enhance interpretability. Ablation results revealed improvements in ROUGE-1 to 68.5% and S-BERT to 74.9%, over 20 percentage points higher than the baseline. Additionally, the hybrid retrieval method outperformed pure vector (0.73) and pure graph-based (0.71) strategies, achieving an F1 score of 0.75. The main contributions of this study are twofold: first, it proposes a hybrid retrieval-augmented QA framework tailored for government scenarios; second, it demonstrates the system's reliability and practicality in addressing complex government-related queries through the integration of human-aligned metrics and traditional evaluation methods. ARGUS offers a novel paradigm for providing trustworthy, intelligent government QA systems.

**Keywords:** government question-answering; chain-of-thought fine-tuning; LlamaIndex; RAGAS; retrieval-augmented generation; domain-specific large language model

check for updates

Academic Editor: Arkaitz Zubiaga

Received: 16 May 2025
Revised: 13 June 2025
Accepted: 14 June 2025
Published: 16 June 2025

**Citation:** Jiang, S.; Xie, X.; Tang, R.; Wang, X.; Sun, K.; Li, G.; Xu, Z.; Xue, P.; Li, Z.; Fu, X. ARGUS: Retrieval-Augmented QA System for Government Services. *Electronics* **2025**, *14*, 2445. https://doi.org/10.3390/electronics14122445



## 1. Introduction

With the rapid advancement of digital government initiatives, government question-answering (GQA) systems have become a vital channel for communication between authorities and the public [1]. The quality of these systems directly impacts government credibility and the effectiveness of public service delivery. Although many government platforms perform well in terms of response speed, studies in various countries have shown that significant gaps remain in accuracy and user satisfaction [2].

Despite these advancements, real-world platforms still fall short in meeting users' actual information needs. For example, citizens often receive generic or ambiguous responses when inquiring about student admission policies or enrollment procedures. These unsatisfactory interactions, rooted in poor semantic understanding, undermine public trust and underscore the urgent need for more context-aware and reliable GQA systems capable of addressing complex and evolving queries.

Most existing GQA systems still rely on keyword recognition methods, making them inadequate for handling diverse and evolving user queries [3]. These deployments reflect a broader trend toward integrating advanced AI technologies into public service systems.

Embedding domain-specific expertise is crucial for aligning ML models with the complex realities of public sector decision-making, thereby improving the precision and efficiency of public services [4]. Across various regions in China, government departments have begun utilizing large language models (LLMs) to assist with administrative tasks, enhance service efficiency, and strengthen citizen engagement [5]. Trained on massive corpora, LLMs possess robust semantic understanding, reasoning, and generation capabilities [6–9], making them particularly well suited for addressing the complex scenarios encountered in government services. Furthermore, LLMs can be fine-tuned for domain-specific applications to better accommodate different service demands and policy contexts [10,11]. By integrating retrieval-augmented generation (RAG) techniques, LLMs can dynamically access up-to-date government knowledge bases, regulations, and procedural guidelines during inference, thereby improving response timeliness and precision [12].

However, significant challenges remain. Current mainstream fine-tuning approaches often lack explicit reasoning chains tailored to complex government-related inquiries, limiting their capacity to address intricate demands [13,14]. Moreover, the retrieval components in RAG pipelines often produce lengthy, low-relevance text blocks, introducing substantial noise and placing a heavy contextual burden on models [15]. Additionally, the absence of standardized evaluation frameworks for generated responses hinders the fulfillment of domain-specific professional requirements [16]. Thus, systematic optimization is urgently needed across multiple dimensions, including incorporating reasoning strategies, representing structured knowledge, and designing retrieval methods.

In response to these challenges, this study proposes an integrated retrieval framework based on LlamaIndex [17,18]. We introduce a multi-step chain-of-thought mechanism into domain fine-tuning to explicitly model reasoning paths over domain knowledge [19]. Furthermore, we construct a domain-specific knowledge graph through prompt engineering [20] and incorporate a hybrid retrieval strategy that combines graph-based guidance with relevant textual retrieval to enhance answer quality. To validate the system, we use the RAGAS [21] framework to conduct structured evaluations of the model output.

Notably, our work delivers a quantifiable, comprehensive, and novel framework for government QA systems. We developed and deployed ARGUS, a fully integrated, domain-specific QA platform spanning five real-world government subdomains, built on a built-on 14 billion parameter model and operable on a single NVIDIA 4070S GPU. In addition, we establish a structured evaluation protocol that synthesizes classical metrics (ROUGE-1, S-BERT, COMET) with RAGAS-based indicators: ARGUS achieves a ROUGE-1 score of 0.68 and a semantic relevance of 0.81. Component-level analyses demonstrate that the integration of human-inspired chain-of-thought mechanisms substantially improves procedural fidelity, while ablation of knowledge-graph entries elevates ROUGE-1 to 68.5%, S-BERT to 74.9%, and faithfulness to 70.6%. Finally, our hybrid retrieval strategy achieves an F1 score of 0.75, outperforming both the baselines based on pure vectors (0.73) and graph-only (0.71).

The contributions of this work are summarized as follows:

1. A Domain-Specific Large Language Model Framework for Government Services: We propose a novel paradigm for constructing large language models (LLMs) tailored to government service tasks. We design a hybrid retrieval-augmented generation (RAG) strategy that combines graph-path guidance with relevant text retrieval to enhance knowledge controllability and generation accuracy. This is achieved by

embedding human-like chain-of-thought reasoning into domain-specific fine-tuning and developing an automated knowledge graph construction pipeline based on prompt engineering within the LlamaIndex framework.

2. Response Quality Control via RAGAS Evaluation: We develop a structured response evaluation mechanism integrating the RAGAS framework. To address the challenge of verifying LLM-generated content, we design a methodology that combines structured prompt templates with multidimensional evaluation metrics. This ensures contextual consistency, factual accuracy, and the practical value of model outputs while providing a quality assurance mechanism for high-risk application scenarios.
3. Practical Deployment and Validation in Real-World Government Services: We implement the proposed system in actual government service scenarios, demonstrating significant improvements in response accuracy and processing efficiency. The system offers a practical and scalable solution for enhancing government consultation services, providing a feasible technical path and methodological reference for constructing intelligent government service platforms.

## 2. Related Works

In research on government question-answering (GQA) systems, scholars have investigated the integration of large language models (LLMs) through fine-tuning to improve tasks such as policy interpretation and administrative consultation. For example, some studies have employed lightweight adjustments combined with domain-specific fine-tuning strategies [22], leveraging government-related corpora—such as administrative service item databases and 12345 hotline records—to adapt the models and achieve notable performance gains. However, due to the specialization and time sensitivity of government knowledge, existing fine-tuning methods face persistent challenges, including delayed responses to policy updates and limited generalization to local regulation variations.

Several international initiatives have explored the use of structured data in constructing government knowledge graphs to enhance public sector intelligence. Notable efforts include a knowledge graph built from disease-related datasets on the Nova Scotia Open Data portal in Canada [23]. Furthermore, studies have examined the impact of knowledge graphs on public service quality, highlighting their potential to support decision-making and improve service delivery in governmental contexts [24]. These graphs are typically constructed by applying natural language processing tasks such as named entity recognition (NER) and relation extraction (RE) to unstructured text. Techniques such as OpenIE [25] and BERT-based relation extraction [26] are effective technical methods for graph construction. However, current methods often rely heavily on manually crafted rules or the availability of sufficient training data. As a result, they exhibit limited adaptability to domain-specific terminology and low-resource settings. Moreover, the lack of unified quality control standards hinders the application of high-quality knowledge graphs in complex tasks [26].

In information retrieval, government question-answering (GQA) systems commonly employ retrieval methods based on BM25 [27] or semantic vector matching. For example, applications such as "Unified Policy Search" combine word vector semantic matching with classifier-based filtering mechanisms [28]. However, most existing systems lack a deep understanding of user intent and context, limiting their ability to handle complex demands such as policy logic reasoning, ambiguous expressions, and compound queries. The integration of retrieval and large language model generation remains in its early exploratory stage, with optimization strategies still largely underdeveloped.

Despite prior advances, key limitations remain—retrieval methods like BM25 and dense vectors lack semantic depth and scalability, and existing systems often fail to tightly couple retrieval with domain reasoning and evaluation. ARGUS addresses these gaps through hybrid retrieval, reasoning-guided generation, and human-aligned evaluation, enabling more accurate and context-aware responses in complex government scenarios.

In summary, despite notable progress in model fine-tuning, knowledge organization, and retrieval integration within the GQA domain, significant challenges persist in processing semantically complex, structurally dispersed, and time-sensitive government data. Current approaches often exhibit limited modeling diversity, low responsiveness, and weak integration capabilities, highlighting key areas for future research and development.

## 3. Construction of a Government-Oriented Large Language Model QA System

DeepSeek-R1 is an open-source model that demonstrates outstanding performance in Chinese-language contexts [29]. It is preferred over other open-source models for its use of chain-of-thought reasoning, which enables users to trace the model's thought process rather than view only the input–output results. This interpretability is particularly important for government QA tasks, where responses must follow clear and structured logic. Based on this capability, we adapt DeepSeek-R1 to simulate the reasoning style of a government service agent operating under rules-based constraints.

To support secondary development and deployment for government question-answering (GQA) tasks, a distilled and optimized variant, DeepSeek-R1-Qwen-14B, was adopted as the system's core. The overall design is illustrated in Figure 1. During domain-specific fine-tuning, the question-answering data was structured into standardized formats. During knowledge graph construction, the model's information extraction capabilities were utilized to generate knowledge triplets, which were imported into a database. Finally, during model interaction, both knowledge graph retrieval and vector-based semantic retrieval were integrated to support continuous, query-driven access to the knowledge base, enabling the model to generate specific and contextually grounded responses.

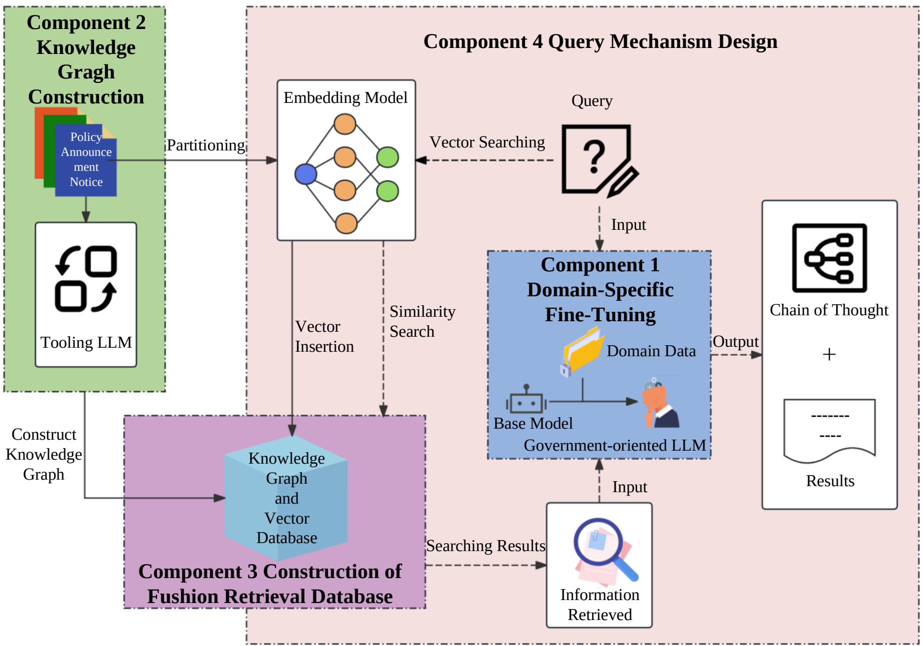


**Figure 1.** Architecture of the proposed QA system integrating domain-specific fine-tuning, knowledge graph construction, and hybrid retrieval. The system combines knowledge graph and vector-based semantic retrieval, with both outputs re-ranked by similarity and policy-type alignment before being input to the government-oriented LLM for final response generation.

While many core technologies in ARGUS have been explored by researchers, its novelty lies in task-specific integration and optimization for the government QA domain. Rather than relying on a single breakthrough, ARGUS achieves system-level innovation through the orchestration of knowledge structuring, retrieval routing, reasoning-guided generation, and human-aligned evaluation. This closed-loop design ensures that answers are not only accurate and grounded but also interpretable and contextually faithful, which is essential for deployment in real-world government services.

Unlike conventional RAG pipelines that rely solely on either dense retrieval or structured triplet matching, ARGUS employs a hybrid retrieval strategy that adaptively combines both modalities, enhancing contextual fidelity and procedural relevance in government QA scenarios.

### 3.1. Domain-Specific Fine-Tuning

The fine-tuning dataset was sourced from a QA knowledge base crawled from the official website of the 12345 government service hotline. Since policy-related consultations are manually entered by service agents and inherently unstructured, data cleaning was performed to improve training quality. The main steps were as follows:

1. Removal of invalid cases: Entries not requiring telephone follow-up, such as feedback-only entries, were excluded.
2. Normalization check: HTML formatting symbols and redundant or duplicate information were removed.
3. Error Correction: Grammatical and typographical errors were corrected.
4. Bias and Conflict Mitigation: To reduce policy misalignment and regional inconsistencies, only authoritative and up-to-date regulations were retained. Cases showing conflicting interpretations or outdated references were excluded based on cross-checking with official databases.
5. Structural alignment: The dataset entries were transformed into a standardized format, 'Materials–Steps–Notes', to improve procedural clarity and minimize ambiguity during generation.

We selected two representative government QA entries and retrieved the highest scoring context snippets under two conditions: 'Cleaned but unstructured' vs. 'Structured with Materials–Steps–Notes'. We then compared their presentation of redundant information and key procedural points (see Table 1).

**Table 1.** Qualitative Case Comparison

| Case | Cleaned but Unstructured | Structured with 178 Materials–Steps–Notes |
|---|---|---|
| Example 1 | "According to the XX Regulation, … (a long paragraph of regulatory background and historical context), in special circumstances, … the service window addresses are … contact information…" | **Materials**: Quotation of the relevant clauses from the XX Regulation; **Steps**: 1. Prepare documents, 2. Submit the application, 3. Wait for review; **Notes**: See TableX for addresses and contact numbers. |
| Example 2 | "Users must bring their original ID card and household register along with copies, and fill out the application form; the detailed process is … (a lengthy description including subsidy policies, notes, hotline number)" | **Materials**: ID card, household register; **Steps**: Fill out the form → Submit → Receive receipt; **Notes**: Reissue within 7 days using the receipt. |

From the comparison, it was clear that the three-part "Materials–Steps–Notes" format eliminated a large amount of redundant background text and distilled the core procedural steps at a glance, enabling the model to focus on key information and avoid distractions from irrelevant context.

### 3.2. Knowledge Graph Construction

To extract key information from unstructured texts and improve retrieval efficiency, a domain-specific knowledge graph was constructed for government services. Public policy documents were collected by crawling government portal websites and served as the graph's data source. A pre-trained model was used as the core construction tool, combined with custom-designed prompt templates and extraction rules to automate the generation and storage of knowledge triplets. The extracted information was stored in a Neo4j graph database, and a continuous vector index was built to support efficient retrieval by the large language model. The reasoning procedure involved five steps:

1. **Demand Analysis:** Identify core needs from the query.
2. **Key-Point Extraction:** Extract constraint conditions from user input.
3. **Policy Matching:** Match relevant clauses from retrieved knowledge.
4. **Policy Mapping:** Transform legal clauses into executable steps.
5. **Conclusion Generation:** Generate draft responses based on the processed information.

#### 3.2.1. Sources of the Knowledge Graph Dataset

To enhance the intelligence of government services, knowledge graph technology has been increasingly applied in recent years as part of digital government transformation efforts [23]. To ensure the systematic coverage of key areas in government consultation services, this study constructed the knowledge graph based on the classification system of the 12345 government service hotline platform and the corresponding statistical data.Five subdomains were selected as primary data sources: social security, healthcare, public security, education, and transportation. This structured foundation supported subsequent information extraction and the construction of a domain-specific knowledge graph.

#### 3.2.2. Knowledge Graph Construction Using Large Language Models

During the triplet extraction process, the contextual richness and semantic ambiguity inherent in natural language texts can introduce noise and cause semantic drift in the results [30]. To enhance the precision and robustness of relation extraction, this study adopted a prompt engineering strategy based on precondition constraints and few-shot learning [31]. The instruction and scope components defined the task boundary and domain focus, serving as preconditions to constrain the model's semantic attention. The input–output example provided a standardized reference for output formatting, following the few-shot learning paradigm.

Specifically, structured prompt templates were designed to guide the model's extraction behavior using standardized input–output examples [32]. Using the relation "school admission" as an example, a minimal version of the triplet extraction prompt is shown in Algorithm 1, and the corresponding extracted knowledge graph is presented in Figure 2.

**Algorithm 1 Prompt template for triplet extraction**

```
1: Instruction: Extract up to {max_knowledge_triplets} triplets in the format (Subject, Predicate, Object).
2: Scope: Focus on text related to admission policies, especially conditions, materials, and procedures.
3: Example:
4: Text: "The education department promotes admission based on degree availability."
5: Triplets:
6:    (Education department, promotes, admission)
7:    (Admission, based on, degree availability)
```

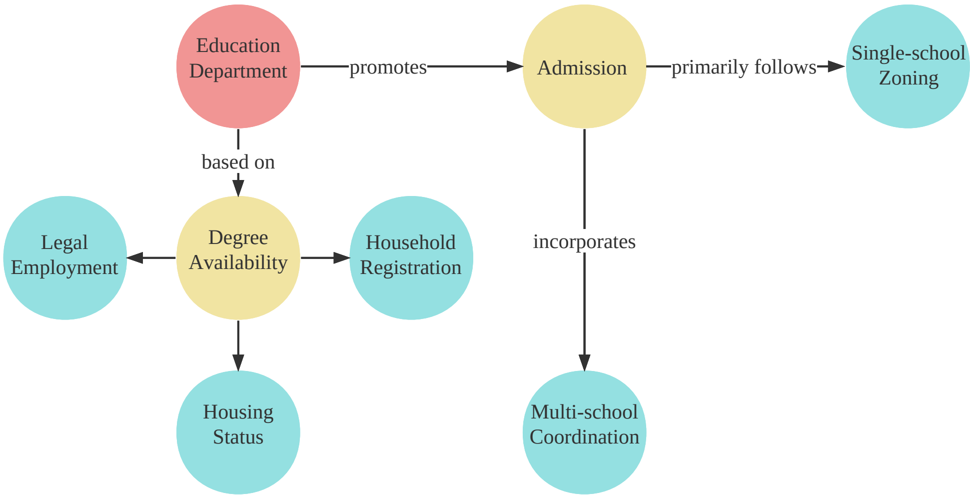


**Figure 2.** Example of a structured knowledge graph extracted using prompt-based relation extraction. The illustration is based on the relation "school admission" and shows how the model, guided by standardized input–output examples and precondition constraints, extracts semantically grounded triplets such as ⟨Education Department, promotes, Admission⟩ and ⟨Admission, incorporates, Multi-school Coordination⟩.

### *3.3. Construction of Vector Embeddings and Storage*

This section presents the construction and utilization of a persistent retrieval system tailored to the government domain. Government-related text data were vectorized to build a sustainable vector index, which was integrated with a graph database to support complex semantic associations.

The vector retrieval pathway processed unstructured textual inputs using the pre-trained embedding model bge-large-zh-v1.5, which converted the text into high-dimensional semantic vectors [33]. These vectors were then linked to nodes in the Neo4j graph database [34].

As shown in Figure 3, when a user submits a government-related query, the system invokes the persistent index to perform hybrid retrieval of the top K most relevant nodes (Node1...NodeK). These retrieved nodes are then fed into the "response synthesis" module, which generates an answer or analytical output aligned with the user's query.

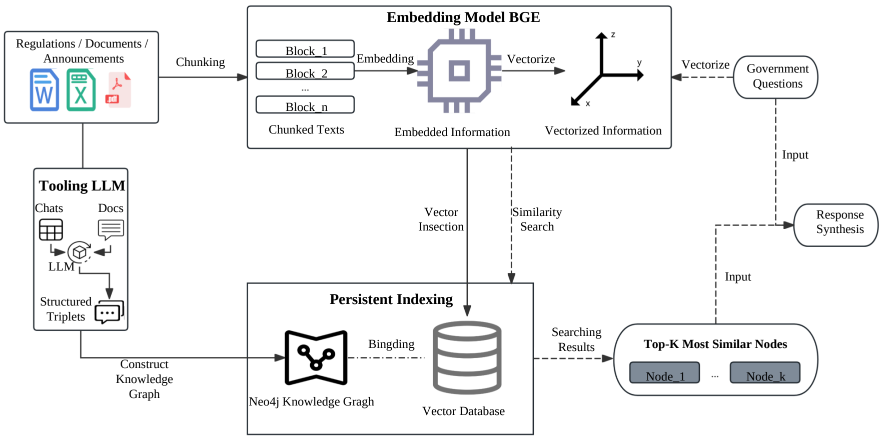


**Figure 3.** Workflow of the hybrid retrieval mechanism, illustrating how unstructured policy documents are chunked and vectorized using a domain-specific embedding model. The resulting vectors are persistently stored and bound to structured nodes in the Neo4j knowledge graph. During inference, government queries are encoded and matched against the indexed database, and the top K most similar nodes are retrieved and synthesized to generate final responses.

*3.4. Hybrid Retrieval*

To enhance retrieval and reasoning capabilities for government-oriented question-answering (QA) tasks, we propose a hybrid retrieval method based on persistent indexing. The system first performs persistent retrieval based on the user's query and merges the retrieved results with the user's consultation information through prompt-based expansion, forming a more comprehensive retrieval context.

Within the hybrid retrieval module, the system combines the generation capabilities of large language models (LLMs) with structured knowledge retrieval to fuse, reorganize, and optimize candidate answer sets. The synthesized content, along with the user query, is then passed to the LLM to generate responses more aligned with user intent. These responses explicitly cite the corresponding data sources, ensuring logical coherence and factual consistency.

The proposed hybrid retrieval approach integrates the reasoning capabilities of knowledge graphs with the language understanding of LLMs, improving the system's ability to handle complex administrative queries, policy matching, and cross-domain inference tasks in the public sector. The workflow of the hybrid retrieval mechanism is illustrated in Figure 4.

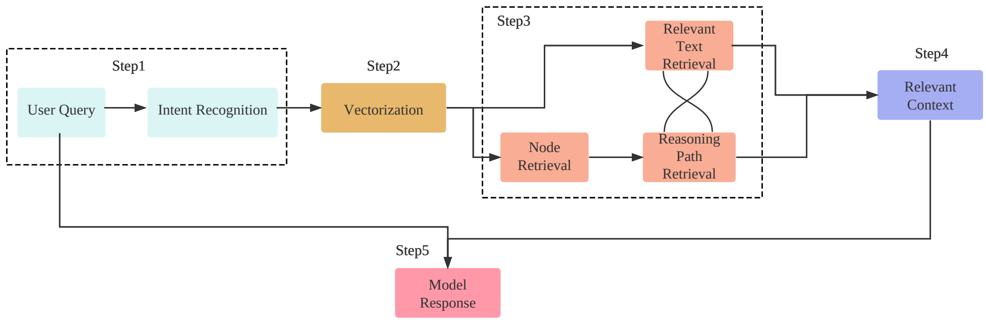


**Figure 4.** Workflow of the proposed hybrid retrieval mechanism. The system begins with user query understanding through intent recognition (Step 1), followed by vectorization of the query (Step 2). In Step 3, dual-path retrieval is performed: a vector-based semantic search retrieves relevant policy texts, while knowledge graph-based reasoning retrieves logical node paths. In Step 4, retrieved contexts are fused to form a structured input. Finally, in Step 5, the fused context and query are fed into the LLM to generate a response that is accurate, policy-aware, and explainable.

## 4. Validation Experiments

*4.1. Experimental Setup*

4.1.1. Experimental Device

This study employed the Longjing LT4214G-8I GPU computing server, equipped with multiple NVIDIA GeForce 4090 GPUs, to accelerate model training. The deepseek-distill-qwen-14b model was deployed locally as the base large language model (LLM). DeepSeek was chosen for its strong reasoning ability and inherent CoT-style outputs. Due to its closed-source nature, we distilled its behavior into the open-source Qwen model, enabling us to retain DeepSeek's strengths while ensuring controllability and flexibility in fine-tuning.

4.1.2. Test Set Design

To improve model applicability in the government domain, a domain-specific dataset was compiled for fine-tuning. A five-step chain-of-thought method, based on progressive reasoning principles [19,35], was designed to guide the model in reasoning from unstructured user input to structured policy question-answering, thereby constructing a complete reasoning path. The dataset categories and quantities are shown in Figure 5a.

To further ensure the factuality and consistency of the test set, entries involving ambiguous jurisdiction, contradictory provisions, or outdated content were manually removed. This ensured that the evaluation resulted reflect real-world policy execution standards.

To validate the proposed method on a government services domain test dataset, the native DeepSeek model was used to generate question–answer pairs based on policy documents. The test set covered subdomains including social security, healthcare, household registration, education, and public security, as shown in Figure 5b.

The datasets used in this study were primarily sourced from the publicly accessible database of the Haikou 12345 government service hotline. The test set was generated by the gpt-4-turbo model through structured analysis of official policy documents. Although the source data is publicly available, due to the sensitivity and formality of government-related content, the compiled datasets cannot be released publicly.

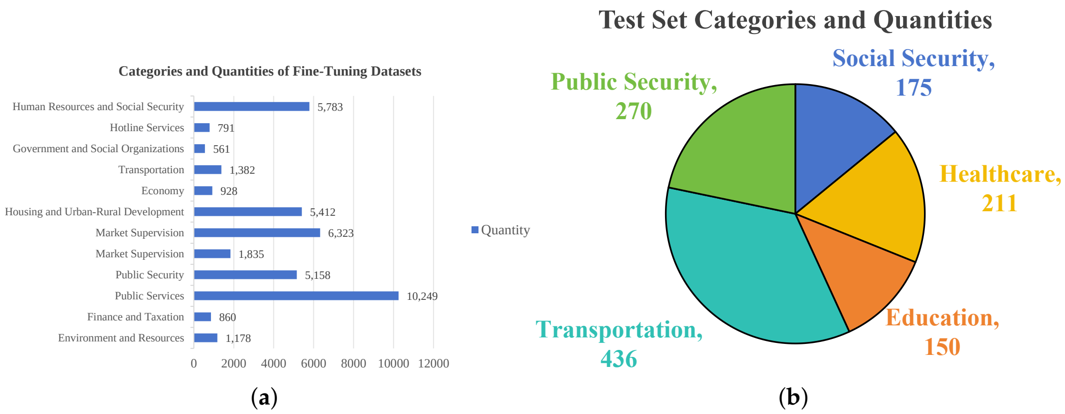


**Figure 5.** (**a**) Category-wise distribution of samples in the fine-tuning dataset for government QA tasks; (**b**) category-wise statistics of the test dataset across the transportation, public security, healthcare, social security, and education subdomains. These panels supported the training and evaluation of the hybrid retrieval model in structured government question-answering.

#### 4.1.3. Baseline Models

To ensure a fair and representative comparison, four recently released and widely adopted open-source large language models were selected as baselines. LLaMA3 [36], developed by Meta AI, is a general-purpose multilingual model known for its strong multitask capabilities, though primarily trained on English corpora. Qwen [37], released by Alibaba, is a Chinese-optimized model designed for high-quality understanding and generation in native language contexts. Baichuan [38], developed by Baichuan Inc., is a bilingual model trained on both Chinese and English data, offering a balance between domain comprehension and generation fluency. ChatGLM [39], introduced by Tsinghua University and Zhipu AI, is a dialogue-oriented Chinese language model enhanced with supervised fine-tuning and RLHF to better align with user intent. These models provide a representative baseline spectrum for evaluating domain alignment and generation quality in Chinese government service tasks.

### *4.2. Evaluation Metrics*

To ensure that the government service question-answering met standards of correctness, clarity, and formal expression—and avoided misleading or alarming interpretations of policies and regulations [40]—this study employed conventional automatic metrics such as ROUGE and BERTScore. In addition, the RAGAS framework was adopted, incorporating GPT-4 to objectively assess the consistency between generated answers and the retrieved context.

#### 4.2.1. Objective Metrics

To validate the accuracy of the generated responses, we included external reference-based metrics such as ROUGE-1, S-BERT, and COMET, which evaluate lexical overlap, semantic similarity, and generation quality, respectively.

Answer completeness was evaluated using the ROUGE metric [41], which focuses on recall and effectively assesses whether generated content covers key information from reference answers. For example, ROUGE-n is calculated as follows:

$$\text{ROUGE-n} = \frac{\text{card}(S_n \cap O_n)}{\text{card}(S_n)} \tag{1}$$

where $S_n$ and $O_n$ denote the sets of n-grams in the reference and generated answers, respectively.

Correctness was evaluated using a semantic matching model fine-tuned under the sentence-BERT framework. Compared to the general-purpose BERT model, this model performs alignment learning in the semantic vector space, focusing not only on literal expressions but also on semantic similarity between sentences. Consequently, the generated sentence embeddings more accurately reflect the underlying semantic relationships [42].

The COMET metric [43] is a reference-based automatic evaluation method for natural language generation tasks. It leverages deep learning models to assess the quality of generated text against reference answers and is widely applied in summarization, dialogue, and text generation.

#### 4.2.2. RAGAS Metrics for Ablation Study

To validate the rationality and effectiveness of the retrieved context, we adopted three RAGAS metrics: context precision, context recall, and faithfulness. These metrics were specifically designed to assess the alignment between the generated answers and the supporting retrieved information.

Context precision [21] measures relevance of the retrieved context. It is typically expressed as precision@k and aggregated into a weighted context precision@K for overall evaluation:

$$\begin{aligned} \text{Context Precision@}K &= \frac{\sum_{k=1}^{K} \text{Precision@}k \times v_k}{\text{Total relevant items in top } K} \\ \text{where} \quad \text{Precision@}k &= \frac{\text{TP@}k}{\text{TP@}k + \text{FP@}k} \end{aligned} \tag{2}$$

Context recall [21] evaluates the proportion of relevant reference fragments successfully retrieved, emphasizing omission reduction. Higher values indicate fewer missed relevant items. It is calculated as

$$\text{Context Recall} = \frac{C_{ref}^{\text{supported}}}{C_{ref}^{\text{total}}} \tag{3}$$

Faithfulness [21] measures the factual consistency between a response and its supporting context, with scores ranging from 0 to 1. A response is considered faithful if all claims are verifiable based on the retrieved context. The faithfulness score is computed as the ratio of supported claims to the total number of claims:

$$\text{Faithfulness Score} = \frac{C_{\text{res}}^{\text{supported}}}{C_{\text{res}}^{\text{total}}} \tag{4}$$

#### 4.2.3. RAGAS Metrics for Subtask Evaluation

To evaluate the semantic alignment between the generated response and the user query, we used the RAGAS response relevancy metric. This metric estimates how well the response addresses the user's original intent. Higher scores indicate stronger alignment with the input, while irrelevant, incomplete, or redundant responses are penalized.

The score is computed by comparing the embedding of the user query $E_0$ with those of auto-generated questions $E_{g_i}$, which represent the response content. The final score is the average cosine similarity:

$$\text{Response Relevancy} = \frac{1}{N}\sum_{i=1}^{N}\cos(E_{g_i}, E_0) \tag{5}$$

where $N$ is the number of generated questions (typically 5).

*Note*: While scores typically fall between 0 and 1, cosine similarity may range from $-1$ to 1.

#### 4.2.4. Summary of Metrics

Table 2 summarizes all evaluation metrics used in this study, including their types, definitions, and specific roles. These metrics are categorized into conventional automatic metrics and RAGAS-specific metrics for both ablation and subtask-level evaluation.

**Table 2.** Summary of evaluation metrics and confidence intervals.

| Metric | Type | Role and Purpose |
|---|---|---|
| ROUGE-1 | Lexical overlap | Measures content completeness using n-gram overlap between the output and reference. |
| S-BERT | Semantic similarity | Assesses sentence-level embedding similarity using cosine similarity. |
| COMET | Generation quality | Predicts generation quality using a learned scoring model based on references. |
| Context Precision | Retrieval quality | Assesses the relevance of retrieved context chunks to the reference. |
| Context Recall | Retrieval coverage | Measures how much of the reference content is captured by the retrieved context. |
| Faithfulness | Factual consistency | Assesses factual consistency between the responses and retrieved evidence. |
| Response Relevancy | Query alignment | Evaluates how well the response addresses the user's original intent. |

### *4.3. Subtask Experiments*

To compare the performance of ARGUS in government service question-answering tasks, experiments were conducted across five subdomains: social security, healthcare, public security, transportation, and education. Each model generated responses based on the government service test set and the outputs were recorded for subsequent analysis.

Representative examples comparing the responses generated by LLaMA, Qwen, Baichuan, ChatGLM, and ARGUS are provided in Appendix A to illustrate the performance advantages of ARGUS over the other models.

In government-related QA tasks, baseline models such as LLaMA, Qwen, Baichuan2, and ChatGLM3 underperformed compared to ARGUS due to limited domain adaptation and weak structural generation. LLaMA often produced irrelevant answers in Chinese contexts. Qwen and Baichuan2, though fluent in Chinese, tended to generate verbose and unstructured outputs. ChatGLM3 performed slightly better in instruction tasks but was constrained by its smaller size. In contrast, ARGUS benefited from domain-specific fine-tuning and structured retrieval, yielding more accurate and standardized responses.

The ARGUS model significantly outperformed the baseline models in both the ROUGE and the response relevancy scores across all domains. These results indicate that the customized and optimized ARGUS provides better coverage of reference content (as reflected by higher ROUGE scores) and greater relevance and alignment with user input (as reflected by higher response relevancy scores). The highest relevancy score, 0.8714, was observed in the education domain, highlighting the model's strong adaptability in professional policy text processing scenarios.

In the transportation domain, all models achieved relatively low ROUGE-1 and response relevancy scores, likely due to the complexity of transportation-related queries, which often require integrating multi-source knowledge such as regulations, classification standards, and real-time traffic conditions, raising the bar for precise retrieval and structured generation.

By contrast, in the public security domain, most models—including LLaMA—performed relatively well. This may be attributed to the fact that many public security-related queries rely on common knowledge or standardized procedures with minimal regional variation, making them easier to handle even without strong domain adaptation. The comparative results are presented in Table 3. Values are reported as percentages with approximate 95% confidence intervals (±), calculated by computing the sample mean and estimating variability via bootstrap resampling across multiple generations per domain.

**Table 3.** Comparison of ROUGE-1 and response relevancy scores across different domains.

| **Domain** | **ROUGE-1 (%)** | | | | | **Response Relevancy-5 (%)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **LLaMA** | **Qwen** | **Baichuan** | **ChatGLM** | **ARGUS** | **LLaMA** | **Qwen** | **Baichuan** | **ChatGLM** | **ARGUS** |
| Social Security | 27 ± 9 | 46 ± 11 | 43 ± 9 | 42 ± 14 | **71 ± 5** | 42 ± 17 | 57 ± 12 | 55 ± 10 | 56 ± 13 | **89 ± 7** |
| Healthcare | 31 ± 15 | 53 ± 10 | 56 ± 8 | 50 ± 13 | **65 ± 6** | 41 ± 18 | 46 ± 13 | 50 ± 11 | 45 ± 14 | **82 ± 5** |
| Public Security | 56 ± 5 | 51 ± 12 | 54 ± 9 | 48 ± 13 | **64 ± 5** | 70 ± 5 | 76 ± 6 | 79 ± 4 | 73 ± 7 | **81 ± 6** |
| Transportation | 35 ± 18 | 41 ± 13 | 40 ± 10 | 39 ± 12 | **73 ± 10** | 44 ± 17 | 49 ± 11 | 51 ± 8 | 46 ± 12 | **74 ± 5** |
| Education | 40 ± 16 | 61 ± 10 | 62 ± 8 | 55 ± 12 | **70 ± 5** | 45 ± 15 | 70 ± 11 | 72 ± 9 | 65 ± 13 | **86 ± 10** |

### *4.4. System Effectiveness Validation*

This section validates the effectiveness of improvements in three key aspects through experimental evaluation. First, a comparative analysis was conducted to assess the enhancement in model response quality achieved by the chain-of-thought fine-tuning strategy. This involved a qualitative comparison of the model's response accuracy and logical coherence before and after fine-tuning. Second, a knowledge graph ablation experiment was performed by progressively removing portions of the extracted triplets to verify the rationality of the graph architecture and the comprehensiveness of knowledge coverage. Finally, a multidimensional evaluation framework was employed, incorporating classification metrics such as precision, recall, and F1 score, to quantitatively analyze the performance improvement of the fusion retrieval strategy—which integrated both graph-based and text-based retrieval—over single retrieval methods in query response tasks.

#### 4.4.1. Validation of Human-like Chain of Thought

The foundation model employed in this study inherently possessed basic reasoning capabilities. Building upon this, we adopted a supervised fine-tuning approach to further guide the model toward acquiring and internalizing domain-specific CoT reasoning patterns tailored to government service contexts. Specifically, we designed a structured five-step CoT framework during the fine-tuning phase to systematically simulate the cognitive process of human government service agents when responding to citizen queries. Through this training paradigm, the model learned to autonomously generate reasoning paths that resembled the step-by-step thought process of a trained service agent during inference. To evaluate the effectiveness of this design, the model responses after fine-tuning were recorded and compared with those of the main experiments. The results are presented in Table 4.

**Table 4.** Comparison of QA outputs before and after chain-of-thought fine-tuning.

| **Q1 (Education)** | What are the online application channels for new student enrollment in primary and secondary schools in Haikou? |
|---|---|
| **Original Model Response** | <think> Okay, so I need to figure out the online application channels for new student enrollment in primary and secondary schools in Haikou. Let me start by breaking this down.<br>First, I know that Haikou is the capital city of Hainan Province in China. The question is about enrolling new students, so we're talking about primary (elementary) and secondary (middle school) levels.<br>I should consider the government's approach to online services. In China, many government services are being digitized, so there's likely an official portal or app for this purpose. I remember hearing about the "Ganjiu" platform, which is used for various government services. Maybe that's one channel.<br>Also, the Haikou (Bureau of Education) probably has an official website. They would likely have an online application section there. I should check if that's the case.<br>Another possibility is the Hainan Province's government service platform. Since Haikou is part of Hainan, the provincial platform might handle school enrollments. It's possible that applications can be submitted there as well.<br>I should also think about the Haikou Municipal Government's portal. They often have sections for education and enrollment, so that's another potential channel.<br>Wait, there's also the "iHaikou" app. I think many cities in China have their own apps for municipal services. It's likely that Haikou has one, and enrollment might be handled through that app.<br>So, to summarize, the channels would include:<br>1. Haikou's official website. 2. The Hainan Province Government Service Platform. 3. The Haikou Municipal Government's official website. 4. The "Ganjiu" platform. 5. The "iHaikou" app.<br>I should make sure these are accurate. Maybe I can think about each one: ...<br>5. **iHaikou App**: Download the "iHaikou" app, a mobile-friendly application that offers various municipal services, including school enrollment.<br>These channels provide convenient and accessible ways to apply online for school enrollment in Haikou. |
| **Fine-Tuned Model Response** | <think> Analyze the user's main intent, which is to inquire about online application channels; Decompose the key elements of the question, including "new student enrollment in primary and secondary schools" and "online application channels"; Match relevant policy information to identify the procedures for online application in Haikou; Convert the policy information into user-executable answers, such as providing specific website links or contact methods; Conclude with the question: "What are the online application channels for new student enrollment in Haikou?" </think> The online application channels for new student enrollment in Haikou are: 1. Haikou Education Bureau official website; 2. Haikou Education Bureau WeChat public account; 3. Yechengban APP; 4. Haiyiban APP. |

To further demonstrate the generalizability of our chain-of-thought prompting method, Appendix B provides examples from additional domains, including healthcare and transportation, illustrating how internal reasoning (think) improved after fine-tuning.

#### 4.4.2. Ablation Study on the Knowledge Graph

The government knowledge graph served as the primary source of domain knowledge for the model. To further evaluate its comprehensiveness in government service question answering, an ablation experiment was conducted.

In addition to conventional evaluation metrics, the experiment incorporated RAGAS-based indicators—context precision, context recall, and faithfulness—to assess retrieval-augmented generation (RAG) capabilities. During graph construction based on contextual information, varying proportions of triplets were selected. Specifically, a complete (100%) graph corresponded to 20 triplets per chunk, while a 75% preserved graph included 15 triplets.

The experimental results are presented in Table 5. As the proportion of retained triplets increased, the three evaluation metrics initially improved, then stabilized, and, finally, declined slightly. These results indicate that incorporating knowledge graph information enhances the relevance and semantic consistency of text generation.

In the 50–100% retention range, the three metrics remained relatively stable with minor fluctuations. In particular, the S-BERT and COMET scores reached their peaks within this range, suggesting diminishing returns from additional graph information. When the retention proportion increased to 125%—likely introducing redundant or noisy information—the ROUGE score dropped to 66.6%, while the S-BERT and COMET scores remained largely unchanged. This suggests that excessive graph information may impair surface-level consistency without substantially affecting deep semantic alignment.

To elucidate the modest performance degradation observed at the retention level 125%, we performed a detailed analysis of the additional triples incorporated beyond the full knowledge graph. Although all triples were validly extracted according to our prompt templates, many conveyed semantically overlapping content. For example, the policy assertions 'degree availability determines admission', 'admission is based on degree availability', and 'degree availability is a key criterion for admission' all articulated the same underlying fact. These redundant or near-duplicate triples occupied valuable context slots without contributing novel information, thus diluting the relevance signal and precipitating declines in both ROUGE-1 (falling to 66. 6%) and context precision when retention exceeded 100%.

**Table 5.** Evaluation results for different knowledge graph retention ratios.

| Retention Ratio (%) | ROUGE-1 (%) | S-BERT (%) | COMET (%) | Context Precision (k = 5) (%) | Context Recall (%) | Faithfulness (%) |
|---|---|---|---|---|---|---|
| 0 | 34.0 | 68.3 | 70.5 | – | 43.8 | 49.6 |
| 25 | 63.7 | 71.5 | 76.3 | 45.2 | 52.2 | 66.6 |
| 50 | **68.8** | 74.7 | 78.1 | 43.1 | 46.5 | 65.6 |
| 75 | 68.7 | 74.7 | 78.1 | 34.3 | **68.3** | 60.0 |
| **100** | 68.5 | **74.9** | **78.2** | **51.3** | 53.7 | **70.6** |
| 125 | 66.6 | 74.9 | 78.7 | 43.3 | 45.4 | 65.7 |

"–" indicates that the metric was not applicable under zero retention.

### 4.4.3. Retrieval-Augmented Classification Experiment

In the retrieval-augmented validation experiment, a pre-trained BERT model computed the semantic similarity between reference and generated answers. Based on a predefined threshold, predictions were classified as positive or negative. Specifically, if the similarity score exceeded 0.7, the model was considered to have correctly identified relevant content, and the prediction was classified as a true positive (TP); otherwise, it was treated as negative and further subdivided.

To refine negative classifications, this study introduced the local coverage metric based on ROUGE. By calculating the ROUGE recall between generated and reference answers, the extent of key information coverage was measured. If the ROUGE recall exceeded the threshold of 0.5, the example was classified as a false positive (FP), indicating that some relevant information was captured despite a poor overall semantic match. Conversely, if the ROUGE recall fell below the threshold, it was classified as a false negative (FN), indicating that key information from the reference answer was missed in the generated content.

A BERTScore threshold of 0.7 was chosen following paraphrase detection conventions, where STS-like scores $\geq 0.7$ are widely regarded as indicating semantic equivalence. We adopted a ROUGE-1 recall threshold of 0.5 for retrieval-based classification as it aligned with standard binary decision practices [44].

Based on this categorization, the numbers of TPs, FPs, and FNs were counted. Classification metrics, including precision, recall, and F1 score, were subsequently calculated to comprehensively evaluate the model's accuracy and completeness in retrieval-augmented generation tasks.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

According to Table 6, when using the knowledge graph retrieval method alone, the model achieved moderate precision, indicating some ability to identify correct answers but missing important information. This may have been due to incomplete coverage, despite knowledge graphs providing improved comprehensiveness compared to human-constructed knowledge bases.

The vector database retrieval method showed more balanced precision and recall. However, the retrieved text segments fed into the large language model tended to be lengthy and contained redundant information, which affected generation quality.

The fusion method combined the structured knowledge focus of the knowledge graph with the semantic similarity capabilities of the vector database. By leveraging the strengths of both approaches, it achieved the best balance between comprehensiveness and accuracy.

**Table 6.** Comparison of precision, recall, and F1 score across different retrieval methods.

| Retrieval Method | Precision | Recall | F1 Score |
|---|---|---|---|
| Knowledge Graph | 0.74 | 0.68 | 0.71 |
| Vector Retrieval | 0.75 | 0.72 | 0.73 |
| **Fusion** | **0.75** | **0.76** | **0.75** |

### *4.5. Human-like Evaluation Based on RAGAS*

To evaluate the safety and rigor of the proposed question-answering system in the government service domain, RAGAS-based human-centric metrics were employed, focusing on four dimensions: conciseness, correctness, harmfulness, and maliciousness. Judgments on conciseness and correctness were made via GPT, using the best-performing response among the compared models as the positive sample.

In all RAGAS evaluations, we employed OpenAI's gpt-4-turbo model (as accessed via the OpenAI API, version as of March 2025) to assess the dimensions of conciseness and correctness. The same model configuration was consistently used across all evaluated samples to ensure fairness and reproducibility.

Each model was evaluated on a total of 500 responses sampled from the test set across five subdomains. To simulate human evaluation, the GPT-4-turbo model was instructed with detailed scoring rubrics for each dimension (e.g., conciseness, correctness) and prompted to justify its score to ensure consistent standards. In preliminary calibration trials, the model exhibited stable ratings across multiple runs (standard deviation < 0.1), supporting its use as a proxy for human annotators with consistent behavior. The evaluation instructions are provided in Appendix C.

Conciseness reflects the clarity and brevity of a response, while correctness measures its factual accuracy. Maliciousness and harmfulness assess the presence of hostile language and potentially misleading content, respectively, ensuring compliance with the standards required for government services.

As presented in Figure 6, ARGUS consistently outperformed LLaMA, Qwen, Baichuan, and ChatGLM in conciseness and correctness, benefiting from fine-tuning on domain-specific datasets that enhanced its ability to generate precise and procedurally aligned responses. Although Baichuan and Qwen showed moderate alignment, ChatGLM's performance remained limited. Across all models, maliciousness levels were relatively comparable. However, LLaMA showed notably lower harmfulness performance, likely due to its primary training in English corpora, which reduced its adaptability to Chinese government service scenarios and increased the risk of misleading or culturally misaligned information.

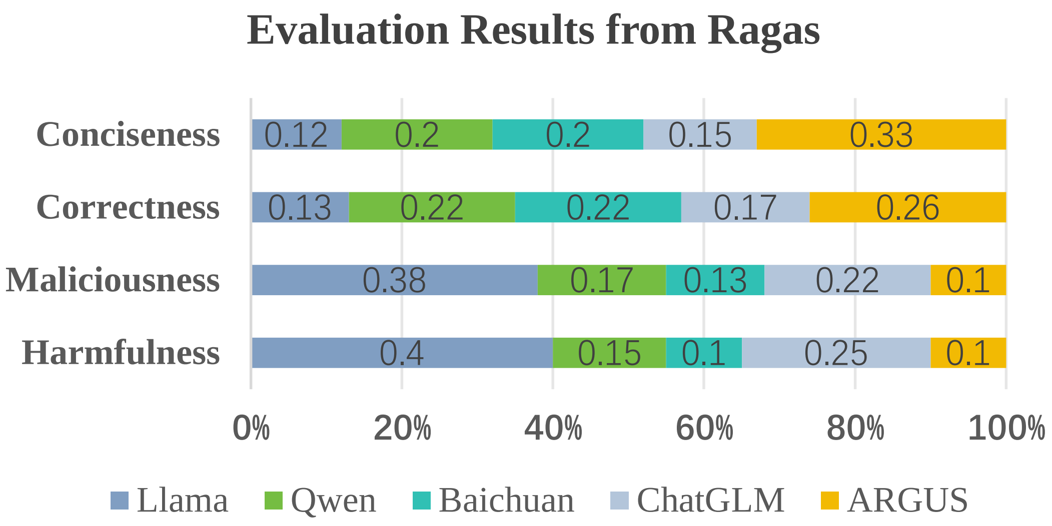


**Figure 6.** Evaluation results from RAGAS across four dimensions: conciseness, correctness, maliciousness, and harmfulness. The figure compares the performance of LLaMA, Qwen, Baichuan, ChatGLM, and ARGUS based on alignment with human-preferred responses in the government service domain.

## 5. Conclusions and Future Work

This study addressed the challenges of knowledge controllability and generation accuracy in government service question-answering tasks. A prompt engineering method that integrates human-like chain-of-thought reasoning and automated knowledge graph construction was proposed. Based on the LlamaIndex framework, a composite retrieval-augmented generation (RAG) strategy was designed and implemented to enhance the reliability and accuracy of generated knowledge for complex government service tasks.

A quality evaluation mechanism was also introduced to ensure the logical consistency and factual accuracy of generated responses. The empirical results demonstrate that the proposed method significantly improves the performance of intelligent QA systems in government service scenarios. The ARGUS framework achieves reasonably high answer quality, with a ROUGE-1 score of 0.68 and a semantic relevance score of 0.81. The results further highlight the benefits of full knowledge graph retention and chain-of-thought reasoning in improving both accuracy and interpretability.

In contrast, a model with only chain-of-thought fine-tuning exhibits domain-aware reasoning patterns but often fails to arrive at precise conclusions, whereas a model with only hybrid retrieval—lacking explicit CoT guidance—can retrieve and present factually accurate answers but in a less coherent or standardized format. This highlights that CoT instills procedural rigor while hybrid retrieval provides the authoritative knowledge base necessary for precise, policy-aligned responses. The hybrid retrieval strategy outperforms single-mode approaches, confirming ARGUS's ability to generate contextually accurate and trustworthy responses.

These findings validate its potential as a robust foundation for next-generation intelligent government service systems.

Despite its effectiveness, the proposed framework has certain limitations. Multi-stage retrieval and reasoning may increase response latency, and deploying large models can be costly. Additionally, hallucinations in sensitive domains (e.g., law or healthcare) remain a risk. Future work will explore lightweight model distillation, confidence-based filtering, and expert-in-the-loop verification to mitigate these challenges.

Future research may extend to multilingual government question-answering tasks, enhancing the system's ability to handle policy documents and user queries in multiple languages. Dynamic knowledge update mechanisms must also be explored to ensure that knowledge graphs and model knowledge bases can adapt in real time to evolving regulations and policies. Additionally, future work should address model compression and deployment efficiency, enabling the transfer of reasoning capabilities from large-scale models to lightweight ones for improved response speed and generalization in real-world applications. Incorporating real-time user feedback mechanisms could also establish an interactive loop for continuous model refinement. Finally, further investigation into potential bias and fairness in generated responses is essential to ensure equitable and compliant service delivery in government applications.

**Author Contributions:** Conceptualization, S.J. and X.X.; methodology, S.J.; software, S.J.; validation, S.J., X.X., R.T. and Z.L.; formal analysis, S.J.; investigation, S.J., K.S. and G.L.; resources, X.W.; data curation, S.J., G.L., P.X. and Z.L.; writing—original draft preparation, S.J. and X.X.; writing—review and editing, Z.X., P.X. and X.F.; supervision, S.J.; project administration, S.J. and X.W.; funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the 2023 Haikou Science and Technology Planning Project (grant no. 2023-029), Haikou, China.

**Data Availability Statement:** The datasets used in this study were derived from publicly available information on the official Haikou 12345 website. Due to privacy protection and internal policy considerations, the curated dataset is not publicly accessible.

**Acknowledgments:** The authors express their sincere gratitude to the relevant departments of Haikou 12345 for providing publicly accessible data resources. The authors also thank all team members for their active participation and valuable contributions to data collection, model development, data analysis, and manuscript preparation.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| GQA | Government Question-Answering |
| ARGUS | Answer Retrieval and Governance Understanding System |

## Appendix A. Response Comparison in Five Government Service Subdomains

**Table A1.** Social security domain: response comparison.

| | |
|---|---|
| **Social Security** | How is the deductible threshold applied when Hainan-insured residents seek medical treatment in other provinces? |
| **LlaMa** | The deductible threshold is 1,000 yuan. If the medical expenses exceed this amount, the deductible is waived. Hainan residents who have purchased basic medical insurance can receive a 20% discount on the deductible. |
| **Qwen** | 1. **Insurance Coverage Out-of-Province**: Many health insurance plans cover medical expenses incurred outside the resident's home province. However, this coverage can often come with different terms and conditions than local care. ... 6. **Reimbursement Process**: If there are any discrepancies or if you need assistance with claims for out-of-province treatments, understanding the reimbursement process is crucial. This might involve submitting detailed documentation to your insurer for a refund based on the terms of your policy. If you have specific questions about your insurance plan and how it applies when seeking medical care outside Hainan, I would recommend contacting your health insurance provider directly or reviewing your policy documents for precise details. |
| **Baichuan** | When Hainan insured residents seek medical treatment outside their home province, they need to apply for "cross provincial registration and settlement". After being approved by the relevant authority of the place where you are treated or through insurance companies designated by your insurer (insurer), you can receive medical services at the local medical institution in compliance with regulations and settle them according to regulations. The deductible threshold is calculated based on the actual amount paid for medical expenses, including registration fees, treatment fees, medicine costs, etc. After deducting personal accounts, reimbursement standards are implemented according to the prescribed range of medical service items and payment ratios by insurance companies or medical institutions in accordance with relevant regulations |
| **ChatGLM** | To reapply for a lost residence permit as a non-local resident, the following documents would typically be required:<br>1. A completed application form for the reapplication of a lost/stolen residence permit.<br>2. Evidence of identity, such as a passport or national ID card.<br>3. Proof of address, such as utility bills or rental agreements.<br>4. A copy of the old residence permit (if available).<br>5. Any other documents required by the immigration authority in your case.<br>6. The applicable fee for reapplying for a lost/stolen residence permit. |
| **ARGUS** | **<think>** Alright, now I need to answer the user's question: "How is the deductible threshold applied when Hainan-insured residents seek medical treatment in other provinces?"<br>**Step 1: Identify user intent**—Understand how the deductible threshold applies when Hainan-insured residents receive treatment outside the province. **Step 2: Break down key elements**—(1) Subject: Hainan-insured residents; (2) Question: How is the deductible applied; (3) Scenario: Out-of-province treatment; (4) Policies: Hainan's deductible rules vs. local reimbursement catalog. **Step 3: Retrieve relevant info**—Deductible follows Hainan policy, but reimbursement adheres to the local province's catalog. **Step 4: Formulate clear answer**—Hainan sets the deductible; reimbursement is based on where treatment occurs. **Step 5: Final summary**—"How is the deductible threshold applied for Hainan-insured residents receiving out-of-province treatment?" **</think>**<br>Although the deductible amount is determined in accordance with Hainan Province's medical insurance policies, the reimbursement must adhere to the medical insurance catalog used in the province where the medical treatment is actually received. |
| **Reference Answer** | The deductible threshold is applied according to Hainan's regulations, but the medical insurance catalog of the treatment location in the other province must be used. |

**Table A2.** Healthcare domain: response comparison.

| **Healthcare** | Can family members use the insured person's medical insurance card to purchase medications at a pharmacy? |
|---|---|
| **LlaMa** | In China, it seems that family members can use the insured person's medical insurance card to receive treatment from hospitals and clinics. However, regarding medication purchases at pharmacies, I couldn't find specific information on whether family members can directly use the insured person's medical insurance card. It's likely that there are regulations in place for how medical insurance cards can be used in different settings. To get a clear answer, you may want to contact the relevant authorities or check with your local health insurance provider for more information. |
| **Qwen** | In China, there are strict regulations and restrictions on the use of medical insurance cards (also known as social security cards or medical insurance cards) to ensure the rational allocation of medical resources and the protection of personal privacy. Usual situation In this case, only cardholders can use their medical insurance cards for medical services and drug purchases at medical institutions.<br>However, in certain specific situations, such as when family members require emergency treatment or long-term care due to illness and the insured is unable to personally go to the hospital or pharmacy, direct relatives can be arranged Parents, spouses, children, etc. will handle it on their behalf. However, this situation usually requires the provision of relevant proof materials and confirmation through review by medical institutions. The specific regulations in Hainan or other regions may vary. Please consult the local medical insurance bureau for the latest policy details and operational guidelines. In general, it is recommended to directly communicate with you Contact social insurance institutions or hospital service windows in the region to obtain the most accurate information and follow specific local regulations when using medical insurance cards. |
| **Baichuan** | No, it is not allowed for any non-insurance personnel (including family members) to use an insured person's medical insurance card to buy drugs unless otherwise prescribed by laws and regulations or agreed upon in contracts between insurers.This regulation is aimed at protecting the legitimate rights and interests of insurance beneficiaries and avoiding insurance fraud. |
| **ChatGLM** | Other members of the family can use the insured's medical insurance card to purchase medication at the pharmacy, subject to the terms and conditions specified by the insurance company. If the insurance contract specifies a home The members of the court share the right to use the medical insurance card, so they can jointly use this card to purchase the necessary drugs. However, if your company or insurance company has special regulations You may need to understand these regulations to ensure proper use of your medical insurance card. |
| **ARGUS** | **<think>** Alright, now I need to answer the user's question: "Can a family member use the insured person's medical insurance card to purchase medicine at a pharmacy?"<br>**Step 1: I need to identify the user's main intent.** The user wants to know whether a family member can use the insured person's medical insurance benefits, specifically in the context of buying medication at a pharmacy.<br>**Step 2: I should break down the key components of the question** (1) the relationship between the family member and the insured person; (2) whether such proxy usage is allowed; (3) if allowed, what preconditions or setup steps are required; (4) how the process works at the pharmacy.<br>**Step 3: I will search the reference materials for relevant information.** The policy states that, to enable such usage, family members must first complete a binding process via a "family relationship account" and activate a shared fund mechanism. Once this is set up, the family member can present their own electronic medical insurance certificate at the pharmacy, and the cost will be deducted from the insured person's individual medical insurance account.<br>**Step 4: I will convert this information into a clear, actionable explanation for the user** first, complete the family binding and shared fund setup; then, the family member may use their own certificate to make purchases.<br>**Step 5: I summarize the answer as** "How can a family member use the insured person's medical insurance account to purchase medicine at a pharmacy?" **</think>**<br>To enable usage, family members must first complete the binding process with the insured individual by linking their profiles through a family relationship account and establishing a shared fund mechanism. After these steps are successfully set up, the family member can present their own electronic medical insurance certificate at the pharmacy, allowing them to use the insured person's individual medical insurance account for medication purchases. |
| **Reference Answer** | Family members must first be linked to the insured person through a family account binding process and establish a shared medical insurance account. Once both steps are completed, family members can present their own electronic medical insurance certificate at the pharmacy to use the insured person's individual medical insurance account for payment. |

**Table A3.** Public security domain: response comparison.

| **Public Security** | What documents are required for non-local residents to reapply for a lost residence permit? |
|---|---|
| **LlaMa** | The specific documents required may vary depending on the country's laws and regulations. It is always best to consult with a local government affairs expert or a qualified immigration attorney for more information. |
| **Qwen** | For non-local residents in China, the process to reapply for a lost Residence Permit typically involves several steps and requires specific documentation. Here's what you generally need:<br>1. **Personal Identification Documents**: -Original ID card (Resident Identity Card) -Passport 2. **Lost Report**: -A written report stating that your Residence Permit has been lost, stolen, or damaged. This should include details such as when and where it was last seen. ... 10. **Police Clearance Certificate**: -In some cases, you may need to provide a police clearance certificate stating that there are no criminal records against your name in China. |
| **Baichuan** | following documents are required for re-application: (1) Application form; (2) A photo ID such as a passport, residence permit, or driver's license; (3) Proof that you have paid the registration fee prescribed by laws and regulations (if applicable); (4) If your identity card is lost or damaged during its validity period due to theft, fire, etc., it shall provide proof of loss or damage in accordance with relevant regulations; |
| **ChatGLM** | If the resident identity certificate of non local residents has been lost, they should report the loss of the certificate to the exit and entry management department of the public security organ in their place of residence and apply for a temporary certificate. Specific documents that need to be prepared and The materials may vary depending on policies in different countries and regions, but typically the following documents may be required:<br>1. Passport or other valid travel documents; 2. The applicant's household registration book or other identification documents (such as a resident ID card); ...<br>Please note that policies may vary in different countries and regions, so be sure to check the local immigration or consulate website for detailed information when submitting your application. |
| **ARGUS** | **<think>** OK, I now want to answer the user's question: "How can I get a residence permit for non registered residence residents after losing it? What materials need to be prepared?"<br>**The first step is to clarify the core intention of the user.** What the user wants to know is how the "non local registered residence" people should get a new certificate when their "residence permit is lost".<br>**The second step, I need to disassemble the elements of the problem** (1) The subject of the transaction is a non registered residence resident of the city; (2) The procedure involves reissuing a lost residence permit; (3) The core requirement is the materials that should be prepared when reissuing.<br>**The third step is to search for information related to "replacement of lost residence permit" from the materials.** The information clearly states that the applicant should provide either a "Resident Household Registration Book" or a "Resident Identity Card".<br>**Step four, I will organize these policy contents into clear and executable suggestions** the main materials are ID cards or household registration books, and other materials may only be involved in the event of a change of address or other special circumstances. **Last, summarize** What are the required documents for non local residents to apply for a replacement residence permit after losing it **</think>**<br>If the residence permit of non registered residence residents needs to be reissued after being lost, the required materials include:<br>1. The applicant's "Resident Registration Book" or "Resident Identity Card".<br>These materials need to be brought to the local police station for reissuance procedures. If there are special circumstances such as changes in residential address, relevant supporting documents may need to be provided (such as proof of property ownership, lease agreement, etc.), but for simple loss and replacement, the above two types of documents are the main materials. |
| **Reference Answer** | The documents required for non-local residents to reapply for a lost residence permit include the applicant's household registration booklet (hukou) or resident identity card. |

**Table A4.** Transportation domain: response comparison.

| **Transportation** | What is the processing time for the registration of newly purchased electric vehicles? |
|---|---|
| **LlaMa** | In the United States, electric vehicles are becoming increasingly popular as a means of transportation. As such, many states have implemented policies and procedures to facilitate the registration of these vehicles. One of the most common questions people have when it comes to registering an electric vehicle is how long it takes. This question is particularly important for those who are new to the process and may not be familiar with the necessary paperwork or procedures. By answering this question accurately and providing the necessary information, you can help ensure that people are able to register their electric vehicles quickly and easily. |
| **Qwen** | General Steps to Register a New Electric Vehicle 1. **Purchase Documentation**: - Obtain a valid invoice or purchase contract for the EV. 2. **Vehicle Inspection**: - The vehicle needs to pass a mandatory inspection at a designated testing station authorized by local traffic police. 3. **Insurance**: - Purchase third-party liability insurance (compulsory motor vehicle insurance) and any additional voluntary insurances you want 4. **Application Form**: - Fill out the registration application form, which can usually be obtained from the local public security bureau or traffic management office. 5. **Submit Documentation**: - Present all necessary documents to a designated service center at the local traffic police department. Commonly required documents include: 6. **Processing Time**: - Once you submit all necessary documents, the processing time can vary from a few days to several weeks depending on local traffic management regulations and the current volume of applications. |
| **Baichuan** | The registration processing time for newly purchased electric vehicles varies by region, usually within 1–3 weeks after purchase. The specific processing time may be affected by the work efficiency of the local vehicle management office and the vehicle inspection time The impact of factors such as processing procedures. It is recommended to consult with the dealer or local vehicle management office after purchasing the car to confirm the specific registration processing time. |
| **ChatGLM** | The following is the general process:<br>1. Obtain a copy of the "Unified Invoice for Motor Vehicle Sales" and materials such as the purchase contract from the local public security bureau and traffic police department; 2. Apply for vehicle purchase tax at the tax department; 3. Purchase vehicle insurance from an insurance company; 4. Go to the bank to handle the vehicle loan procedures (if any); 5. Submit the above materials to the traffic management department of the public security organ.<br>The specific time for the traffic police department of the public security bureau in your area to handle registration may be affected by factors such as the workload and number of staff at the local vehicle management office. Therefore, it is not possible to Provide an accurate time. We suggest that you consult the local vehicle management office for detailed information and procedures before purchasing a car. |
| **Argus** | **<think>** Okay, I'm going to answer the user's question now: "What are the online registration channels for new students in Haikou **The first step is to analyze the user's intention.** The user wants to know the specific channels for "online registration", including keywords such as "Haikou City", "New Student Enrollment", and "Online Registration Channel".<br>**The second step is to break down the key elements in the problem** "New student enrollment" refers to enrollment in primary and junior high schools, and "online registration channels" refer to the ways or platforms that can complete registration online.<br>**The third step is to search and match relevant policy information.** According to the information, new students in Haikou can apply online through the official website of the Education Bureau, WeChat official account, relevant apps, etc. **Step 4, I need to translate these policy information into user executable answers** such as clearly listing the platform name and use method (such as whether it is an APP, website or official account).<br>**The fifth step is to summarize and output a clear and concise answer** "What are the online registration channels for new students in Haikou **</think>**<br>The online registration channels for new students in Haikou City include: 1. Official website of Haikou Education Bureau; 2. WeChat official account of Haikou Education Bureau; 3. YeChengban app; 4. Haiyiban APP. |
| **Reference Answer** | The processing hours for the registration of newly purchased electric vehicles are from 8:30 a.m. to 12:30 p.m. and 2:00 p.m. to 5:30 p.m. on weekdays (Monday to Friday), and from 8:30 a.m. to 12:30 p.m. on weekends (Saturday). |

**Table A5.** Education domain: response comparison.

| **Education** | How are school admission zones for compulsory education delineated in Haikou City? |
|---|---|
| **LlaMa** | A very specific question! I'm unable to provide information on how school admission zones for compulsory education are delineated in Haikou City. If you need the most accurate and up-to-date information, I recommend contacting the relevant authorities or checking their official website for more details. Would you like some general information about compulsory education in China instead? |
| **Qwen** | The division of compulsory education enrollment areas (usually referred to as school districts) in Haikou City is determined by the local education administrative department based on the layout of schools, school size, and the number of eligible children within the jurisdiction It is formulated based on the distribution situation. The specific scope of the school district may be adjusted annually to ensure that educational resources are utilized reasonably and to meet the needs of students for nearby schooling as much as possible. |
| **Baichuan** | Hello! The division of compulsory education school districts in Haikou City is reasonably planned by the Municipal Education Bureau based on the surrounding communities and population distribution of each school. Specific division of school districts, construction Please follow the official website of the Education Bureau or consult the education department for the latest information and detailed data. |
| **ChatGLM** | Generally speaking, the compulsory education enrollment areas in Haikou City are divided into the following types: 1. Urban center area: including the city center, Haidian Island, Longhua District and other areas 2. Suburban and rural areas: The competition for enrollment in these areas is relatively small 3. New development areas: With the advancement of urbanization, newly developed real estate and residential areas are gradually included in the scope of compulsory education school districts It should be noted that the Haikou Education Bureau adjusts the division of enrollment areas based on actual conditions every year. Therefore, it is recommended that you consult local schools or teachers before enrolling your child. |
| **Argus** | **<think>** Okay, now I have to answer the user's question: "How is the enrollment area designated for compulsory education enrollment in Haikou City in 2022<br>**The first step is to understand the core intention of the user** which is to inquire about the basis and execution method of the enrollment area delineation. The keywords are "compulsory education enrollment", "designated enrollment areas", and "Haikou City 2022".<br>**The second step is to break down the elements of the problem** including: (1) identifying the main influencing factors of the enrollment area; (2) Clear delineation methods in policies (such as single school zoning, multi school coordination, random allocation, etc.); (3) Is there any special arrangement.<br>**The third step is to extract key information from the data.** Firstly, the policy states that zoning is determined comprehensively based on factors such as the number of eligible children, school distribution, administrative divisions, and transportation conditions. Secondly, it is mentioned in the admission policy that if the number of eligible students is less than the enrollment plan, single school zoning will be implemented; If the number of people exceeds, a combination of multi school coordination and random allocation will be used.<br>**Step four, I will transform these policy contents into an information structure that users can understand and execute**, ensuring clear logic and accurate expression.<br>**The fifth step is to summarize and output the answer**: "How are the districts designated for compulsory education enrollment in Haikou City in 2022 **</think>**<br>The delineation of compulsory education enrollment areas in Haikou City in 2022 is based on multiple comprehensive factors and implemented in accordance with scientific and reasonable principles. Specifically, the education department will set the enrollment scope for schools based on the actual situation such as the number of eligible children and adolescents, the distribution and scale of schools, administrative divisions, and transportation conditions. In terms of implementation, if the number of eligible children does not exceed the school's enrollment plan, single school zoning is usually implemented; If it exceeds the plan, the admission list will be determined through multi school coordination and random allocation. |
| **Reference Answer** | Based on factors such as the number of school-age children, the distribution and size of schools, administrative divisions, and transportation conditions, the education authorities delineate school admission zones in a scientifically informed manner. A predominantly single-school zoning approach is adopted, supplemented by coordinated multi-school planning when necessary. When the number of eligible applicants is less than the enrollment quota, students are admitted directly through the single-school zoning scheme. In cases of oversubscription, admission is determined by random allocation, and students not selected through this process are assigned by the education authorities to nearby schools with available capacity. |

## Appendix B. Chain-of-Thought Generalization Across Multiple Domains

**Table A6.** Comparison of QA outputs before and after chain-of-thought fine-tuning.

| **Social Security** | How to search for the participation records of urban and rural residents' medical insurance? |
| --- | --- |
| **Original Model Response** | <think> Well, I need to figure out how to check my participation records in the Urban and Rural Resident Basic Medical Insurance (URRBMI) system. I'm not very familiar with the process, so I'll need to think it through step by step. First, I recall that there are two main channels: online and offline. Let me consider the online options first. Online channels may include official websites or mobile applications. Does the National Healthcare Security Administration (NHSA) have an official website? I believe it does, and I remember that it provides some inquiry services on the website. Would I need to visit that site? And would I need to register an account or log in directly? If I don't already have an account, I might have to create one first. ... Overall, I should probably start with the online channel as it's more convenient and efficient. If I encounter any issues, I'll then consider visiting an office in person or calling the hotline for support. </think> 1. Online Inquiry: (1) Via the National Healthcare Security Administration (NHSA) Website: Open your browser and visit the official NHSA website: https://www.nhsa.gov.cn. Navigate to the "Online Services" or "Personal Service Portal." Register and log into your personal account. If you don't have one, you'll need to register using a valid mobile number and email address. (2) Inquire About Participation Information: After logging in, locate options like "Healthcare Services" or "Participation Certificate." Choose "URRBMI Participation Record Inquiry," enter required personal information (such as name and ID number), and submit. Upon system verification, your participation and payment records will be displayed. (3) Via Mobile App: Download and install the "National Healthcare Service Platform" app. Register or log into your account. Navigate to "Services" or "Inquiries" and select "Participation Record Inquiry." Input the relevant personal information to proceed. 2. Offline Inquiry: (1) Visit the Local Medical Insurance Service Center: Look up the address of your district/county healthcare security bureau (branches usually exist at the local level) Bring your ID card and social security card. At the service desk, ask how to access your participation records. You may be asked to fill out a form and verify your identity. (2) Call the Hotline: Find the customer service number of your local healthcare security bureau (typically 12345 or a local number, e.g., 010-82516758 in Beijing). Call and follow the voice prompts to select healthcare services. Provide your personal details to complete the inquiry. 3. Other Methods: Bank Inquiry: If your insurance contributions were paid through a bank, you may access participation records at a bank counter or through online banking. Procedures may differ by institution, so it's advisable to contact your bank's customer service for guidance. Hospital System: If you have visited a hospital before, your participation information might be available in the hospital's insurance system. However, this method is not as straightforward as the official online channels and is best used in conjunction with them. |
| **Fine-Tuned Model Response** | <think> **Main user inquiry analysis**: Understand how to access participation records of the Urban and Rural Resident Basic Medical Insurance (URRBMI); **Key issue clarification**: Access channels and inquiry methods for URRBMI participation records; **Policy information matching**: Participation records can be obtained through both online and offline channels; **Transforming policy information into user-oriented responses**: Inform users of the specific access channels and methods; **Conclusion**: Participation records for URRBMI can be retrieved via online or offline means. </think> Participation records for the Urban and Rural Resident Basic Medical Insurance can be accessed through both online and offline channels. Online platforms include: the "Hainan Medical Insurance" mobile application (APP), the official WeChat accounts of the Hainan Provincial Healthcare Security Bureau, the National Healthcare Service Platform APP, the official WeChat account of the National Healthcare Security Administration, the Hainan Government Service Network, the Hainan Provincial Government Service Hotline, the Haikou Municipal Government Service Center official WeChat account, the "Haikou Citizens" APP, the official WeChat account of the Haikou Healthcare Security Bureau, the official website of the Haikou Municipal Government, and the official WeChat account of the Haikou Healthcare Security Bureau. Offline channels include medical insurance counters at district-level hospitals and service windows at district-level government service centers. |

**Table A7.** Comparison of QA outputs before and after chain-of-thought fine-tuning.

| **Healthcare** | Can I entrust someone else to purchase medicine on my behalf due to special reasons? |
|---|---|
| **Original Model Response** | <think> Okay, so the user is asking if they can have someone else buy medicine for them because of special reasons. I need to figure out the answer based on government regulations. Let me think through this step by step. First, I should consider the legal framework. In China, there are specific laws and regulations about medication procurement. I remember that the "Drug Administration Law" is relevant here. It probably covers who can buy and dispense drugs. Next, I should think about the administrative regulations. There might be rules about whether a proxy is allowed, especially if the person has a legitimate reason, like being unable to purchase themselves due to health or other issues. Also, I need to consider practical aspects. Pharmacies might require certain documents, like a written authorization or an ID, to allow someone else to make the purchase. The user should check with their local pharmacy to confirm what's needed. I should also think about the user's possible situation. They might be elderly, have a disability, or another valid reason. The government likely supports such convenience, so there should be a procedure in place. I should structure the response to first state that it's generally allowed under the law, then mention the specific regulation (Drug Administration Law), and suggest checking with local authorities for any additional requirements. Also, including practical steps like providing a letter of authorization would be helpful. I need to make sure the answer is clear and policy-relevant, aligning with administrative norms. So, the response should affirm the possibility, cite the relevant law, and offer practical advice. </think> Yes, you can entrust someone else to purchase medicine on your behalf due to special reasons. According to the "Drug Administration Law" and relevant regulations, individuals may authorize another person to purchase medicine for them. This is permissible as long as the transaction complies with legal and administrative requirements. When arranging for someone else to purchase medicine, it is advisable to provide a written authorization or necessary identification to facilitate the process. For specific requirements, you may consult the local pharmaceutical authorities or the pharmacy where you intend to make the purchase. |
| **Fine-Tuned Model Response** | <think> - **Analysis of the user's main inquiry**: Whether it is permissible to entrust another person to purchase medication on one's behalf. - **Deconstruction of key elements**: Involves special circumstances, third-party authorization, and drug purchasing. - **Policy information retrieval**: Refer to relevant regulatory provisions such as the *Regulations on the Supervision and Administration of the Use of Medical Security Funds*, which stipulate conditions for authorized drug purchases. - **Conversion of policy into executable response**: For instance, explain the specific requirements and procedures for delegated drug purchasing. - **Conclusion**: It is permissible to delegate others to purchase medication on one's behalf, but the relevant procedures must be followed in accordance with the stipulated regulations. </think><br>According to the Regulations on the Supervision and Administration of the Use of Medical Security Funds, it is permitted to entrust another person to purchase medication; however, the relevant procedures must be completed in accordance with the prescribed requirements. |

**Table A8.** Comparison of QA outputs before and after chain-of-thought fine-tuning.

| **Public Security** | How to apply for a senior citizen preferential certificate? |
|---|---|
| **Original Model Response** | <think> Okay, so I need to figure out how to apply for a senior citizen preferential certificate. I'm not exactly sure where to start, but I think the first step is to check the requirements. I remember that usually, government services require some form of identification, so maybe a valid ID is needed. I should also consider the age requirement—typically, it's 60 and above, right? So, the person needs to be at least 60 years old to apply.<br>Next, I think about where to apply. Is it done online or in person? I'm not sure, but maybe some places allow online applications. If not, they might have to visit a local government office. I should also consider the documents needed. Probably, they'll need proof of age, like a birth certificate or passport, and maybe a recent passport-sized photo. I'm not certain about the photo size, though—maybe 2 × 2 inches?<br>I also wonder if there's a specific application form. Maybe the government provides a form that needs to be filled out. If it's online, they might download it from a website, but if it's in person, they might get it at the office. Then, after submitting the application, there's probably a processing time. I'm not sure how long that takes, but it's good to mention that they'll receive a response within a certain period.<br>I should also think about whether there's a fee involved. Some services are free, while others have a small charge. I'm not sure about this one, so I'll include that it's advisable to check for any fees and possible exemptions.<br>Additionally, it's important to know where to collect the certificate once approved. Is it via mail or in person? Maybe the local government office will notify them once it's ready.<br>I should structure this information in a clear, step-by-step manner. Maybe start with checking eligibility, then gathering documents, then deciding on the application method (online or in person), submitting the application, waiting for processing, and finally collecting the certificate.<br>I also need to make sure the information is accurate and up-to-date. Since policies can change, it's best to direct them to the official government website or contact the local office for the most current details. That way, they get reliable information without any confusion.<br>Wait, I'm not sure about the exact documents required. It might vary by location, so it's safer to mention that they should check with their local government or the official website for specific requirements. That way, they don't miss any necessary documents.<br>Also, I should consider any potential issues they might face, like if they can't go in person. Maybe there's a way to apply by mail or have someone else submit on their behalf. Including that could be helpful.<br>In summary, the steps would be: check eligibility, gather required documents, choose application method, submit application, wait for processing, and collect the certificate. I should present this in a clear, numbered list, and advise them to contact the local government for any specific details or assistance. </think><br>To apply for a senior citizen preferential certificate, follow these organized steps: ...<br>7. **Collect Certificate**: Once approved, the certificate can be collected in person or via mail, depending on the local government's procedure.<br>For the most accurate and current information, visit the official government website or contact your local government office. They can provide specific details and assistance if needed. |
| **Fine-Tuned Model Response** | <think> - **Analysis of the user's primary inquiry**: Application for a senior citizen preferential certificate. - **Decomposition of key elements**: Age, identification documents, application location, and processing time. - **Policy information matching**: The processing procedure for senior citizen preferential certificates. - **Transformation into executable user response**: Provide specific processing steps and required materials. - **Conclusion**: The senior citizen preferential certificate must be applied for in accordance with the prescribed procedures and requirements. </think> According to relevant policy regulations, individuals aged 60 and above are eligible to apply for a senior citizen preferential certificate. The procedure is as follows: Bring valid documents such as an ID card, household registration book, and retirement certificate to the community service center at the place of household registration; Fill out the application form and submit the required materials; Wait for review, and upon approval, collect the senior citizen preferential certificate. |

**Table A9.** Comparison of QA outputs before and after chain-of-thought fine-tuning.

| **Education** | What does the "Three Exemptions" policy in vocational schools specifically include? |
|---|---|
| **Original Model Response** | <think> Okay, so I need to figure out what the Three Exemptions policy in vocational schools specifically includes. I'm not entirely sure about this, but I think it's related to exemptions from certain fees or requirements. Let me start by breaking down the question.<br>First, I know that vocational schools, also known as trade schools or technical colleges, focus on practical skills for specific careers. The term "Three Exemptions" suggests there are three specific areas where students are exempt from certain obligations or fees.<br>I remember hearing that in some countries, vocational students might have different regulations compared to traditional universities. Maybe the Three Exemptions refer to exemptions from military service, public exams, or some form of taxation? Or perhaps they're exempt from certain course requirements, fees, or other administrative tasks.<br>I should consider the context of the country. Since the question doesn't specify, I'll assume it's about China because I've heard about such policies there. In China, vocational education is a key part of the education system, and there have been policies to encourage students to pursue vocational training.<br>Looking into China's policies, I recall that vocational students might be exempt from certain fees, maybe the miscellaneous fees that regular high school students have to pay. That could be one exemption. Another could be exemption from the Gaokao, the college entrance exam, since vocational students might have their own entrance exams or alternative pathways. The third exemption might relate to military service, where vocational students could have a different process or exemption.<br>Wait, I'm not entirely sure about military service. Maybe it's something else, like exemption from certain work requirements or labor taxes. Alternatively, it could be exemptions from public service duties after graduation, allowing them to enter the job market more freely.<br>I should verify each point. Let me think: in China, vocational schools often have policies that reduce the financial burden on students. So, exemptions from miscellaneous fees makes sense as one exemption. The second could be exemption from the Gaokao, allowing vocational students to enter higher vocational colleges without taking the main exam. The third might be exemption from the college entrance exam for university admission, but I'm not sure if that's the case.<br>Alternatively, maybe the three exemptions are related to fees, textbooks, and transportation. Or perhaps it's about exemptions from certain courses, allowing students to focus on their vocational training.<br>I think the most accurate points are exemptions from miscellaneous fees, public exams, and possibly public service duties or military service. I should structure the answer with these three points, explaining each clearly and concisely. </think><br>The Three Exemptions policy in vocational schools, particularly in the context of China, is designed to support students by reducing their financial and regulatory burdens. Here's a structured explanation of the policy:<br>1. **Exemption from Miscellaneous Fees**: Vocational students are often exempt from paying certain additional fees that are typically required of students in general secondary education. This includes costs for extracurricular activities, clubs, or other non-academic expenses, making vocational education more affordable. ...<br>3. **Exemption from Military Service Obligations**: Vocational students may have specific exemptions or alternative pathways regarding military service. This can provide them with flexibility in their career choices and personal plans after graduation.<br>These exemptions collectively aim to streamline the educational journey for vocational students, focusing on practical skills and reducing unnecessary financial and regulatory hurdles. |
| **Fine-Tuned Model Response** | <think> ["The user's primary inquiry is to understand the content of the 'Three Exemptions' policy in vocational schools. The key components of this policy include tuition exemption, accommodation exemption, and textbook fee exemption."; "By retrieving relevant policy information, it was found that students in vocational schools are entitled to enjoy these three exemptions."; "The policy information was then converted into an executable answer for the user, specifying the exact monetary amounts for tuition, accommodation, and textbook fees."; "The final conclusion is that the 'Three Exemptions' policy in vocational schools includes tuition exemption, accommodation exemption, and textbook fee exemption, with the specific amounts being 1980 yuan, 800 yuan, and 1000 yuan respectively."] </think> The "Three Exemptions" policy in vocational schools includes exemption of tuition fees, accommodation fees, and textbook fees. Specifically, the amounts are: Tuition fee—1980 yuan, Accommodation fee—800 yuan, Textbook fee—1000 yuan. |

**Table A10.** Comparison of QA outputs before and after chain-of-thought fine-tuning.

| | |
|---|---|
| **Transportation** | Which categories of motor vehicle driving licenses are applicable for individuals with monocular visual impairment, provided their corrected vision or uncorrected vision in the better eye reaches 5.0 or above on the standard visual acuity chart, and their horizontal field of vision reaches 150 degrees? |
| **Original Model Response** | <think> Okay, so I need to figure out which categories of motor vehicle driving licenses are applicable for individuals with monocular visual impairment, given their corrected or uncorrected vision in the better eye is 5.0 or above and their horizontal field of vision is at least 150 degrees.<br>First, I should understand what monocular visual impairment means. Monocular means only one eye is impaired, so the other eye is typically normal or has better vision. The user's better eye has a vision of 5.0 or above, which is 20/20 vision in the Snellen chart, so that's good. Also, the horizontal field of vision is 150 degrees, which is pretty wide, almost normal.<br>Next, I should recall the different categories of driving licenses. In many jurisdictions, licenses are categorized based on the type of vehicles you can drive, such as cars, motorcycles, buses, etc. But in some places, especially in China, the categories are labeled from A1 to D. I think A1 is for large passenger buses, A2 for heavy trucks, A3 for medium-sized buses, B1 for medium trucks, B2 for regular trucks, and C1 for small cars, C2 for small automatic cars, D for low-speed vehicles.<br>Now, considering visual requirements. Typically, for a regular car license, the vision requirement is at least 5.0 in the better eye, which the user meets. The field of vision is also important for safe driving, especially for maneuvering and avoiding obstacles. 150 degrees is quite wide, so that should be sufficient.<br>I should check if monocular vision is allowed. In many places, as long as the better eye meets the criteria, monocular vision is acceptable. Some jurisdictions might have stricter rules, but in general, if the user's vision is good and the field is adequate, they should qualify for standard licenses.<br>So, the user can apply for C1, C2, or D categories. C1 is for small cars, C2 for automatic cars, and D for low-speed vehicles. These are the most common categories for individuals without special vehicle needs.<br>I should also consider if there are any restrictions or additional requirements. Sometimes, they might need a medical certificate or special testing, but the user already meets the vision criteria, so that shouldn't be an issue.<br>In conclusion, the applicable categories are C1, C2, and D. These are suitable for individuals with monocular impairment who meet the specified vision standards. </think><br>The applicable categories of motor vehicle driving licenses for individuals with monocular visual impairment, who meet the specified vision standards, are:<br>- **C1**: Applicable for small cars. - **C2**: Applicable for small automatic cars. - **D**: Applicable for low-speed vehicles.<br>These categories are suitable given the individual's corrected or uncorrected vision of 5.0 or above in the better eye and a horizontal field of vision of 150 degrees. |
| **Fine-Tuned Model Response** | <think> - **Transformation of policy information into executable user guidance**: Individuals with monocular visual impairment whose better eye meets or exceeds 5.0 on the standard visual acuity chart and whose horizontal visual field reaches 150 degrees are eligible to apply for driving licenses for the following vehicle categories: small passenger cars, small automatic transmission cars, low-speed freight vehicles, three-wheeled vehicles, and motor vehicles adapted for individuals with disabilities (small automatic passenger vehicles). - **Conclusion**: Individuals with monocular visual impairment whose corrected or uncorrected vision in the better eye reaches 5.0 or above and who possess a horizontal visual field of 150 degrees or more may apply for licenses for small passenger vehicles, small automatic transmission vehicles, low-speed freight vehicles, three-wheeled vehicles, and specially adapted small automatic passenger vehicles for persons with disabilities. </think><br>According to the relevant regulations, individuals with monocular visual impairment who have corrected or uncorrected vision in the better eye reaching 5.0 or above, and whose horizontal visual field reaches 150 degrees, are eligible to apply for driving licenses for the following vehicle categories: small passenger vehicles, small automatic transmission vehicles, low-speed freight vehicles, three-wheeled vehicles, and small automatic passenger vehicles specifically adapted for individuals with disabilities. |

## Appendix C. Human-like Evaluation Details

To ensure fairness and reproducibility in human-like evaluations, we used OpenAI's GPT-4-turbo model (API version as of March 2025) to simulate structured annotation. The model was provided with detailed instructions and scoring rubrics to evaluate system responses in the following dimensions: ** consciseness **, ** correctness **, ** maliciousness **, and ** harmfulness **.

### *Appendix C.1. Evaluation Criteria*

The evaluation dimensions were defined as follows:

- **Conciseness**: Measured the brevity and clarity of a response while preserving its informativeness.
- **Correctness**: Assessed the factual alignment between the response and the ground-truth reference or supporting documents.
- **Maliciousness**: Detected whether the response contained hostile or offensive content.
- **Harmfulness**: Evaluated the potential of the response to mislead users or cause administrative misinterpretation.

### *Appendix C.2. Scoring Instructions*

The model was prompted to rate each response on a scale from 0 to 1, with a brief justification of the score to encourage self-consistency. The scoring rubric was embedded in the prompt to maintain evaluation consistency across all samples and model runs.

### *Appendix C.3. Evaluation Scope*

A total of **500 test samples** were selected from five policy subdomains: public security, healthcare, education, social security, and transportation. Each model (LLaMA3, Qwen, and ARGUS) was evaluated on the same set of questions.

### *Appendix C.4. Reliability Measures*

To assess the consistency of automatic evaluation, we conducted calibration runs using GPT-4-turbo on a subset of 100 responses. The results showed low variance (standard deviation < 0.1) in repeated scores, suggesting high stability of the automated evaluator. While no human annotators were involved, the model exhibited behavior comparable to reliable inter-rater agreement.

## References


1. Ion, R.; Avram, A.M.; Păiş, V.; Mitrofan, M.; Mititelu, V.B.; Irimia, E.; Badea, V. An open-domain QA system for e-governance. *arXiv* **2022**, arXiv:2206.08046. [CrossRef]
2. Pham, L.; Limbu, Y.B.; Le, M.T.T.; Nguyen, N.L. E-government service quality, perceived value, satisfaction, and loyalty: Evidence from a newly emerging country. *J. Public Policy* **2023**, *43*, 812–833. [CrossRef]
3. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 494–514. [CrossRef]
4. Sundberg, L.; Holmström, J. Fusing domain knowledge with machine learning: A public sector perspective. *J. Strateg. Inf. Syst.* **2024**, *33*, 101848. [CrossRef]
5. Wang, Y.; Hu, M.; Ta, N.; Sun, H.; Guo, Y.; Zhou, W.; Guo, Y.; Zhang, W.; Feng, J. Large language models and their application in government affairs. *J. Tsinghua Univ. Sci. Technol.* **2024**, *64*, 649–658. [CrossRef]
6. Yang, S.; Chen, F.; Yang, Y.; Zhu, Z. A Study on Semantic Understanding of Large Language Models from the Perspective of Ambiguity Resolution. In Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence, Shanghai, China, 21–23 July 2023; pp. 165–170. [CrossRef]
7. Tang, X.; Zheng, Z.; Li, J.; Meng, F.; Zhu, S.C.; Liang, Y.; Zhang, M. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv* **2023**, arXiv:2305.14825. [CrossRef]

8. Matarazzo, A.; Torlone, R. A Survey on Large Language Models with some Insights on their Capabilities and Limitations. *arXiv* **2025**, arXiv:2501.04040. [CrossRef]
9. Gao, S.; Gao, L.; Li, Q.; Xu, J. Application of large language model in intelligent Q&A of digital government. In Proceedings of the 2023 2nd International Conference on Networks, Communications and Information Technology, Qinghai, China, 16–18 June 2023; pp. 24–27. [CrossRef]
10. Lin, H. Designing Domain-Specific Large Language Models: The Critical Role of Fine-Tuning in Public Opinion Simulation. *arXiv* **2024**, arXiv:2409.19308. [CrossRef]
11. Lu, W.; Luu, R.K.; Buehler, M.J. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Comput. Mater.* **2025**, *11*, 84. [CrossRef]
12. Wang, X.; Tan, Y.; Yang, T.; Yuan, M.; Wang, S.; Chen, M.; Ren, F.; Zhang, Z.; Shao, Y. Efficient Large Language Model Application Development: A Case Study of Knowledge Base, API, and Deep Web Search Integration. *J. Comput. Commun.* **2024**, *12*, 171–200. [CrossRef]
13. Chen, Z.; Li, Y.; Wang, K. Optimizing reasoning abilities in large language models: A step-by-step approach. *TechRxiv* **2024**. [CrossRef]
14. Vaillancourt, E.; Thompson, C. Instruction tuning on large language models to improve reasoning performance. *TechRxiv* **2024**. [CrossRef]
15. Fang, F.; Bai, Y.; Ni, S.; Yang, M.; Chen, X.; Xu, R. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv* **2024**, arXiv:2405.20978. [CrossRef]
16. Sun, C.; Lin, K.; Wang, S.; Wu, H.; Fu, C.; Wang, Z. LalaEval: A Holistic Human Evaluation Framework for Domain-Specific Large Language Models. *arXiv* **2024**, arXiv:2408.13338. [CrossRef]
17. Pasquarelli, L.; Koutcheme, C.; Hellas, A. Comparing the Utility, Preference, and Performance of Course Material Search Functionality and Retrieval-Augmented Generation Large Language Model (RAG-LLM) AI Chatbots in Information-Seeking Tasks. *arXiv* **2024**, arXiv:2410.13326.
18. Sung, C.W.; Lee, Y.K.; Tsai, Y.T. A new pipeline for generating instruction dataset via RAG and self fine-tuning. In Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), Osaka, Japan, 2–4 July 2024; pp. 2308–2312. [CrossRef]
19. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (accessed on 15 May 2025).
20. Trajanoska, M.; Stojanov, R.; Trajanov, D. Enhancing knowledge graph construction using large language models. *arXiv* **2023**, arXiv:2305.04676. [CrossRef]
21. Es, S.; James, J.; Anke, L.E.; Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julian's, Malta, 17–22 March 2024; pp. 150–158. Available online: https://aclanthology.org/2024.eacl-demo.16/ (accessed on 13 June 2025).
22. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
23. Rajabi, E.; Midha, R.; de Souza, J.F. Constructing a knowledge graph for open government data: The case of Nova Scotia disease datasets. *J. Biomed. Semant.* **2023**, *14*, 4. [CrossRef]
24. Pellegrino, M.A.; Rula, A.; Tuozzo, G. KGHeartBeat: An Open Source Tool for Periodically Evaluating the Quality of Knowledge Graphs. In Proceedings of the International Semantic Web Conference. Springer, Hanover, MD, USA, 11–15 November 2024; pp. 40–58. [CrossRef]
25. Dong, K.; Sun, A.; Kim, J.J.; Li, X. Syntactic multi-view learning for open information extraction. *arXiv* **2022**, arXiv:2212.02068. [CrossRef]
26. Su, P.; Vijay-Shanker, K. Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction. *BMC Bioinform.* **2022**, *23*, 120. [CrossRef] [PubMed]
27. Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Found. Trends® Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
28. Alhabashneh, O.; Iqbal, R.; Doctor, F.; James, A. Fuzzy rule based profiling approach for enterprise information seeking and retrieval. *Inf. Sci.* **2017**, *394*, 18–37. [CrossRef]
29. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948. [CrossRef]
30. Gharagozlou, H.; Mohammadzadeh, J.; Bastanfard, A.; Ghidary, S.S. Semantic relation extraction: A review of approaches, datasets, and evaluation methods with looking at the methods and datasets in the Persian language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–29. [CrossRef]

31. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
32. Hang, C.N.; Yu, P.D.; Tan, C.W. TrumorGPT: Graph-Based Retrieval-Augmented Large Language Model for Fact-Checking. *IEEE Trans. Artif. Intell.* **2025**. [CrossRef]
33. Cao, H. Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark. *arXiv* **2024**, arXiv:2406.01607. [CrossRef]
34. Pelofske, E.; Liebrock, L.M.; Urias, V. Cybersecurity threat hunting and vulnerability analysis using a Neo4j graph database of open source intelligence. *arXiv* **2023**, arXiv:2301.12013. [CrossRef]
35. Wang, B.; Deng, X.; Sun, H. Iteratively prompt pre-trained language models for chain of thought. *arXiv* **2022**, arXiv:2203.08383. [CrossRef]
36. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
37. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen Technical Report. *arXiv* **2023**, arXiv:2309.16609. [CrossRef]
38. Wang, B.; Zhao, H.; Zhou, H.; Song, L.; Xu, M.; Cheng, W.; Zeng, X.; Zhang, Y.; Huo, Y.; Wang, Z.; et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv* **2025**, arXiv:2502.12671. [CrossRef]
39. GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv* **2024**, arXiv:2406.12793. [CrossRef]
40. Ibtasham, M.S.; Bashir, S.; Abbas, M.; Haider, Z.; Saadatmand, M.; Cicchetti, A. ReqRAG: Enhancing Software Release Management through Retrieval-Augmented LLMs: An Industrial Study. In *Proceedings of the International Working Conference on Requirements Engineering: Foundation for Software Quality*; Springer: Cham, Switzerland, 2025; pp. 277–292.
41. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81. [CrossRef]
42. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084. [CrossRef]
43. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A neural framework for MT evaluation. *arXiv* **2020**, arXiv:2009.09025. [CrossRef]
44. Sarmah, B.; Li, M.; Lyu, J.; Frank, S.; Castellanos, N.; Pasquali, S.; Mehta, D. How to Choose a Threshold for an Evaluation Metric for Large Language Models. *arXiv* **2024**, arXiv:2412.12148. [CrossRef]