

RG-CoT: A Rule-Guided Chain-of-Thought Evaluation Framework for Evaluating and Aligning LLM Reasoning

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) exhibit remarkable generative capabilities, evaluating their internal reasoning processes remains significantly more challenging than assessing superficial outputs. Prevailing automated metrics predominantly focus on surface-level textual quality, failing to capture critical dimensions such as logical coherence and interpretability. To bridge this gap, we propose RG-CoT, a novel evaluation framework that employs rule-guided distillation to extract interpretable Chain-of-Thought (CoT) rationales from target models. RG-CoT establishes a multi-dimensional verification system, assessing correctness, logical consistency, reliability, format compliance, and self reflection. This enables efficient model evaluation and alignment through parameter-efficient fine-tuning on high-quality distilled data. We conduct extensive experiments on the preprocessed COIG-PC dataset using a suite of state-of-the-art models. Our study spans 10 task categories, involving over 180,000 distilled CoT samples and 5,500 test queries. Results demonstrate that RG-CoT effectively evaluates high-quality model data; notably, fine-tuning on merely 5.6% of the curated high-quality data significantly outperforms training on the entire 180k corpus. Furthermore, out-of-distribution tests on the BANK-AUDIT benchmark confirm the robust generalizability of our framework.

1 Introduction

In recent years, the rapid advancement of Large Language Models (LLMs) has continuously driven technological breakthroughs in generative tasks. How to scientifically and comprehensively evaluate the practical capabilities of LLMs in such tasks has attracted increasing attention from both academia and industry (Sébastien Bubeck and Zhang, 2023; Morris et al., 2024). Although human evaluation has long been regarded as the gold standard, its

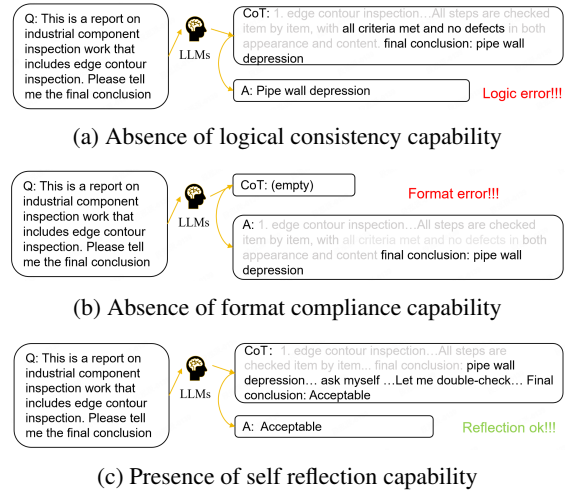


Figure 1: Examples of LLMs reasoning. This figure illustrates how different dimensions of CoT affect the correctness of LLMs reasoning through three examples in the quality inspection report.

high cost and time-consuming nature have limited large-scale applications, thus making the exploration of automated evaluation methods a research priority (Sai et al., 2023; Gehrmann et al., 2023; Zhou et al., 2023).

Early related studies have achieved remarkable progress: GPTScore (Fu et al., 2024) leverages the emergent abilities of generative pre-trained models (e.g., zero-shot instructions) to score generated texts; Chiang and Lee (2023) adopted a 1–5 Likert scale to evaluate story generation and adversarial attack scenarios via InstructGPT (Long Ouyang and Lowe, 2022) and ChatGPT, verifying the consistency between LLMs evaluation results and those of expert human raters; Kim et al. (2024) carefully constructed high-quality evaluation datasets and then fine-tuned open-source foundation LLMs through specific methods, aiming to achieve human-level evaluation performance; InteractEval (Chu et al., 2025) innovatively integrates attributes generated by humans and LLMs, generating evaluation questions and outputting final scores

with the help of the think-aloud method. Based on existing studies, relevant evaluation methods can be explicitly categorized into four types according to the core application modes of LLMs in evaluation (Gao et al., 2025): LLMs-derived metrics (ES et al., 2024; Murugadoss et al., 2025), prompting LLMs (Shu et al., 2024; Kotonya et al., 2023), fine-tuning LLMs (Hu et al., 2024; Li et al., 2024), and human-machine collaborative (Rastogi et al., 2023; Shankar et al., 2024) evaluation. Although existing methods have broken through the limitations of traditional evaluation to a certain extent, they still suffer from several critical drawbacks when confronted with the complex generative mechanisms and deep reasoning requirements of LLMs.

First, the utility of traditional evaluation metrics is often constrained by uninterpretable evaluation results (Zhang et al., 2004) and an inability to adapt to specific evaluation criteria (Wiseman et al., 2017), making it difficult to capture the reasoning logic and semantic relevance underlying texts generated by LLMs. Second, most existing evaluation methods prioritize the final outcome over the underlying reasoning process. This misalignment often leads to Goodhart’s Law effects (Manheim and Garrabrant, 2018), specifically metric hacking, where models are optimized to maximize scores through superficial heuristics rather than through robust reasoning. (Wang et al., 2025; Mizrahi et al., 2024). Chain-of-Thought (CoT) prompting is one of the key techniques that enable large language models to achieve outstanding performance in numerous generative tasks (Wei et al., 2022); it guides language models to generate reasoning grounds in a step-by-step manner and then derive final predictions (Kojima et al., 2022). The quality of CoT directly reflects the logical consistency, factual accuracy and reasoning depth of the model as illustrated in Figure 1. However, existing evaluation methods fail to fully leverage CoT as an interpretable carrier, making it hard to establish a connection between generated results and the model’s intrinsic capabilities. This leads to a lack of interpretability in evaluation and thus prevents the provision of precise guidance for model optimization.

To address the aforementioned limitations, this study proposes a CoT-based evaluation framework for LLMs capabilities, aiming to break through the constraints of existing evaluation methods. The specific scheme is as follows: First, leveraging LLMs distillation techniques, we initially construct a raw CoT dataset (for LLMs capability assess-

ment), which addresses the challenge of large-scale supply of evaluation data. Second, for five core evaluation dimensions including factual accuracy, logical consistency, and self-reflective ability, we design customized evaluation templates tailored to the characteristics of each dimension. These templates are then used to systematically filter the raw dataset, enabling precise identification and extraction of high-quality CoT samples. Finally, we take the filtered high-quality CoT samples as training data to conduct targeted fine-tuning on the foundation model. By quantitatively analyzing the correlation between evaluation results and the improvement of task performance after model fine-tuning, we establish a stable mapping relationship between CoT quality and the intrinsic capabilities of LLMs.

This framework is designed to overcome the limitation that traditional evaluation methods only focus on surface-level results, and compensate for the inadequacy of existing LLMs-based evaluation methods in measuring reasoning processes. It provides an interpretable, low-cost, and efficient solution for LLMs capability assessment, thereby facilitating the construction of trustworthy artificial intelligence systems.

Our main contributions are as follows:

- We propose a novel rule-guided Chain-of-Thought evaluation framework (RG-CoT), and use this framework to collect human-annotated data and construct synthetic datasets.
- We provide fine-grained and interpretable evaluation methods for vertical domains such as government affairs and banking, which have special requirements for logical rigor and industry regulations, addressing the lack of such tools in these fields.
- With minimal human involvement, we leverage prompt engineering and fine-tuning methods to achieve a more automated approach to model evaluation.

To facilitate further research, all data and code are available at <https://anonymous.4open.science/r/eval-78C1dfywnfcucc/>.

2 Related Work

The rapid development of Large Language Models (LLMs) has promoted the wide application of Chain-of-Thought (CoT) technology in complex tasks such as reasoning, code generation, and math-

emational problem-solving (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022; Törnberg, 2023; Ostyakova et al., 2023; Zhuang et al., 2023). As the core reasoning carrier for LLMs to solve complex problems, the quality of CoT directly maps the intrinsic capabilities of the model, reflecting whether it can decompose complex tasks, follow logical rules, or correct intermediate errors. Therefore, accurately evaluating the core capabilities of LLMs by analyzing the CoT reasoning process has become one of the key directions in current research.

However, existing evaluation systems still struggle to achieve this goal. On the one hand, traditional automatic evaluation metrics (e.g., BLEU (Papineni et al., 2002), BARTScore (Yuan et al., 2021)) ignore the model capability signals carried by the output process. Even if two models produce the same answer, one relying on rigorous step-by-step reasoning and the other on random guessing or memory, traditional metrics cannot distinguish the difference in their capabilities. On the other hand, although there have been relevant explorations on LLM evaluation in recent years (Gilardi et al., 2023; Törnberg, 2023; Ostyakova et al., 2023), none of them focus on the core goal of inferring model capabilities through the CoT process: Mendonça et al. (2023) rely on LLMs zero-shot prompting to achieve score prediction for dialogue responses, but fail to capture and analyze the model’s reasoning process. He et al. (2024) harness the powerful capabilities of GPT-4 and optimize prompt design through the Socratic method to achieve reference-free evaluation of reasoning chain quality. Gong and Mao (2023) propose a dimension chain method that introduces multi-dimensional scoring, but this method focuses on the attributes of the answer itself rather than the dimensions of the CoT in the reasoning process. Golovneva et al. (2023) develop the ROSCOE framework that performs fine-grained decomposition of step-by-step reasoning, but this framework targets the evaluation of the reliability of results rather than the inverse mapping of model capability strength.

Compared with previous studies, this paper proposes an LLM capability evaluation system based on CoT inference process analysis and applies it to more rigorous vertical domains.

3 Problem Formalization

This study aims to address the following core problem: in vertical domains \mathcal{D} lacking large-scale and high-quality human-annotated datasets \mathcal{A} , how to construct a trustworthy and automated evaluation framework \mathcal{F} . Specifically, we first distill the capabilities of the generative model \mathcal{M}_G to generate training data \mathcal{D}_{CoT} , and then achieve efficient selection and construction of high-quality CoT data $\mathcal{D}_{\text{CoT}}^*$ through framework \mathcal{F} evaluation. This high-quality CoT data is then used for model fine-tuning. Ultimately, we verify the effectiveness of the framework \mathcal{F} by measuring the performance improvement of the base model \mathcal{M}_{Θ_0} within the domain \mathcal{D} .

Definition 1. Core Elements We define the domain knowledge as $\mathcal{K} = \mathcal{K}_{\text{exp}} \cup \mathcal{K}_{\text{imp}}$, where \mathcal{K}_{exp} and \mathcal{K}_{imp} denote explicit rules and implicit heuristics, respectively. Let \mathcal{I} , \mathcal{O} , and \mathcal{T} represent the input, answer, and Chain of Thought spaces. Each instance is a triplet $\langle i, o, t \rangle \in \mathcal{I} \times \mathcal{O} \times \mathcal{T}$.

Definition 2. Framework Formulation The RG-CoT framework \mathcal{F} is formalized as a tuple $\mathcal{F} = \langle \mathcal{R}, \mathcal{V} \rangle$:

- **Rule Transformer \mathcal{R}** : Formalizes \mathcal{K} into rule-guided instructions r encapsulating reasoning steps, $\mathcal{R} : \mathcal{K} \rightarrow r$.
- **Model Behavior Verifier \mathcal{V}** : Performs multi-dimensional assessment of $\langle i, o, t \rangle$ under r to compute a score $S \in [0, 10]$:

$$\mathcal{V} : (\mathcal{I} \times \mathcal{O} \times \mathcal{T} \mid r) \rightarrow S \quad (1)$$

Definition 3. Verification Objective The objective is to curate a high-quality dataset $\mathcal{D}_{\text{CoT}}^*$ via \mathcal{V} based on a threshold $S_{\text{threshold}}$:

$$\mathcal{D}_{\text{CoT}}^* = \left\{ \begin{array}{l} \langle i, o, t \rangle \mid \langle i, o, t \rangle \in \mathcal{D}_{\text{CoT}} \\ \text{and } \mathcal{V}(\langle i, o, t \rangle) \geq S_{\text{threshold}} \end{array} \right\} \quad (2)$$

This dataset is used for supervised fine-tuning to align \mathcal{M}_{Θ_0} with domain requirements. The ultimate goal is to achieve $E(\mathcal{M}_{\Theta^*}) > E(\mathcal{M}_{\Theta_0})$, where $E(\cdot)$ is the evaluation metric.

This performance gain serves as an empirical proxy to assess the framework’s capacity to accurately identify and leverage high-quality reasoning paths that comply with domain-specific logical constraints.

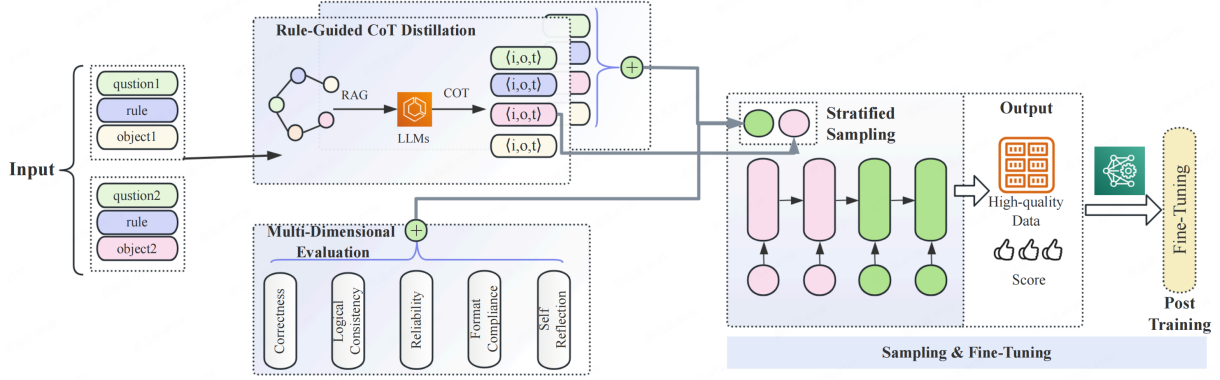


Figure 2: The Process of RG-CoT for LLM intrinsic reasoning evaluation. Rule-guided CoT distillation (Section 4.1): convert domain knowledge to guide rule-based CoT dataset zero-shot generation. Multi-dimensional evaluation (Section 4.2): verifier scores CoT samples via five weighted dimensions for quality control. Sampling & fine-tuning (Section 4.3): score-stratified sampling to build high-quality dataset for base model LoRA-SFT and domain benchmark validation.

4 The RG-CoT Framework

RG-CoT aims to evaluate the intrinsic reasoning capabilities of Large Language Models (LLMs). The core idea of this framework lies in leveraging rule-based domain knowledge \mathcal{K} to provide bidirectional guidance for the generation and evaluation of Chain-of-Thought (CoT) reasoning paths. As illustrated in Figure 2, the RG-CoT framework comprises three key modules: rule-guided CoT distillation, multi-dimensional evaluation, and sampling & fine-tuning. This section elaborates on its technical implementation in detail.

4.1 Rule-Guided CoT Distillation

To mitigate the reliance on human annotations, we leverage a generator model \mathcal{M}_G to autonomously synthesize a training dataset \mathcal{D}_{CoT} that encompasses both reasoning trajectories and final answers.

Rule Base Construction and Transformation We first employ a knowledge transformer \mathcal{R} to convert heterogeneous domain knowledge \mathcal{K} into machine-readable and highly prescriptive rules r . For explicit knowledge \mathcal{K}_{exp} , we formalize it into a structured format [Placeholder: JSON Schema]. For implicit criteria \mathcal{K}_{imp} , we integrate them into the prompts via [Placeholder: describe the transformation method]. The resulting rule base r ensures that the generated CoT remains domain-aligned.

Generation Procedure Let $P(i, r) = \text{zero-shot prompt}(i, r)$ denote the rule-guided zero-shot prompting function. We utilize an instruction set $\mathcal{I}_{\text{sample}}$ covering core domain tasks and perform zero-shot CoT prompting on \mathcal{M}_G in conjunction with rules r . Specifically, we incorporate instructions such as "<think>", "<rule>", and

"<answer>" within the prompt to explicitly activate the model's internal reasoning mechanisms. To further reinforce the guidance of r , the prompt includes specific directives requiring the model to emulate human-like step-by-step comprehension and adherence to the rules. Consequently, \mathcal{M}_G generates the raw dataset \mathcal{D}_{CoT} based on i and r :

$$\mathcal{D}_{\text{CoT}} = \{ \langle i, o, t \rangle \mid \langle o, t \rangle = \mathcal{M}_G(P(i, r)) \} \quad (3)$$

During the distillation phase, we attempt to generate the raw dataset five times. That is, one question corresponds to multiple distilled CoTs, and the size of the final dataset is increased by five times.

4.2 Multi-Dimensional Evaluation

The model behavior verifier \mathcal{V} serves as the backbone of the RG-CoT framework. It performs automated assessment on each sample in \mathcal{D}_{CoT} across five key dimensions to ensure that the distilled data is factually accurate, logically rigorous, and domain-compliant. The aggregate score S is computed as a weighted sum:

$$S = \sum_{k \in \{\text{Cor}, \text{LC}, \text{RL}, \text{FC}, \text{SR}\}} w_k \cdot \text{Eval}_k(\langle i, o, t \rangle \mid r) \quad (4)$$

where w_k represents the weight assigned to each dimension, such that $\sum w_k = 1$, see Appendix A.1 for details.

Correctness (Cor) measures the fidelity of the final output o to objective facts or domain-standard answers. We employ a binary scoring system (0/10) for closed-ended tasks and an LLM-as-a-judge mechanism for generative tasks.

Logical Consistency (LC) assesses the causal soundness and coherence of the reasoning steps

within the CoT t . We implement an atomic decomposition technique to deconstruct t into a sequence of irreducible reasoning units $\{u_1, u_2, \dots, u_N\}$. The score is derived from strong relevance computation: we evaluate not only the degree to which each atomic step is grounded in the input i and directed toward the output o , but also the inferential succession between consecutive units u_j and $u_{<j}$. Formally, the score is defined as the average relevance across all atomic steps:

$$\text{Eval}_{LC} = \frac{1}{N} \sum_{j=1}^N \text{Rel}(u_j | u_{<j}, i, o) \quad (5)$$

where $\text{Rel}(\cdot)$ denotes the relevance scoring function. This mechanism enables a rigorous examination of the logical integrity of each step, ensuring high internal consistency across the input, the output, and all intermediate reasoning nodes.

Reliability (RL) characterizes the model’s consistency in generating stable outputs under identical or perturbed inputs. This is quantified by performing N samplings via temperature variation for a given i and calculating the semantic similarity across the resulting outputs.

Format Compliance (FC) determines whether o and t strictly adhere to pre-defined structural constraints, such as JSON Schema. The scoring involves granular validation of delimiters and content validity, utilizing regular expression matching to ensure the presence of all mandatory fields.

Self Reflection (SR) evaluates whether the model demonstrates proactive error detection or corrective behavior within the CoT t . The metric detects the presence of specific linguistic cues (see Appendix B.1) and assigns incremental scores based on the effectiveness of the reflection in refining the final answer.

4.3 Sampling & Fine-Tuning

Score-based Stratified Sampling Guided by the evaluation results from the assessment framework, we first conduct quality screening on the multiple CoT outputs corresponding to each question. Specifically, we select the CoT that best reflects the model’s capability to address the target question. This selection is predicated on the comprehensive assessment outcomes, ensuring the retained data effectively characterizes the model’s response competence for each question. We then map the selected optimal CoT to its corresponding question, thereby constructing the final evaluation dataset.

Upon obtaining the aggregate scores S for all samples in this constructed evaluation dataset (derived from \mathcal{D}_{CoT}), we implement a stratified sampling strategy. We partition the data into three distinct tiers: a high-score stratum ([8, 10]), a mid-to-high stratum ([5, 8)), and a low-score stratum ([0, 5)). To ensure high confidence in the resulting training set, we enforce a global quality criterion $S \geq S_{\text{threshold}}$, yielding the finalized dataset $\mathcal{D}_{\text{CoT}}^*$.

Post-training and Validation We utilize $\mathcal{D}_{\text{CoT}}^*$ as supervision data to perform Supervised Fine-Tuning (SFT) on the base model \mathcal{M}_{Θ_0} . Specifically, we employ Low-Rank Adaptation (LoRA) to optimize the parameters Θ_0 , resulting in the fine-tuned model \mathcal{M}_{Θ^*} . To validate the efficacy of the RG-CoT framework in the absence of large-scale human annotations, we evaluate the performance $E(\cdot)$ of \mathcal{M}_{Θ^*} against the baseline \mathcal{M}_{Θ_0} on the standard benchmark of domain \mathcal{D} .

5 Experiments

5.1 Experimental Setup

Model We select Yi-1.5-34B (Young et al., 2024) as the baseline model (Bai et al., 2025) due to its superior performance in existing evaluations, and distill the evaluated model to fine-tune the Qwen2.5-7B (Yang et al., 2024) and Qwen3-8B models (Yang et al., 2025). Qwen2.5-7B has a pre-training data volume of 18 trillion tokens and supports a 128K context window. Qwen3-8B has a pre-training data volume twice that of Qwen2.5-7B and also supports a 128K context window. The evaluated models (distillation models) adopt Llama-3.1-70B-Instruct (Team, 2024), Llama-3.3-70B-Instruct (Meta AI, 2024), GPT-4o (Hurst et al., 2024), Deepseek-R1-671B (DeepSeek-AI, 2025), Qwen2.5-72B (Yang et al., 2024), Qwen2.5-32B (Yang et al., 2024), and QwQ-32B (The Qwen Team, 2025). We use the Huggingface Transformers (Wolf et al., 2019) and DeepSpeed (Rasley et al., 2020) libraries for training, and the vLLM library (Kwon et al., 2023) for model inference. For model distillation, we use temperature = 0.7, following Cobbe et al. (2021), attempts = 5.0. Involving LLMs as a judge in evaluation, we use a sampling temperature of 0.1.

Dataset (1) In-domain Evaluation and Training: COIG-PC (BAAI, 2023; Zhang et al., 2023) is a Chinese natural language processing dataset covering over 3,000 Chinese tasks. We use the preprocessed COIG-PC dataset, which includes data for five gen-

Base Model	CoT Generator	ACC. (%)	Acc. (%) (Post-SFT)	
		(Pre-SFT)	180k (Full)	10k (Sub)
Yi-1.5-34B (Baseline)	—	63.7	76.0 (\uparrow 12.3)	—
Qwen2.5-7B	DeepSeek-R1-671B	45.1	77.0	78.5 (\uparrow 33.4)
	GPT-4o		74.5	76.8
	Qwen2.5-72B		75.5	76.5
	Llama-3.3-70B-Instruct		73.8	75.5
	QwQ-32B		70.0	72.0
	Llama-3.1-70B-Instruct		69.2	71.0
	Qwen2.5-32B		68.0	70.3
Qwen3-8B	DeepSeek-R1-671B	61.0	80.0	83.2 (\uparrow 22.2)
	GPT-4o		77.4	80.8
	Llama-3.3-70B-Inst.		77.9	79.8
	Qwen2.5-72B		76.7	79.5
	QwQ-32B		75.5	76.5
	Llama-3.1-70B-Inst.		75.2	77.1
	Qwen2.5-32B		70.0	71.0

Table 1: Experimental results on the COIG-PC dataset. Both student models surpass the baseline when fine-tuned by the most capable teacher model. Pre-SFT denotes the model’s zero-shot capability, while Post-SFT indicates performance after instruction tuning. Acc. refers to accuracy.

421 eral task types (question answering, reading comprehension, summarization, information extraction, and natural language inference) and five government affairs task types (intent recognition, type classification, domain classification, transfer classification, and government QA). There are a total of 180,000 training samples. We select 5% of the data from government affairs tasks to form the test set, resulting in a total of 5,500 test samples. More details are in Appendix A.2.1.(2) Out-of-Domain Eval: To verify the generalization ability of the evaluation framework beyond government affairs, we test on bank domain tasks. The dataset is from desensitized BANK-AUDIT see Appendix A.2.2, including five violation inspections: bank demonstration activity violation inspection(Activ-V), bank announcement compliance inspection(Annc-C), bank report violation inspection(Report-V), bank illegal assessment inspection(Assess-I), and bank business application violation inspection(Bus-V), with 60,000 instances.

442 5.2 Experimental Results

443 **Evaluation on Government Affairs and General NLP** As shown in Table 1, we utilize post-SFT accuracy to evaluate model capability and data quality. First, regarding Teacher Assessment, the student models’ performance strictly correlates with the strength of the generator, identifying DeepSeek-R1-671B as the superior reasoner, achieving a peak of 83.2%. Second, regarding Student Potential, the framework reveals that smaller models like

452 Qwen2.5-7B (78.5%) can outperform larger base- 453 lines like Yi-1.5-34B (76.0%) when steered by 454 high-quality reasoning paths. Finally, regarding 455 Data Quality, the consistent superiority of the 10k 456 curated subset over the 180k fullset (e.g., +3.2% 457 gain for Qwen3-8B) validates our RG-CoT frame- 458 work’s ability to filter noise and select the most 459 representative data for efficient alignment.

460 **Robustness in Financial Domain Audit** The out- 461 of-domain evaluation results, summarized in Table 462 2, demonstrate that While the Qwen3-8B shows 463 limited performance (52.40%), all teacher mod- 464 els demonstrate strong zero-shot capabilities (81%- 465 86%), with DeepSeek-R1-671B leading at 86.38%. 466 Crucially, fine-tuning with subset curated by RG- 467 CoT consistently outperforms the fullset approach 468 across all teacher sources, validating the frame- 469 work’s ability to select data that best represents the 470 model’s capabilities. Furthermore, data distilled 471 from the strongest teacher, Deepseek-R1-671B, en- 472 ables Qwen3-8B to achieve the highest accuracy 473 of 79.02%, demonstrating that this fine-tuning per- 474 formance serves as an effective metric to evaluate 475 which model is stronger.

476 5.3 Analysis

477 **CoT Quality as a Proxy for Capability** Experi- 478 mental results demonstrate that the key to evalu- 479 ating model capabilities lies in analyzing the gen- 480 eration quality of their CoT reasoning paths. We 481 observe that relying solely on the correctness of 482 final answers fails to fully distinguish the actual

Training Method	Activ-V	Annc-C	Report-V	Assess-I	Bus-V	Avg.
<i>Qwen3-8B (Baseline)</i>						
Pre-SFT baseline	54.20	61.50	48.30	50.40	47.60	52.40
<i>DeepSeek-R1-671B</i>						
Base accuracy	91.40	88.60	82.30	85.50	84.10	86.38
Fullset	76.50	79.20	72.40	74.80	73.50	75.28
Subset	80.20 (↑26.0)	82.40 (↑20.9)	76.80 (↑28.5)	78.50 (↑28.1)	77.20 (↑29.6)	79.02
<i>GPT-4o</i>						
Base accuracy	89.20	87.40	79.50	82.60	81.40	84.02
Fullset	74.30	77.10	70.80	72.40	71.20	73.16
Subset	77.80 (↑23.6)	79.50 (↑18.0)	74.20 (↑25.9)	75.90 (↑25.5)	74.80 (↑27.2)	76.44
<i>Llama-3.3-70B-Instruct</i>						
Base accuracy	86.50	84.20	80.80	81.40	82.50	83.08
Fullset	71.80	74.50	71.20	71.80	72.40	72.34
Subset	75.10 (↑20.9)	77.20 (↑15.7)	75.40 (↑27.1)	74.60 (↑24.2)	75.80 (↑28.2)	75.62
<i>QwQ-32B</i>						
Base accuracy	87.80	85.10	76.40	79.20	78.50	81.40
Fullset	73.40	75.80	68.50	70.10	69.80	71.52
Subset	76.20 (↑22.0)	78.40 (↑16.9)	72.60 (↑24.3)	73.50 (↑23.1)	73.20 (↑25.6)	74.78

Table 2: Experimental results on the BANK-AUDIT task. The Pre-SFT baseline denotes the student model’s initial capability, while base accuracy serves as the upper bound derived from the teacher model. The subset model achieves substantial gains over the baseline (marked in green) and significantly approaches the upper bound of base accuracy. Avg. refers to average.

performance of models. For instance, with identical accuracy rates, models with high scores in *LC* and *SR* exhibit stronger robustness when handling unseen complex tasks. This finding validates the core hypothesis of this paper: high-quality CoT reasoning processes serve as explicit manifestations of a model’s intrinsic capabilities. The high-score samples selected by the RG-CoT framework essentially represent the upper bound of a model’s ability to adhere to logical rules and domain knowledge; thus, fine-tuning with these samples can induce the foundation model to achieve significant improvements in capability as illustrated in Figure 3.

Supervised Weight Optimization of Evaluation Operators

In this framework, the calculation of the comprehensive score S is not fixed; instead, the weights w_k assigned to the five evaluation dimensions (Cor, LC, RL, FC, SR) are task-adaptive and domain-specific. For instance, in government affairs domain tasks, higher weights are allocated to Cor, FC and LC, as these tasks prioritize strict adherence to standardized document formats and logical rigor of policy interpretation processes. By contrast, in banking security review tasks, Cor, RL and SR serve as the core operators for filtering high-quality samples, since such scenarios demand precise alignment with regulatory rules and trustworthy judgment of risk-related information. This

adaptive weighting mechanism ensures that RG-CoT can capture the differentiated requirements of core model capabilities between government affairs and banking security domains, thereby demonstrating strong generalization performance in cross-domain experiments as illustrated in Figure 4a.

Impact of Data Composition on Performance Improvement

This section investigates the regulatory effect of data composition on the final performance of models. We conduct comparative experiments between the fullset of 180,000 samples and the filtered high-quality subset of 10,000 samples. The results reveal that larger dataset sizes do not necessarily yield better performance; on the contrary, incorporating a large proportion of low-quality or logically redundant CoT data introduces noise and limits the performance gains of foundation models. As shown in Table 1, fine-tuning with only 5.6% of high-score samples achieves a significantly higher accuracy (e.g., 83.2% for Qwen3-8B), outperforming the fullset with potential erroneous reasoning paths (80.0%). This finding indicates that in resource-constrained vertical domains, prioritizing the purity of data quality over pursuing the scale of data quantity plays a more critical role in aligning model capabilities as illustrated in Figure 4b.

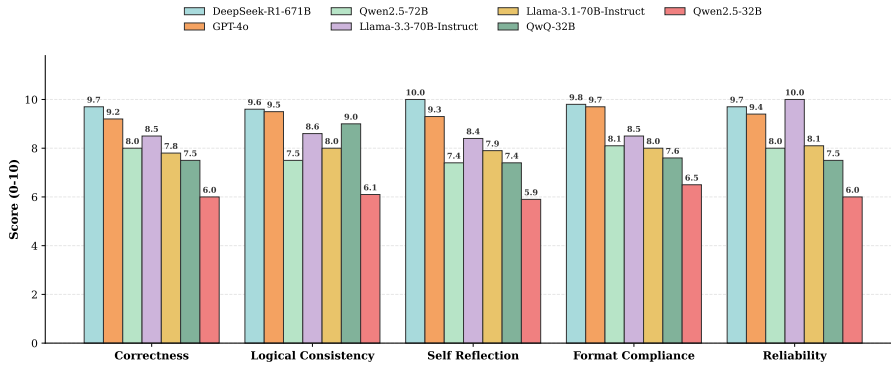
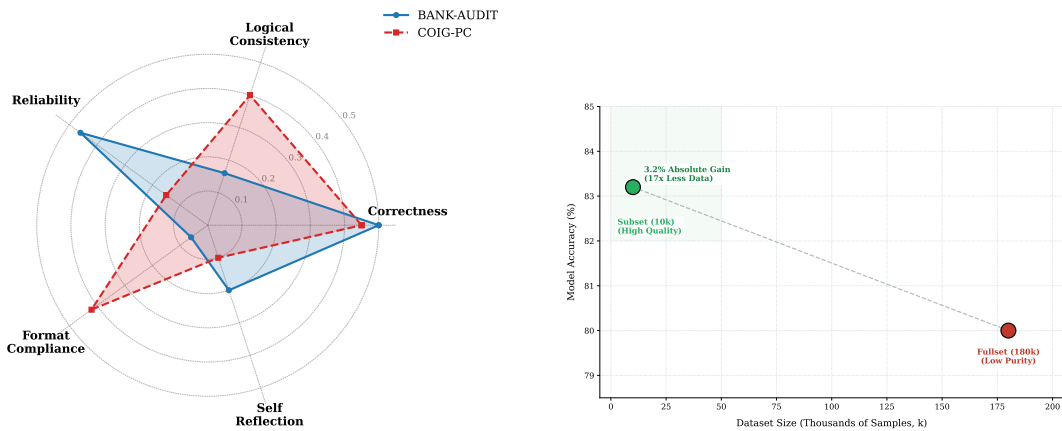


Figure 3: Performance comparison of seven models across five CoT evaluation dimensions in the government affairs vertical. DeepSeek-R1-671B and GPT-4o demonstrate dominant performance overall, particularly in self reflection and correctness. Notably, Llama-3.3-70B-Instruct achieves a perfect score (10.0) in Reliability, outperforming larger models, while QwQ-32B shows strong competitiveness in logical consistency.



(a) Task-specific weight distribution: the BANK-AUDIT task (blue) focuses on RL, while COIG-PC (red) targets FC and LC.

(b) Comparison of data strategies. The subset (green) faithfully captures the teacher model’s capability, whereas the fullset (red) suffers from noise.

Figure 4: Weight allocation analysis of RG-CoT, and efficiency analysis of distilled models in government affairs tasks

6 Conclusion

This paper introduced RG-CoT, an evaluation framework designed to address the lack of process-oriented support in current Large Language Model assessments. Unlike traditional metrics that focus exclusively on final outputs, our research utilized CoT as an interpretable vehicle that reflected logical depth and adhered to domain-specific rules. By constructing a rule-guided distillation pipeline and a multi-dimensional verification mechanism, we demonstrated that the quality of the reasoning process reliably maps a model’s intrinsic capabilities. Experimental results confirmed that fine-tuning models on high-scoring CoT samples curated by this framework yielded performance improvements that surpassed baseline methods. Future research will explore the extensibility of RG-CoT to broader multi-modal reasoning tasks and further optimize

automated judgment models to mitigate potential biases in the evaluation process.

Limitations

The current implementation of RG-CoT prioritizes rigorous evaluation standards by leveraging explicit, structured domain knowledge. While this design choice guarantees high precision for high-stakes vertical domains (such as banking audits), it entails a necessary trade-off regarding the manual formalism required for tasks with implicit or ambiguous reasoning standards. Furthermore, our multi-dimensional verification process introduces an upfront computational investment during the data preparation phase due to the generation of multiple CoT paths. We view this as a strategic cost to facilitate parameter-efficient deployment, as the resulting high-quality distilled data enables smaller models to achieve performance compara-

574 ble to larger baselines, thereby reducing the long-
 575 term inference overhead in practical applications.
 576 Regarding the evaluation metrics, the 'Self Reflec-
 577 tion' dimension utilizes specific linguistic cues as
 578 proxies for introspective reasoning. While we ac-
 579 knowledge the theoretical possibility of surface-
 580 level mimicry, the consistent performance gains
 581 observed in downstream tasks suggest that these
 582 markers successfully trigger necessary computa-
 583 tion to refine outputs. Finally, as a distillation-
 584 based framework, the performance boundary of the
 585 student model is naturally aligned with the capa-
 586 bilities of the teacher model. RG-CoT focuses on
 587 enhancing the consistency of reasoning alignment
 588 and is designed to be modular, allowing it to scale
 589 with and distill the capabilities of increasingly pow-
 590 erful foundation models.

591 Ethical considerations

592 AI-assisted language refinement tools were used
 593 to improve writing clarity, academic expression,
 594 and translation during manuscript preparation, with
 595 all AI contributions reviewed and verified by the
 596 authors to maintain scientific integrity and trans-
 597 parency.

598 Acknowledgments

599 References

600 BAAI. 2023. COIG-PC: Chinese open instruction gen-
 601 eralist - prompt collection. <https://huggingface.co/datasets/BAAI/COIG-PC>.
 602

603 Yuelin Bai, Xeron Du, Yiming Liang, Leo Jin, Junt-
 604 ing Zhou, Ziqiang Liu, Feiteng Fang, Mingshan
 605 Chang, Tianyu Zheng, Xincheng Zhang, Nuo Ma,
 606 Zekun Moore Wang, Ruibin Yuan, Haihong Wu,
 607 Hongquan Lin, Wenhao Huang, Jiajun Zhang,
 608 Chenghua Lin, Jie Fu, and 3 others. 2025. **COIG-
 609 CQIA: quality is all you need for chinese instruction
 610 fine-tuning**. In *Findings of the Association for Com-
 611 putational Linguistics: NAACL 2025, Albuquerque,
 612 New Mexico, USA, April 29 - May 4, 2025*, pages
 613 8190–8205. Association for Computational Linguis-
 614 tics.

615 Cheng-Han Chiang and Hung-yi Lee. 2023. Can large
 616 language models be an alternative to human evalua-
 617 tions? *arXiv preprint arXiv:2305.01937*.

618 SeongYeub Chu, Jong Woo Kim, and Mun Yong Yi.
 619 2025. Think together and work better: Combining
 620 humans' and LLMs' think-aloud outcomes for ef-
 621 fective text evaluation. In *Proceedings of the 2025
 622 CHI Conference on Human Factors in Computing
 623 Systems*, pages 1–23.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,
 624 Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias
 625 Plappert, Jerry Tworek, Jacob Hilton, Reiichiro
 626 Nakano, Christopher Hesse, and John Schulman.
 627 2021. **Training verifiers to solve math word prob-
 628 lems**. *arXiv preprint arXiv:2110.14168*. 629

DeepSeek-AI. 2025. **Deepseek-r1: Incentivizing rea-
 630 soning capability in llms via reinforcement learning**.
 631 *arXiv preprint arXiv:2501.12948*. 632

Shahul ES, Jithin James, Luis Espinosa Anke, and
 633 Steven Schockaert. 2024. **Ragas: Automated evalua-
 634 tion of retrieval augmented generation**. In *In Pro-
 635 ceedings of EACL 2024*, pages 150–158. Association
 636 for Computational Linguistics. 637

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei
 638 Liu. 2024. **GPTScore: Evaluate as you desire**. In
 639 *Proceedings of the 2024 Conference of the North
 640 American Chapter of the Association for Computa-
 641 tional Linguistics: Human Language Technologies
 642 (Volume 1: Long Papers)*, pages 6556–6576. 643

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun
 644 Wan. 2025. **LLM-based NLG evaluation: Current
 645 status and challenges**. *Computational Linguistics*,
 646 pages 1–27. 647

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sel-
 648 lam. 2023. **Repairing the cracked foundation: A sur-
 649 vey of obstacles in evaluation practices for generated
 650 text**. *J. Artif. Intell. Res.*, 77:103–166. 651

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli.
 652 2023. **ChatGPT outperforms crowd workers for
 653 text-annotation tasks**. *Proceedings of the National
 654 Academy of Sciences*, 120(30):e2305016120. 655

Olga Golovneva, Moya Chen, Spencer Poff, Martin
 656 Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi,
 657 and Asli Celikyilmaz. 2023. **ROSCOE: A suite of
 658 metrics for scoring step-by-step reasoning**. In *The
 659 Eleventh International Conference on Learning Rep-
 660 resentations*. 661

Peiyuan Gong and Jiaxin Mao. 2023. **CoAScore: Chain-
 662 of-aspects prompting for NLG evaluation**. *arXiv
 663 preprint arXiv:2312.10355*. 664

Hangfeng He, Hongming Zhang, and Dan Roth. 2024.
 665 **Socreval: Large language models with the socratic
 666 method for reference-free reasoning evaluation**. In
 667 *Findings of the Association for Computational Lin-
 668 guistics: NAACL 2024, Mexico City, Mexico, June
 669 16-21, 2024*, pages 2736–2764. Association for Com-
 670 putational Linguistics. 671

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xi-
 672 aojun Wan. 2024. **Themis: A reference-free NLG
 673 evaluation language model with flexibility and in-
 674 terpretability**. In *Proceedings of the 2024 Confer-
 675 ence on Empirical Methods in Natural Language
 676 Processing, EMNLP 2024, Miami, FL, USA, Novem-
 677 ber 12-16, 2024*, pages 15924–15951. Association
 678 for Computational Linguistics. 679

680	Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card . <i>arXiv preprint arXiv:2410.21276</i> .	
688	Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models . In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024</i> , pages 4334–4353. Association for Computational Linguistics.	
698	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In <i>Advances in Neural Information Processing Systems</i> , volume 35, pages 22199–22213.	
703	Neema Kotonya, Saran Krishnasamy, Joel R. Tetreault, and Alejandro Jaimes. 2023. Little giants: Exploring the potential of small llms as evaluation metrics in summarization in the eval4nlp 2023 shared task . In <i>Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2023, Bali, Indonesia, November 1, 2023</i> , pages 202–218. Association for Computational Linguistics.	
711	Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention . In <i>Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023</i> , pages 611–626. ACM.	
719	Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment . In <i>The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024</i> . OpenReview.net.	
725	Xu Jiang Diogo Almeida Carroll L. Wainwright Pamela Mishkin Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell Peter Welinder Paul F. Christiano Jan Leike Long Ouyang, Jeffrey Wu and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In <i>Advances in Neural Information Processing Systems</i> , volume 35, pages 27730–27744.	
734	David Manheim and Scott Garrabrant. 2018. Categorizing variants of goodhart’s law . <i>arXiv preprint arXiv:1803.04585</i> .	
	John Mendonça, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. In <i>Proceedings of the Eleventh Dialog System Technology Challenge</i> , pages 133–143.	737 738 739 740 741 742
	Meta AI. 2024. Llama 3.3: Supercharging the open source ecosystem. https://ai.meta.com/blog/meta-llama-3/ .	743 744 745
	Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation . <i>Transactions of the Association for Computational Linguistics</i> , 12:933–949.	746 747 748 749 750
	Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clément Farabet, and Shane Legg. 2024. Position: Levels of AGI for operationalizing progress on the path to AGI. In <i>Forty-first International Conference on Machine Learning</i> .	751 752 753 754 755 756
	Bhuvanashree Murugadoss, Christian Pölit, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2025. Evaluating the evaluator: Measuring llms’ adherence to task evaluation instructions . In <i>AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA</i> , pages 19589–19597. AAAI Press.	757 758 759 760 761 762 763 764
	Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In <i>Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue</i> , pages 242–254.	765 766 767 768 769 770 771
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics</i> , pages 311–318.	772 773 774 775 776
	Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters . In <i>KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020</i> , pages 3505–3506. ACM.	777 778 779 780 781 782 783
	Charvi Rastogi, Marco Túlio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms . In <i>Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023</i> , pages 913–926. ACM.	784 785 786 787 788 789
	Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems . <i>ACM Computing Surveys</i> , 55(2):26:1–26:39.	790 791 792 793

794	Shreya Shankar, J. D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya G. Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of LLM outputs with human preferences. In <i>Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16, 2024</i> , pages 131:1–131:14. ACM.	852
795		853
796		854
797		855
798		856
799		857
800		
801		
802	Lei Shu, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, and Lei Meng. 2024. Fusion-eval: Integrating assistant evaluators with llms. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024</i> , pages 225–238. Association for Computational Linguistics.	858
803		859
804		860
805		861
806		862
807		863
808		864
809		
810	Ronen Eldan Johannes Gehrke Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott M. Lundberg Harsha Nori Hamid Palangi Marco Túlio Ribeiro Sébastien Bubeck, Varun Chandrasekaran and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. <i>arXiv preprint arXiv:2303.12712</i> .	865
811		866
812		867
813		868
814		869
815		870
816		871
817	Llama Team. 2024. The llama 3 ferd of models. <i>arXiv preprint arXiv:2407.21783</i> .	872
818		873
819		874
820		875
821		876
822		877
823		878
824		
825		
826		
827		
828		
829		
830		
831		
832		
833		
834		
835		
836		
837		
838		
839		
840		
841		
842		
843		
844		
845		
846		
847		
848		
849		
850		
851		
	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. <i>arXiv preprint arXiv:1910.03771</i> .	
	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. <i>arXiv preprint arXiv:2505.09388</i> .	
	An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. <i>arXiv preprint arXiv:2412.15115</i> .	
	Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, and 11 others. 2024. Yi: Open foundation models by 01.ai. <i>arXiv preprint arXiv:2403.04652</i> .	
	Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In <i>Advances in Neural Information Processing Systems</i> , volume 34, pages 27263–27277.	
	Ge Zhang, Yemin Shi, Ruiibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. 2023. Chinese open instruction generalist: A preliminary release. <i>arXiv preprint arXiv:2304.07987</i> .	
	Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In <i>Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal</i> . European Language Resources Association.	
	Yongxin Zhou, Fabien Ringeval, and François Portet. 2023. A survey of evaluation methods of generated medical textual reports. In <i>Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023</i> , pages 447–459. Association for Computational Linguistics.	
	Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Zixian Feng, Weinan Zhang, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In <i>Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)</i> , pages 88–109.	

A Appendix

A.1 Weights Update Details

Regarding the preprocessed COIG-PC dataset, ten typical chain-of-thought (CoT) cases were manually annotated with scores across all dimensions to serve as the initial dataset. Evaluation results for each dimension were normalized to the range [0, 1] to ensure consistent numerical scales across different dimensions. Additionally, fifty representative Chain-of-Thought (CoT) cases were manually annotated by three domain experts, who independently scored each case across all five evaluation dimensions to constitute a seed set for evaluating annotation consistency and for the subsequent bootstrap-based verification of weight convergence. A consistency check using Fleiss’ Kappa was implemented to verify inter-annotator agreement (IAA), ensuring the reliability and statistical significance of the annotation outcomes. The BANK-AUDIT dataset adopts the identical weight processing pipeline, including the same annotation protocol and normalization method.

Experimental Setup

- **Experimental Group 1 (Gaussian Distribution Initialization):** Five weights were randomly sampled from a Gaussian distribution with a mean $\mu = 0.2$ (corresponding to an average weight of 0.2 across the five dimensions) and a standard deviation $\sigma = 0.05$. These sampled weights were then normalized to ensure their sum equals 1.
- **Experimental Group 2 (Uniform Distribution Initialization):** Five weights were randomly sampled from the interval [0.1, 0.3] and subsequently normalized to a sum of 1.
- **Experimental Group 3 (Entropy Weight Method Initialization):** Weights were calculated based on the degree of dispersion in the manually annotated data, which objectively reflects the discriminability of each dimension.

Experimental Results

- **Group 1:** Random sampling failed to correlate with the discriminability of each dimension; weight assignment was entirely determined by probability, deviating from the expected outcomes. No significant differences were observed between the weights of correctness and logical consistency versus those of

reliability and format compliance. Notably, an unintended result emerged where the weight of format compliance exceeded that of logical consistency.

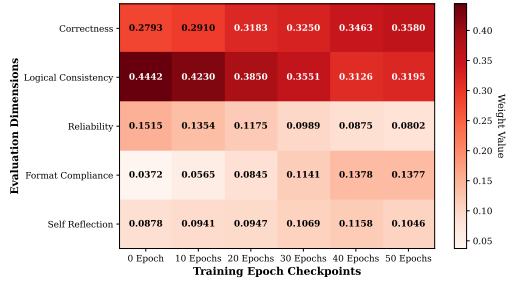
- **Group 2:** Forcing weights to approximate an average across all dimensions neglected the discriminability differences between dimensions. All five dimensions yielded weights in the range [0.18, 0.22], meaning correctness and logical consistency could not be assigned higher weights, while format compliance and other dimensions could not be assigned lower weights.
- **Group 3:** A lower entropy value corresponds to higher discriminability and thus a higher weight. The weights of correctness and logical consistency were significantly higher than those of the other three dimensions.

A.1.1 Weight Initialization via Entropy Weight Method

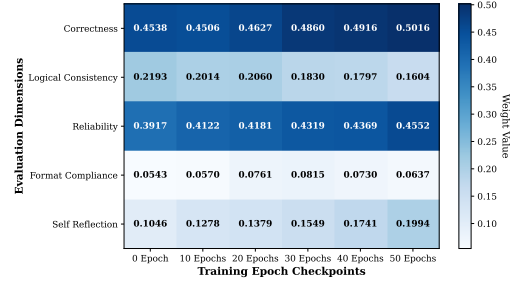
Since the quality evaluation of CoT requires an objective reflection of each dimension’s discriminative contribution to the final assessment result, classical random initialization methods (Gaussian/uniform) fail to capture the discriminability differences between dimensions, leading to weight assignments that deviate from requirements. In contrast, the entropy weight method quantifies discriminative contributions via the degree of data dispersion, and the initialized weights are more aligned with the core objective of CoT quality evaluation, making it a superior weight initialization scheme.

Therefore, the entropy weight method was ultimately adopted for weight initialization: For the standardized data of all dimensions, the information entropy of each dimension was calculated. A smaller information entropy indicates a more dispersed data distribution (i.e., higher discriminability), signifying a greater contribution to the final evaluation result and thus warranting a higher weight. Initial weights were computed based on entropy values, with the constraint that the sum of weights across all dimensions equals 1.

To objectively determine the importance of each evaluation dimension, we employ the Entropy Weight Method (EWM) for weight initialization. The core principle of EWM is that for normalized data across all dimensions, the information entropy



(a) Weight Evolution in Government Domain.



(b) Weight Evolution in Banking Domain.

Figure A.1: Heatmaps illustrating the dynamic evolution of evaluation dimension weights during the model training process for government and banking tasks. Sub-figure (a) presents the weight adjustment strategy for the government domain, while (b) displays the strategy for the banking domain. As training epochs progress, the weights of each dimension adaptively change based on specific task characteristics.

serves as a measure of dispersion. A smaller entropy value indicates greater data dispersion (higher discriminative power), suggesting that the dimension provides more useful information and should contribute more to the final evaluation (i.e., assigned a higher weight). Conversely, higher entropy implies less information and a lower weight. The weights are initialized to ensure that the sum of all dimensional weights equals 1.

Let $n = 10$ denote the number of samples and $m = 5$ denote the number of dimensions. Let x_{ij} represent the raw score of the i -th sample on the j -th dimension (where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$). Furthermore, let y_{ij} be the normalized score, p_{ij} be the probability proportion of the i -th sample under the j -th dimension, e_j be the information entropy of the j -th dimension, and w_j be the final weight assigned to the j -th dimension.

Data Normalization Since the raw scores of the 5 dimensions may differ in scale, they must first be normalized to the $[0, 1]$ interval. For positive indicators (where a higher score indicates better performance), the normalization is defined as:

$$y_{ij} = \frac{x_{ij} - \min_{i=1..n}(x_{ij})}{\max_{i=1..n}(x_{ij}) - \min_{i=1..n}(x_{ij})} \quad (\text{A.1})$$

In the specific case where all samples have the same score for a given dimension (i.e., $\max(x_{ij}) - \min(x_{ij}) = 0$), the discriminative power of that dimension is 0. Consequently, the normalized score is set to $y_{ij} = 0$, leading to a subsequent entropy of 1 and a weight of 0.

Probability Proportion Calculation We calculate the probability proportion p_{ij} . This treats the normalized scores of the j -th dimension as a probability distribution, where the share of the i -th sam-

ple represents its contribution to the total score of that dimension:

$$p_{ij} = \frac{y_{ij} + \epsilon}{\sum_{i=1}^n (y_{ij} + \epsilon)} \quad (\text{A.2})$$

Here, a small smoothing factor $\epsilon = 10^{-10}$ is introduced to prevent meaningless mathematical operations, such as ensuring $y_{ij} \neq 0$ when calculating logarithms in later steps.

Entropy Calculation The information entropy e_j for each dimension is calculated based on Shannon entropy to reflect the degree of data dispersion:

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \cdot \ln(p_{ij}) \quad (\text{A.3})$$

The term $-\frac{1}{\ln n}$ is used to normalize the entropy value to the $[0, 1]$ interval. When all p_{ij} are equal, $e_j = 1$, indicating the lowest discriminative power. Conversely, if one sample has $p_{ij} = 1$ while others are 0, then $e_j = 0$, indicating the highest discriminative power.

Weight Determination Finally, we calculate the weight w_j for each dimension. We first determine the coefficient of variation (or difference coefficient) g_j , where a smaller entropy results in a larger coefficient and a higher weight. The weights are then normalized:

$$w_j = \frac{g_j}{\sum_{j=1}^m g_j}, \quad \text{where } g_j = 1 - e_j \quad (\text{A.4})$$

This ensures that $\sum_{j=1}^m w_j = 1$, and that dimensions with smaller e_j values receive larger weights.

A.1.2 Dynamic Weight Adjustment via Gradient Optimization

Data Composition Our dataset is constructed to cover diverse scenarios. For the government affairs domain, we selected 200 identical workflow queries and collected 200 Chain-of-Thought (CoT) responses from each of the following models: Llama-3.1-70B-Instruct (Team, 2024), Llama-3.3-70B-Instruct (Meta AI, 2024), GPT-4o (Hurst et al., 2024), Deepseek-R1-671B (DeepSeek-AI, 2025), Qwen2.5-72B (Yang et al., 2024), Qwen2.5-32B (Yang et al., 2024), and QwQ-32B (The Qwen Team, 2025). Similarly, for the general domain, we selected 100 identical queries and collected 100 CoT responses from the same set of models.

Composite Scoring Let the scores for a single sample across the five dimensions be defined as x_1 (Correctness), x_2 (Logical Consistency), x_3 (Reliability), x_4 (Format Compliance), and x_5 (Self Reflection). With the corresponding weight vector $(w_1, w_2, w_3, w_4, w_5)$, the predicted composite score \hat{s} is calculated:

$$\hat{s} = \sum_{j=1}^5 x_j \cdot w_j \quad (\text{A.5})$$

Here, \hat{s} represents the model-predicted composite score, while s denotes the ground truth score derived from human annotation in the dataset. The objective is to approximate s as closely as possible using \hat{s} .

Loss Function Construction To maximize the composite score effectiveness for government workflow tasks, we construct a loss function incorporating a penalty term, which can be modified for different tasks:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n (\hat{s} - s)^2 + \lambda \left[(\max(0, w_3 - w_4))^2 + (\max(0, w_3 - w_5))^2 \right] \quad (\text{A.6})$$

In this equation, λ controls the penalty strength. The penalty term is designed to penalize scenarios where the weight of Reliability (w_3) exceeds that of Format Compliance (w_4) or Self Reflection (w_5).

Training and Optimization The format of a training sample is $(x_1, x_2, x_3, x_4, x_5, s)$, where $x_{1...5}$ are the scores for the five dimensions and

s is the quality label. We train for 50 epochs. To prevent cross-task sample mixing from diluting the quality difference signals within a single task, we treat all model outputs for the same original query as a *minimal sample unit*. Each original query corresponds to 7 samples (outputs from all evaluated teacher models).

A single batch contains 2 original queries (totaling 14 samples). In each iteration, the loss is calculated using the Mean Squared Error (MSE) and the penalty term on the batch, and weights are updated accordingly. The learning rate is set to 0.01. The weights stabilize after approximately 40 epochs. The goal of this process is to enable the model to learn the characteristics of high-quality CoT data and to allow the loss function to distinguish between optimal and suboptimal reasoning chains. Figure A.1 illustrates the changes in sub-weights before and after optimization via a heatmap.

Task (Abbr.)	Fixed	Dynamic	Gain
Gov. Intent Recog.	76.8	78.2	+1.4
Gov. Domain Class.	75.3	78.7	+3.4
Gov. Type Class.	74.5	77.1	+2.6
Gov. Transfer Class.	73.2	76.9	+3.7
Gov. QA	77.4	79.0	+1.6
Bank Activ-V	72.8	75.1	+2.3
Bank Annc-C	74.1	76.5	+2.4
Bank Report-V	72.1	75.5	+3.2
Bank Assess-I	71.9	74.3	+2.4
Bank Bus-V	70.6	73.0	+2.4
Average	73.6	75.7	+2.1

Table A.1: Performance: Dynamic vs. Fixed Weights (%)

To further validate the rationality and necessity of dynamic weight adjustment, we conduct three supplementary verification experiments:

Comparative Experiment between Dynamic and Fixed Weights: As shown in Table A.1, we set fixed weights as the average value of the entropy-weighted initial weights (each dimension = 0.2) and compare the model performance of dynamic weights vs. fixed weights across different task types (e.g., government affair domain classification, bank report violation inspection). Statistical significance for the performance differences is confirmed via t-test with $p < 0.05$. The results quantify the performance gain of dynamic adjustment in scenario-adaptive scenarios. **Convergence Verification with Bootstrap Method:** As shown in Table A.2, we sample 100 subsets from the manually annotated seed set with replacement, train the weight

Dim. (Abbr.)	Avg. Fluct.	Max. Fluct.	Conv.
Correctness (Cor)	3.1	4.2	Yes
Logical Cons. (LC)	2.8	3.9	Yes
Reliability (RL)	3.5	4.5	Yes
Format Compl. (FC)	2.6	3.7	Yes
Self-Refl. (SR)	3.3	4.3	Yes

Table A.2: Weight Fluctuation via Bootstrap Test (%)

adjustment module for 40 epochs for each subset, and calculate the weight fluctuation amplitude of each evaluation dimension (Correctness, Logical Consistency, Reliability, Format Compliance, Self Reflection). The fluctuation amplitude is defined as $\frac{|W_{40}-W_{init}|}{W_{init}} \times 100\%$, where W_{40} denotes the weight after 40 training epochs and W_{init} denotes the entropy-weighted initial weight. The verification criterion is that the fluctuation amplitude of each dimension’s weight is 5%, ensuring the stability of the dynamically adjusted weights. Ablation

Framework	Gov.	Bank	Drop
Full (Dynamic)	78.5	77.9	-
Ablated (Fixed)	76.0	76.7	-
Task Drop	-2.5	-1.2	-1.8

Table A.3: Performance: Full vs. Ablated Framework (%)

Experiment of Dynamic Weight Module: As shown in Table A.3, we remove the dynamic weight adjustment module and retain only the entropy-weighted initial fixed weights, then compare the model’s performance on the test set with and without the module. This experiment verifies whether the dynamic weight module contributes incrementally to the overall framework, avoiding methodological redundancy.

A.2 Datasets Construction

A.2.1 Preprocessed COIG-PC Dataset Details

To construct a balanced government-vertical corpus from the COIG-PC benchmark, we implemented a hybrid curation pipeline combining automated classification with generative augmentation. Initially, An LLM-based classifier was employed to categorize the raw samples into five specific sub-tasks, guided by both semantic content and predefined decision rules.. To ensure label accuracy, we conducted human-in-the-loop verification via random sampling of the classified results. Furthermore, to address data scarcity in specific categories and satisfy the pre-defined quantity requirements (e.g.,

40,000 samples for QA), we utilized LLM-driven synthetic data generation to expand the dataset, ultimately achieving a robust and balanced distribution across all tasks, and ensure a diverse coverage of both general NLP tasks and government-affairs applications.

The resulting dataset comprises 180,000 samples, structured into two primary clusters, as shown in Table A.4:

Task Category	Specific Task	Size
General	Open-domain QA	20k
	Reading Comp.	20k
	Summarization	10k
	Info. Extraction	10k
	NLI	10k
Gov. Affairs	Intent Recog.	10k
	Type Class.	10k
	Domain Class.	10k
	Workflow Routing	40k
	Gov. QA	40k
Total	-	180k

Table A.4: Statistics of the Constructed Training Dataset Derived from COIG-PC.

General NLP Tasks: Designed to maintain the model’s fundamental language understanding. This cluster includes 20,000 samples each for Open-domain QA and Reading Comprehension, along with 10,000 samples each for Summarization, Information Extraction (IE), and Natural Language Inference (NLI).

Government Affairs Tasks: Tailored to enhance performance in administrative scenarios. We formulated five specific sub-tasks: Intent Recognition, Type Classification, and Domain Classification (10,000 samples each), as well as Transfer Classification and Government QA (40,000 samples each).

For government affairs tasks, strict adherence to administrative norms and safety protocols is paramount. We conducted extensive preliminary tests on various prompt templates, optimizing factors such as wording, formatting constraints, and symbolic indicators. We designed a unified System Meta-Prompt to define the model’s identity as a helpful, harmless, and polite "Cognitive Service Assistant" developed by the [Center Name] team. This meta-prompt explicitly enforces safety guardrails, prohibiting illegal content and requiring a neutral, non-defensive tone. Furthermore, we

1213	tailored specific instruction prompts for each sub-	for semantic entities and regular expressions for	1262
1214	task, e.g., Intent Recognition, Workflow Routing,	pattern-based IDs to ensure zero leakage of sensi-	1263
1215	to standardize the JSON output format. Detailed	tive business secrets.	1264
1216	definitions of these prompts are provided in Ap-	Dataset Construction & Cleaning	1265
1217	pendix B.	The construction of the dataset followed a three-	1266
1218	Conversely, for the general NLP tasks, we de-	stage pipeline:	1267
1219	liberately abstained from using fixed prompt tem-	Taxonomy Definition We collaborated with fi-	1268
1220	plates. This strategy allows the model to interact	ancial compliance experts to define five core vio-	1269
1221	with diverse instructions, thereby preserving its	lation inspection tasks that require complex reason-	1270
1222	inherent zero-shot generalization ability and pre-	ing. These tasks correspond to critical risk control	1271
1223	venting overfitting to a single prompt style.	points in banking operations:	1272
1224	Upon establishing the prompt system, we em-	• Bank Demonstration Activity Violation In-	1273
1225	ployed DeepSeek-R1-671B as the generator model.	spection (Activ-V): Detecting non-compliant	1274
1226	By integrating the specific prompts with our pro-	marketing behaviors or unauthorized promo-	1275
1227	posed RG-CoT framework, as illustrated in Fig-	tional events, <i>such as unauthorized lot-</i>	1276
1228	ure B.3, Figure B.4, we synthesized a high-quality	<i>tery draws or exaggerated yield claims (e.g.,</i>	1277
1229	dataset embedded with explicit chain-of-thought	<i>promising "guaranteed returns" for non-</i>	1278
1230	reasoning paths, serving as the foundation for sub-	<i>principal-guaranteed products).</i>	1279
1231	sequent fine-tuning.	• Bank Announcement Compliance Inspec-	1280
1232	A.2.2 BANK-AUDIT Dataset Details	tion (Annc-C): Verifying if public announce-	1281
1233	This dataset focuses on the text auditing and com-	ments adhere to regulatory disclosure stan-	1282
1234	pliance inspection tasks within the banking and	dards <i>and ensuring the absence of sensitive</i>	1283
1235	financial sector.	<i>internal data leakage.</i>	1284
1236	Data Source & Collection	• Bank Report Violation Inspection (Report-	1285
1237	The BANK-AUDIT dataset is derived from internal	V): Identifying inconsistencies or prohibited	1286
1238	proprietary archives of a large-scale commercial	terms in internal audit reports, <i>checking for</i>	1287
1239	bank. The raw corpus primarily consists of unstruc-	<i>forbidden phrasing and logical errors in finan-</i>	1288
1240	tured and semi-structured textual data generated	<i>cial figures.</i>	1289
1241	during daily banking operations, spanning a period	• Bank Illegal Assessment Inspection (Assess-	1290
1242	from 2021 to 2024. The data sources include but	D): Flagging potential violations in credit or	1291
1243	are not limited to:	risk assessment procedures, <i>such as approving</i>	1292
1244	• Internal audit reports and risk assessment	<i>loans without required collateral documenta-</i>	1293
1245	memos;	<i>tion.</i>	1294
1246	• Publicly released bank announcements and	• Bank Business Application Violation In-	1295
1247	financial disclosure drafts;	spection (Bus-V): Checking business applica-	1296
1248	• Marketing activity proposals and promotional	tions for missing documentation or policy vi-	1297
1249	materials;	olations <i>regarding budget limits and approval</i>	1298
1250	• Business application forms and credit assess-	<i>hierarchies.</i>	1299
1251	ment logs.	Data Cleaning and Preprocessing The raw text	1300
1252	Given the high sensitivity of financial data, we	contained significant noise, such as HTML tags,	1301
1253	applied a rigorous desensitization process. All	table formatting characters, and irrelevant head-	1302
1254	Personally Identifiable Information (PII) (e.g.,	ers/footers. We employed regular expressions to	1303
1255	customer names, ID numbers, phone numbers)	clean the text and unified the format into standard	1304
1256	and sensitive institutional data (e.g., specific	plain text. Unlike standard datasets, we purpose-	1305
1257	branch codes, exact transaction amounts) were	fully retained long-form document structures to	1306
1258	anonymized using entity replacement and mask	test the model’s ability to locate subtle violations	1307
1259	tokenization (e.g., CLIENT_NAME, AMOUNT). Specif-	within extensive contexts. Text segments shorter	1308
1260	ically, we employed a hybrid approach combin-	than 50 tokens or lacking semantic completeness	1309
1261	ing BERT-based Named Entity Recognition (NER)	were discarded to ensure data quality.	1310

Annotation and Review The final training set comprises 60,000 instances. The annotation process involved a "Human-in-the-Loop" approach. Initially, expert annotators labeled a seed set of 2,000 examples. These experts were junior auditors with at least two years of experience. We calculated the inter-annotator agreement on the seed set using Fleiss' Kappa, achieving a score of $\kappa = 0.86$, indicating substantial agreement. A heuristic rule-based system then propagated these labels to the larger corpus, followed by a manual random check on 10% of the data to ensure an accuracy rate above 95%.

Task	Instances	Avg. Len	Vocab	Violations (%)
Activ-V	10,200	945	8.5k	22.4%
Annc-C	12,500	1,682	12.1k	35.6%
Report-V	15,800	2,150	15.6k	41.2%
Assess-I	11,000	786	6.8k	28.9%
Bus-V	10,500	458	4.2k	15.3%
Total	60,000	1,268	28.5k	29.8%

Table A.5: Statistics of the BANK-AUDIT dataset. "Violations" indicates the proportion of samples containing non-compliant samples.

Statistics

To ensure a robust evaluation, we split the data into training, validation, and test sets in an 8:1:1 ratio. Table A.5 summarizes the statistics. Figure B.5 and Figure B.6 presents two desensitized examples from the dataset, illustrating the input text and the corresponding audit output with reasoning rules.

B Data Examples & Prompt Templates

Figure B.1 presents the lexicon of self-reflection trigger phrases. Figure B.2 illustrates the unified prompt template configuration for government vertical tasks, detailing the global System Meta-Prompt and specific instructions for intent recognition and classification tasks. Figure B.3 and Figure B.4 demonstrate the data construction examples for domain classification and workflow classification, respectively, comparing the standard data format with the distilled CoT format that incorporates explicit reasoning paths (e.g., analyze intent, retrieve domain rules). Finally, Figure B.5 and Figure B.6 display the application in the financial compliance domain, specifically for bank report violation inspection (Report-V) and bank demonstration activity violation inspection (Activ-V), showcasing how the RG-CoT framework analyzes risks and applies regulatory rules within the <think> block

prior to generating the final verdict.

Reflection & Correction Trigger Phrases

Let's break this down, What is the question really asking?, Let's outline a plan, Wait hold on, Let me step back for a moment, I feel like I'm missing something, Let me re-read the problem statement, I made a mistake in the previous step, Ask myself, Let me backtrack, Let me rewind, This leads to a contradiction, Let me correct this, Let me rethink this approach, Let's try a different angle, Let's start over from scratch, Alternative perspective, Let's do a sanity check, The logic holds up, I am confident in this result, Conclusion, Wait there seems to be a logical gap here, Did I skip any step, Let me challenge my own assumptions, This hypothesis seems plausible but, Let me retrace my steps, I need to check, Let me cross-validate, Let me evaluate, Maybe I should, Let me assess, I need to verify, I might be, Let me ensure, Let me re-read, What if...

Figure B.1: The lexicon of trigger phrases used to elicit self-correction and reflection capabilities in the RG-CoT framework.

Full Prompt Configuration for Government Vertical Tasks

1. System Meta-Prompt (Global)

You are a Cognitive Service Assistant named [Name], a government AI assistant developed by the [Center Name] team.

- You aim to provide helpful, harmless, and accurate responses.
- You must strictly refuse to answer any questions involving illegal activities.
- Your responses must NOT be vague, accusatory, rude, controversial, off-topic, or defensive.
- Your responses must remain polite and professional at all times.

2. Task-Specific Instructions

[Task A: Intent Recognition]

Instruction: Please read the following content carefully and summarize it into a short title. Please answer in the format: ["Title": " "].

Input: {input}

[Task B: Domain Classification]

Instruction: Please read the following content carefully and judge which topic it belongs to. Optional topics include: Urban Construction, Traffic, Government Affairs, Employment, Tourism, Education, Healthcare, Environmental Protection, Agriculture/Rural/Farmers, Enterprise, Public Security, Culture & Entertainment, Finance, Sports. Please answer in the format: ["Topic": " "].

Input: {input}

[Task C: Type Classification]

Instruction: Please read the following content carefully and determine the nature of the citizen's appeal. Optional types include: Consultation, Complaint, Suggestion, Request for Assistance, Commendation, Reporting. Please answer in the format: ["Type": " "].

Input: {input}

[Task D: Transfer Classification]

Instruction: Please read the following content carefully and determine which government organization it is relevant to. Optional organizations include: Housing and Urban-Rural Development Bureau, Urban Management Bureau, Public Security Bureau, Human Resources and Social Security Bureau, Tourism/Culture/Sports Bureau, Civil Affairs Bureau, Healthcare Security Bureau, Veterans Affairs Bureau, Commerce Bureau, Agriculture and Rural Affairs Bureau, 12345 Hotline Center, Taxation Bureau, Development and Reform Commission, Industry and Information Technology Bureau, Public Complaints Bureau, Emergency Management Bureau, Justice Bureau, Cyberspace Affairs Commission, Government Services Data Bureau, Financial Supervision Bureau, Statistics Bureau, Science and Technology Bureau, Foreign Affairs Bureau, Discipline Inspection Commission, Finance Bureau, Rural Revitalization Bureau, Big Data Bureau, Ethnic and Religious Affairs Bureau, Audit Bureau, Maritime Safety Administration. Please answer in the format: ["Organization": " "].

Input: {input}

[Task E: Government QA]

Instruction: Please read the user's question carefully, understand their intent, and reply to the user in the tone of a government staff member.

User: {input}

Staff:

Figure B.2: The unified prompt template configuration used in the RG-CoT framework for government domain tasks. The configuration consists of a global System Meta-Prompt ensuring safety and tone, followed by specific instruction templates for Intent Recognition, Domain Classification, Type Classification, Transfer Classification, and Government QA. The placeholders {input} are replaced by the actual query during inference.

Standard Domain Classification Data
<p>Input: The commercial complex above Line 15 at Gulang Road, Putuo District, Shanghai, has been topped out for years but remains unfinished. Construction has stalled for over a decade, seriously affecting the living amenities of nearby residents. Taopu Smart City is a key development area; how can the developer be allowed to leave it unfinished for so long? Last year they promised to start, but what about this year?</p> <p>Output: Urban Construction</p>

Distilled CoT Data (Ours)
<p>Meta Prompt: {See <i>meta_prompt</i> in Figure B.2}</p> <p>Input: The commercial complex above Line 15 at Gulang Road, Putuo District, Shanghai, has been topped out for years but remains unfinished. Construction has stalled for over a decade, seriously affecting the living amenities of nearby residents. Taopu Smart City is a key development area; how can the developer be allowed to leave it unfinished for so long? Last year they promised to start, but what about this year?</p> <p>Output: <think> 1. Analyze Intent: The user reports a commercial project that has been "topped out" but "stalled for over a decade," questioning the lack of progress. 2. Retrieve Domain Rules: According to the government affairs taxonomy, the "Urban Construction" category governs real estate development, construction permits, and project completion oversight. Specifically, disputes involving "unfinished buildings" (<i>lanwei lou</i>) fall under the jurisdiction of the Housing and Urban-Rural Development Commission. 3. Rule Alignment: Although "Taopu Smart City" mentions planning, the core complaint is about the <i>halt of physical construction</i>, not zoning adjustments. Similarly, it is not a "Market Regulation" issue as no sales fraud is mentioned. 4. Conclusion: The terms "developer" and "construction" align strictly with the Urban Construction domain. </think> <answer> Urban Construction </answer></p>

Figure B.3: Example of domain classification. **Top:** The standard dataset maps the input text directly to the final label. **Bottom:** Our RG-CoT approach explicitly reasons about the domain logic using a meta-prompt (wrapped in <think> tags) and aligns the content with specific definitions before generating the final category.

Standard Workflow Classification Data
<p>Input: I live in the Development Zone and would like to inquire about the specific opening dates for Gantang Primary School and Gantang Kindergarten. It has been several years, yet there is no specific news. Thank you.</p> <p>Output: Education Bureau</p>

Distilled CoT Data (Ours)
<p>Meta Prompt: {See <i>meta_prompt</i> in Figure B.2}</p> <p>Input: I live in the Development Zone and would like to inquire about the specific opening dates for Gantang Primary School and Gantang Kindergarten. It has been several years, yet there is no specific news. Thank you.</p> <p>Output: <think> 1. Analyze Intent: The user explicitly asks about the "specific opening time" (operational start date) of a primary school and a kindergarten. 2. Retrieve Domain Rules: Under the government responsibility list, the Education Bureau is the primary authority responsible for educational facility planning, teacher allocation, and determining the official opening schedules of schools. 3. Rule Alignment: While the user lives in a "Development Zone" (often associated with construction), the query does not concern building quality or land zoning (which would fall under Urban Construction or Planning). The core issue is the <i>provision of educational services</i>. 4. Conclusion: Queries regarding school enrollment and operational timelines belong strictly to the Education Bureau. </think> <answer> Education Bureau </answer></p>

Figure B.4: Example of workflow classification. **Top:** The standard dataset maps the inquiry directly to the target department. **Bottom:** Our RG-CoT approach explicitly reasons about the user's intent (school opening time) and matches it with the functional responsibilities of the Education Bureau before assigning the label.

Standard Data Report-V

Input: ...IV. Contract Price and Payment Terms

4.1 The total amount of this Contract shall not exceed RMB 1,734,700 (In words: One Million Seven Hundred Thirty-Four Thousand Seven Hundred Yuan Only) (hereinafter referred to as the "Total Contract Amount"). The Total Contract Amount is inclusive of service fees, corresponding transportation expenses, Party B's profit, applicable taxes, and all costs incurred by Party B in assuming all obligations and liabilities under this Contract.

4.2 Party A shall fulfill the obligation to pay the Contract Price in accordance with the following provisions: Settlement shall be made at the end of each quarter based on the actual person-months invested by Party B during that quarter. If business travel is involved, the travel expenses shall be settled based on the actual costs incurred by Party B....

Output: Non-compliant. The phrase "actual person-months invested" should be changed to "actual services provided and deliverables accepted".

Distilled CoT Data (Ours)

Instruction: You are a senior corporate contract audit expert. Please strictly review the compliance of the provided Contract Content based on the Audit Rules.

Input: ...IV. Contract Price and Payment Terms

4.1 The total amount of this Contract shall not exceed RMB 1,734,700 (In words: One Million Seven Hundred Thirty-Four Thousand Seven Hundred Yuan Only) (hereinafter referred to as the "Total Contract Amount"). The Total Contract Amount is inclusive of service fees, corresponding transportation expenses, Party B's profit, applicable taxes, and all costs incurred by Party B in assuming all obligations and liabilities under this Contract.

4.2 Party A shall fulfill the obligation to pay the Contract Price in accordance with the following provisions: Settlement shall be made at the end of each quarter based on the actual person-months invested by Party B during that quarter. If business travel is involved, the travel expenses shall be settled based on the actual costs incurred by Party B....

Output:

<think> 1. **Initialize Audit:** I need to review the technical service contract against the provided audit rules to identify risks and propose modifications. I will cross-reference each rule one by one. 2. **Analyze Rule 1 (Deliverables):** The rule requires payment to be conditional on "acceptance of deliverables." The current service content only mentions technical support (manpower) without specifying acceptance criteria for final outcomes. This creates a risk where payment relies on time input rather than actual results. An acceptance clause must be added. 3. **Analyze Rule 2 (Payment Basis):** The rule states payment must be based on "service outcomes," not "person-months." Clause 4.2 currently settles based on "actual person-months invested." This is problematic because Party A might have to pay for time spent even if the results are substandard. ... **Refine Strategy:** I must recommend changing the settlement basis to "actual deliverables accepted" and adding a reconciliation process. I also need to check for conflicts in other clauses to ensure the modification is consistent. **Risk Assessment:** Paying for labor time without defining deliverables transfers the inefficiency risk to Party A. </think>

<rule> ... 1. Explicitly state in the Service Content clause that acceptance of deliverables by Party A is a prerequisite for settlement, ensuring obligations are linked to quality outcomes rather than just workload. 2. Payment terms must shift the settlement basis from "person-months invested" to "acceptance of service deliverables" and establish a reconciliation mechanism to prevent unilateral measurement risks. ... </rule>

<answer>Non-compliant. ... 1. Suggest adding to the end of Article 3: "Final deliverables such as research reports and planning schemes provided by Party B shall be subject to expert review and acceptance organized by Party A. Specific acceptance criteria shall be defined separately in writing by Party A."

2. Modify Clause 4.2 to: "Settlement shall be made at the end of each quarter based on the actual service deliverables completed and the corresponding acceptance documents, which must be signed and confirmed by the project managers of both parties. Expenses related to business travel shall be supported by travel vouchers approved by Party A."
...</answer>

Figure B.5: Example of bank report violation inspection (Report-V). **Top:** The standard dataset maps the contract clause directly to the correction advice. **Bottom:** Our RG-CoT approach explicitly analyzes risk factors in the <think> block and references specific compliance rules (<rule>) before generating the final modification advice.

Standard Data Activ-V

Input:

"Date: 2023-11-20\nFrom: XXX Banking Dept - Strategy Team\nTo: All Branch Managers, Wealth Advisors\nSubject: Q4 'Glory Wealth' Year-End Promotion Strategy\n\nDear Team,\n\nAs we approach the year-end closing, we need to push the sales of our new 'Glory Wealth No.5' series (Risk Level R3). To ensure we meet the Q4 targets, please strictly follow the guidelines below:\n\n1. Target Audience: Focus on conservative clients holding large expiring deposits. They are hesitant about market fluctuations, so we need to ease their concerns.\n\n2. Sales Pitch: Although the product contract explicitly states 'Non-Principal Guaranteed', our internal historical data shows a 100% payout record for this series. When communicating with clients, you are authorized to verbally assure them: 'This product functions just like a time deposit, but offers a significantly higher yield of 4.5%.' Do not overemphasize the R3 risk rating to avoid scaring them away.\n\n3. Incentives: We are launching the 'Golden Lottery' event. Any client investing over 500,000 CNY will receive one entry. The top prize is a 50g Gold Bar. Ensure this attractive prize is highlighted in your pitch.\n\nLet's go all out for a strong finish!"

Output: Non-Compliant

Distilled CoT Data (Ours)

Meta Prompt: You are the "Marketing Violation Review Expert System" developed by the Bank's Internal Control and Compliance Department. Your core responsibility is to conduct a comprehensive compliance scan on input [Marketing Proposals / Sales Scripts / Notifications]. Please make determinations based on the following core regulatory principles (including but not limited to): Rule 1: Misleading Sales and False Advertising The use of absolutist terms such as "principal guaranteed," "zero risk," or "guaranteed redemption" is strictly prohibited for non-deposit products (e.g., wealth management products, funds, and insurance). Furthermore, marketing content must not exaggerate past performance or emphasize returns without clearly disclosing the associated risks. Rule 2: Illegal Deposit Gathering and Deposit Management Marketing activities must not involve "gifts for deposits" (e.g., gold bars, shopping cards), interest subsidies, or unauthorized interest rate hikes. It is strictly prohibited to attract deposits by offering disguised high interest rates through the provision of physical goods or cash rebates. Rule 3: Investor Suitability Management Compliance checks must ensure that the target customer group aligns with the product's risk rating. It is strictly prohibited to market high-risk products to investors with low risk tolerance, such as the elderly or conservative investors.

Input:

"Date: 2023-11-20\nFrom: XXX Banking Dept - Strategy Team\nTo: All Branch Managers, Wealth Advisors\nSubject: Q4 'Glory Wealth' Year-End Promotion Strategy\n\nDear Team,\n\nAs we approach the year-end closing, we need to push the sales of our new 'Glory Wealth No.5' series (Risk Level R3). To ensure we meet the Q4 targets, please strictly follow the guidelines below:\n\n1. Target Audience: Focus on conservative clients holding large expiring deposits. They are hesitant about market fluctuations, so we need to ease their concerns.\n\n2. Sales Pitch: Although the product contract explicitly states 'Non-Principal Guaranteed', our internal historical data shows a 100% payout record for this series. When communicating with clients, you are authorized to verbally assure them: 'This product functions just like a time deposit, but offers a significantly higher yield of 4.5%.' Do not overemphasize the R3 risk rating to avoid scaring them away.\n\n3. Incentives: We are launching the 'Golden Lottery' event. Any client investing over 500,000 CNY will receive one entry. The top prize is a 50g Gold Bar. Ensure this attractive prize is highlighted in your pitch.\n\nLet's go all out for a strong finish!"

Output:

<think> 1. **Intent Recognition:** Identified internal strategy email instructing specific sales pitches and incentives for 'Glory Wealth No.5'. 2. **Entity Extraction:** - *Product:* R3 Risk Level (Medium Risk). - *Target:* Conservative clients (Expiring deposits). - *Prize:* 50g Gold Bar (Est. Value ≈ 24,000 CNY in late 2023). - *Pitch:* "Verbally assure... functions just like a time deposit". 3. **Violation Review:** Prize value (24k) is below the 50,000 CNY statutory limit. → **Pass (Compliant).** - *Rule 1 (Misleading Sales):* Explicit instruction to verbally guarantee an R3 (Non-Principal Guaranteed) product as a "Time Deposit" constitutes a "False Guarantee". → **Severe Violation.** - *Rule 3 (Suitability):* Deliberately marketing R3 products to conservative clients while hiding risks violates suitability regulations. → **Severe Violation.** 4. **Severity Grading:** - Although the lottery prize is legal, the core sales strategy involves systemic fraud (verbal misrepresentation) and unsuitability. - **Conclusion:** Level 1 - Severe. Immediately revoke the verbal comparison to deposits. Must disclose "Non-Principal Guaranteed". Shift marketing focus to "Balanced" or "Aggressive" investors only. </think>

<rule> Rule 1: Any verbal or written comparison of non-principal-guaranteed wealth products (like R3 funds) to "time deposits" is an immediate severe violation. "Verbal assurance" cannot override the "Non-Principal Guaranteed" clause in the contract; doing so is considered fraud.... </rule>

<answer> Non-Compliant Revise script to strictly disclose risks; Stop targeting conservative clients </answer>

Figure B.6: Example of bank demonstration activity violation inspection (Activ-V): **Top:** Standard dataset input with a realistic internal email. **Bottom:** Our RG-CoT approach correctly distinguishes between compliant and non-compliant elements within the same text.