

Selective Compliance in Financial Rule Reasoning: A Solver-Verified Chinese Benchmark and Empirical Analysis

Peng Li Yinchun Wang
ailab@chinaamc.com

Abstract

Selective compliance is important for financial LLM applications: an applicable context rule should override the LLM’s parametric rule, while an inapplicable rule should be ignored. Though previous studies recognize the importance of rule selection, dedicated evaluations tailored specifically to the financial domain remain absent. To bridge this gap, we build a controlled benchmark based on the CFLUE dataset, where intervention rules are synthesized and FOL logic is applied to ensure the correctness of the new derived answer.¹ Furthermore, we designed three evaluation settings covering clean inputs, noisy inputs, and multi-hop reasoning scenarios. We conduct a systematic comparison across different rule selection strategies and models. Experiments show that: (1) For recent advanced LLMs, single-pass integrated reasoning approach (CAR-Joint) consistently outperforms the decoupled ‘route-then-answer’ approach (CAR-Route). (2) Our proposed post-hoc review method, CAR-Joint-Review, outperforms CAR-Joint on four of the five evaluated models. Similarly, pre-hoc routing via the parametric rule elicitation method CAR-Elicit-Route surpasses CAR-Route in multi-hop settings. The performance gains are model-dependent. These findings reveal the trade-offs across various methods and LLMs in financial selective compliance.

1 Introduction

Retrieval-Augmented Generation (RAG) equips LLMs with domain knowledge (Lewis et al., 2020), yet retrieved contexts in finance are not always applicable. Blindly following inapplicable provisions risks non-compliant or legally incorrect answers (Mallen et al., 2023; Cuconasu et al., 2024; Yoran et al., 2024). We formalize this challenge as *Financial Selective Compliance*: a model must

adopt a retrieved rule only when it applies and alters the outcome, but retain its parametric knowledge otherwise. Unlike standard memory-conflict benchmarks (Longpre et al., 2021; Neeman et al., 2023; Xu et al., 2024), this task requires verifying complex conditions and propagating logical consequences.

While financial benchmarks evaluate general domain capabilities (Zhu et al., 2024; Guo et al., 2025; Xie et al., 2024b) and rule-following tasks test reasoning over realistic documents (Zhou et al., 2025; Diao et al., 2025), none isolate the decision between applying answer-changing rules and resisting answer-preserving ones. To address this, we construct a solver-verified benchmark from the CFLUE (Zhu et al., 2024) test split. Using rule-dependent examination questions, we build executable First-Order Logic (FOL) backbones, retaining those validated by round-trip symbolic solving. We apply controlled symbolic interventions, labeling answer-changing rules as *conflict* and answer-preserving ones as *supplement*. The benchmark construction process is illustrated in Figure 1. To measure both failure modes—under-compliance (refusing applicable rules) and over-compliance (adopting inapplicable rules)—we evaluate Override Rate (OR) on conflict rules, Retention Rate (RR) on supplement rules, and their harmonic mean, FR. This two-sided evaluation prevents models from optimizing context-following at the expense of answer stability.

We compare direct answering and Chain-of-Thought (CoT) prompting against four applicability-oriented strategies: CAR-Joint, CAR-Joint-Review, CAR-Route, and CAR-Elicit-Route. Specifically, CAR-Joint utilizes a tailored prompt to handle applicability judgment and reasoning simultaneously, whereas CAR-Route adopts a two-stage route-then-answer strategy. Drawing upon the principles of ASTUTE RAG (Wang et al., 2024a), CAR-Joint-Review incorporates a post-hoc

¹The dataset is available at https://github.com/nb003pl-dotcom/fin_select_reason/.

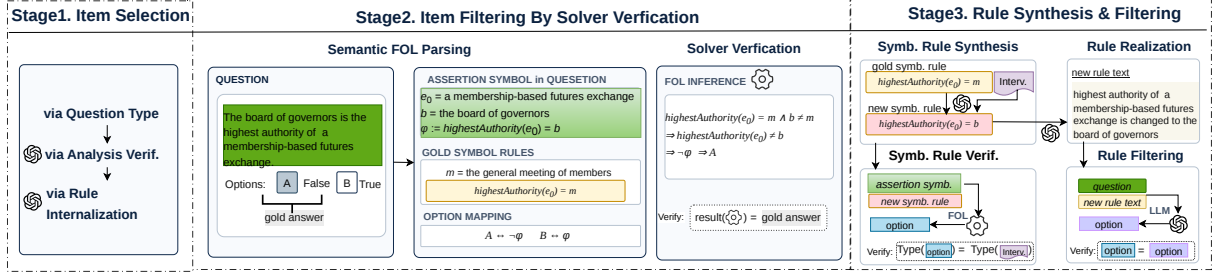


Figure 1: Overview of the solver-verified benchmark construction pipeline. Stage 1 selects rule-dependent questions; Stage 2 retains instances that can be represented in First-Order Logic (FOL) and successfully verified by a round-trip solver; and Stage 3 synthesizes rule interventions and filters them through symbolic and realization checks.

review mechanism that re-introduces context rules to re-evaluate the trajectory. Meanwhile, CAR-Elicit-Route employs a rule-elicitation scheme for routing, requiring the model to first answer based on its internal rules and subsequently determine applicability by comparing its response against the context rules.

Results show that joint reasoning methods (CAR-Joint and CAR-Joint-Review) substantially surpass the route-then-answer paradigm (e.g., CAR-Route) on LLMs except Qwen2.5-7B. CAR-Joint-Review improves mean FR over CAR-Joint on four of five models, with gains ranging from 0.004 to 0.028. Further investigation indicates that CAR-Route fails primarily due to poor routing accuracy under supplement interventions. CAR-Elicit-Route mitigates routing failures in multi-hop settings via internal rule elicitation.

Our primary contributions are threefold: First, we formalize *Financial Selective Compliance* to evaluate a model’s ability to selectively adhere to or resist retrieved rules. Second, we construct a solver-verified Chinese financial benchmark featuring controlled conflict and supplement interventions across Clean, Noisy, and Multi-Hop scenarios. Third, we propose Contextual-Applicability-Aware-Reasoning strategies and empirically evaluate their effectiveness across diverse model capabilities.

2 Related Work

Context-memory knowledge selection for RAG. Our study belongs to the context-memory conflict family surveyed by Xu et al. (2024), where parametric priors interact with input-provided information. Prior studies separate parametric and contextual answers (Longpre et al., 2021; Neeman et al., 2023), evaluate whether LLMs follow conflicting evidence (Xie et al., 2024a; Wu et al., 2024a; Su et al., 2024; Ming et al., 2025), and study whether models should trust, reject, or expose both sides of a con-

flict (Wang et al., 2024b; Huang et al., 2025). Retrieval can also be harmful, introducing irrelevant or misleading passages and encouraging models to over-trust context (Mallen et al., 2023; Cuconasu et al., 2024; Yoran et al., 2024). Several methods try to address knowledge selection in RAG, including selective reliance on non-parametric memory (Mallen et al., 2023), conflict analysis in retrieval-augmented models (Jin et al., 2024), source-aware consolidation of LLM-internal and retrieved knowledge (Wang et al., 2024a), and adaptive decoding for detected conflicts (Yuan et al., 2024). These studies mainly treat external context as factual evidence whose reliability needs to be judged. In our setting, by contrast, answer correctness hinges on whether the retrieved rule is applicable to the question.

Rule following and counterfactual reasoning.

Another line of work studies whether LLMs can follow explicit rules or reason in counterfactual settings. Paired default/counterfactual protocols test whether models reason under perturbed task conditions rather than reciting memorized patterns (Wu et al., 2024b). Rule-focused benchmarks evaluate realistic rule documents with distractors (Zhou et al., 2025), domain guidelines and rule updates (Diao et al., 2025), or legal-style decompositions of rule elements (Servantez et al., 2024). These settings ask whether a model can apply a provided rule set once that rule set is selected. Another line of findings shows that stronger context reliance can help context-only tasks but hurt tasks requiring parametric knowledge (Ying et al., 2024; Sun et al., 2026). Our benchmark evaluates selective rule-following specifically tailored to the financial domain.

Financial benchmarks. Financial NLP benchmarks cover Chinese finance and broad financial capabilities, including CFLUE (Zhu et al., 2024),

FinEval (Guo et al., 2025), CFinBench (Nie et al., 2025), FinBen (Xie et al., 2024b), and CNFinBench (Ding et al., 2026), but mainly measure static knowledge, task performance, or risk-control behavior. Closer work studies financial rule diagnosis (Malarkkan et al., 2026), real-time finance knowledge updating (Tang et al., 2026), or conflicts between financial priors and contextual evidence (Lee et al., 2025). In contrast, we transform CFLUE items into dynamic intervention-rule tests with FOL backbones and solver-derived conflict/supplement labels, targeting post-retrieval selective compliance rather than static financial knowledge or rule-violation detection.

3 Benchmark

3.1 Task Definition

Given a question q , a contextual rule r , and a candidate option set C , our objective is to evaluate whether a model $\mathcal{M}$ correctly predicts an answer $a \in C$, where $a = \mathcal{M}(q, C, r)$. Assuming $\mathcal{M}$ has internalized the default gold rule r^{gold} , then $\mathcal{M}(q, C) = \mathcal{M}(q, C, r^{\text{gold}}) = a^*$, where a^* is gold answer. To test a model’s capacity to dynamically select and apply contextual rules, we formulate two scenarios:

- **Conflict Rules (Rule Updating):** The contextual rule r applies to q but contradicts r^{gold} . A flexible model should prioritize r over its internal prior and produce an updated answer $a' \neq a^*$ ($a' \in C$).
- **Supplement Rules (Rule Noise):** The contextual rule r is inapplicable to q . A robust model should ignore inapplicable and answer-changing rule r , and rely on r^{gold} to produce a^* .

3.2 Evaluation Metrics

Let $\mathcal{D}_{\text{conf}}$ be the set of cases in which context rule is applicable and changes the answer from a^* to a' , and let $\mathcal{D}_{\text{sup}}$ be the set of cases in which the target intervention clause in context is not applicable and no applicable answer-changing update is present, so the internal-rule answer should be retained. For an output $\hat{a}_i = \mathcal{M}(q_i, C_i, r_i)$, we define

$$\text{OR} = \frac{1}{|\mathcal{D}_{\text{conf}}|} \sum_{i \in \mathcal{D}_{\text{conf}}} \mathbf{1}[\hat{a}_i = a'_i],$$

$$\text{RR} = \frac{1}{|\mathcal{D}_{\text{sup}}|} \sum_{i \in \mathcal{D}_{\text{sup}}} \mathbf{1}[\hat{a}_i = a_i^*].$$

$$\text{FR} = \frac{2 \cdot \text{OR} \cdot \text{RR}}{\text{OR} + \text{RR}},$$

Here, OR (Override Rate) measures how often a model follows an applicable rule that conflicts with its internal knowledge. RR (Retention Rate) measures how often the model preserves its original answer when presented with an independently constructed, condition-false distractor clause. RR is therefore a behavioral answer-retention metric; the final answer alone is not used to infer whether the model internally ignored or processed the clause. FR (F1 for robustness) is an aggregation of OR and RR. FR penalizes one-sided strategies: always following context can improve OR but hurts RR, whereas always keeping the internal answer can improve RR but hurts OR.

3.3 Dataset Construction

We build benchmark from the test split of CFLUE, a Chinese financial-domain examination dataset. The CFLUE test set contains 3,864 items drawn from accounting, banking, securities, funds, futures, and related regulatory domains. Each item provides a question, a set of answer choices, a gold answer, and an expert-written analysis. The analysis field contains the rules and reasoning process for answer. The core idea is to synthesize item-specific rules based on the analysis field and the answer choices, where first-order logic (FOL) is applied to ensure the correctness of synthesis.

Stage 1: Item Selection. We select high-quality benchmark items through a three-step filtering pipeline: (1) Question Type Selection: We keep single-choice and true/false items with valid option sets and non-empty expert analyses, excluding those that rely on visual context or exceed 400 Chinese characters. (2) Analysis Verification: An annotator model $\mathcal{M}_{\text{ann}}$ ² verifies that the analysis field supports the gold answer and categorizes items as rule-dependent, computation-based, or non-parametric. We retain only answer-consistent, rule-dependent items governed by modifiable regulatory parameters (e.g., statutory ratios, deadlines, monetary thresholds, or eligibility criteria). (3) Rule Internalization Verification: To ensure the evaluated model $\mathcal{M}_{\text{subj}}$ ³ possesses the required parametric knowledge, we query it three times per item without external context and retain only items for which it consistently predicts the gold answer.

² $\mathcal{M}_{\text{ann}}$: Qwen3.7-Max

³ $\mathcal{M}_{\text{subj}}$: DeepSeek-V4-Flash

Code	Type	Construction operation	Example
CR	Conflict	Constant Replacement	IPO supervision lasts 3 years $\rightarrow$ it lasts 2 years.
CB	Conflict	Conditional Branching	the right covers layout design $\rightarrow$ under an exclusive contract, it covers binding design.
EI	Conflict	Exception Insertion	guarantors without agreed shares have joint liability $\rightarrow$ they have proportional liability.
ID	Supplement	Inapplicable Distractor	supervision lasts 2 years $\rightarrow$ it lasts 3 years only for major restructurings.

Table 1: Intervention types for benchmark construction. Conflict Rules: Synthesized rules whose activation conditions evaluate to true under the question facts, with answer-changing effects verified via symbolic execution. Supplement Rules: An ID supplement is retained only if its activation condition is false under the question facts and symbolic execution confirms that the original answer remains unchanged.

Stage 2: Item Filtering via FOL Verification.

To guarantee that each item admits a deterministic formal derivation, we verify candidate questions using First-Order Logic (FOL). First, the annotation model $\mathcal{M}_{\text{ann}}$ parses each question into a set of FOL clauses $B = (A, R, C)$, where A , R , and C correspond to the question assertions, the gold symbolic rules (extracted from the expert analysis), and the answer option mappings, respectively. Next, we run an automated FOL solver $g(\cdot)$ over these representation clauses. We retain an item if and only if the solver derives a unique candidate option that matches the gold answer, i.e., $g(A, R) = a^*$. This filtering process eventually yields 512 verified items.

Stage 3: Rule Synthesis and Filtering.

This stage synthesizes contextual rules to construct new test items under controlled interventions: (1) Symbol Rule Synthesis: We employ annotation model $\mathcal{M}_{\text{ann}}$ to modify the symbolic rules in Stage 2, yielding a refined rule set: $E = \{p_j \leftarrow e'_j(A, R, C)\}_{j=1}^m$ for each question. Specifically, we establish two directional categories of rule edits: answer-changing conflict operations and answer-preserving supplement operations. The detailed edit types under each direction are summarized in Table 1. To ensure that the new rules remain evaluable, we require that the expected answer after rule edits stay within the predefined option set. The instructions for these specific edits are provided as prompts to $\mathcal{M}_{\text{ann}}$. We uniformly distribute the rule sampling quota across all edit types within each category. (2) Synthesized Symbol Rule Verification: The modified rule set $E(R)$ is executed by the symbolic solver to ensure validity. Let $a' = g(A, E(R), C)$ represent the solver’s derived answer under the modified rules. Conflict synthesized rules are retained if it produces an altered answer ($a' \neq a$). Supplement synthesized rules must satisfy the following two gates: (i) it is inapplicable to question. Suppose ϕ_s is the activation

predicate based on the variables in the verified question facts and synthesized rule parameters, and it must evaluate to false under the current item: $\phi_s(A, R) = \text{False}$. (ii) the solver-derived answer is preserved, $g(A, R \cup \{r_s\}, C) = g(A, R, C) = a^*$. Crucially, these labels and target answers are determined deterministically by symbolic execution rather than LLM judgment. (3) Synthesized Rule Realization: generator model $\mathcal{M}_{\text{ann}}$ converts the synthesized symbolic rules into natural language text. To prevent generating shortcut text, we incorporate prompt constraints forbidding option leakage during conversion. (4) Synthesized Rule Filtering: To make sure natural language realization faithfully expresses symbolic intent, we apply two independent LLM-based verifiers $\mathcal{M}_{\text{ver}}^4$, $\mathcal{M}_{\text{ver}2}^5$. Given the question, options, and rendered natural-language rule, the verifiers predict the answer. A synthesized item is retained only if the verifier’s output aligns perfectly with the solver-derived target answer a' . We manually spot-check a random sample of constructed instances to ensure correctness.

3.4 Evaluation Settings

To evaluate model behavior across diverse reasoning demands, we evaluate models under three distinct settings that vary in context length, noise level, and reasoning depth.

Clean Setting. This serves as the basic benchmark setting, where the model receives a question alongside a single contextual intervention rule (either a conflict or supplement rule). This setting directly measures the fundamental ability of a model to perform selective compliance. Following the procedure described in Section 3.3, this evaluation set comprises 126 curated items.

Noisy Setting. This setting evaluates a model’s ability to localize relevant rules before reasoning.

⁴ $\mathcal{M}_{\text{ver}}$: DeepSeek-V4-Pro

⁵ $\mathcal{M}_{\text{ver}2}$: GPT-5.4-mini

Specifically, we insert the target rule from the Clean setting among 8–12 topically related distractor clauses generated by an LLM. This applies only on Conflict type interventions. The underlying question, symbolic intervention, and target answer remain identical to the Clean setting.

Multi-Hop Setting. This setting focuses on reasoning depth rather than contextual noise. Rules belong to the Multi-Hop only if their symbolic dependency distance to the final choice predicate is at least two. We initially selected 34 items from the Clean setting whose rules satisfy the multi-hop criteria. To complement the dataset size, we repeated the rule synthesis and multi-hop filtering to retrieve an additional 33 items, yielding 67 items in total.

Each base item in the Clean and Multi-Hop settings is paired with two rule intervention types (*conflict* and *supplement*), whereas the Noisy setting contains conflict interventions only. The dataset statistics is provided in Table 2.

Setting	Conflict			Supplement	
	CR	CB	EI	ID	Total
Clean	78	32	16	126	252
Noisy	78	32	16	0	126
Multi-Hop	57	4	6	67	134
All	213	68	38	193	512

Table 2: Dataset statistics across rule intervention types and experimental settings.

4 Experiments

4.1 Compared Methods and Models

Inspired by context-memory selection paradigms in RAG-based factual QA (Wang et al., 2024a), we adapt retrieval-and-selection principles to the task of dynamic rule selection by introducing four Contextual-Applicability-Aware Reasoning (CAR) strategies:

CAR-Joint (Joint Inference). Assesses rule applicability and derives the final answer in a single end-to-end forward pass. The selection constraints and prioritization criteria are directly embedded within the prompt instructions.

CAR-Joint-Review (Joint Inference with Post-Hoc Review). Shown in Algorithm 1, this strategy introduces a secondary verification pass. Following initial generation via CAR-Joint, an independent critique prompt reviews the execution trace τ_0 , re-evaluating the applicability of the contextual rule and revising the answer accordingly.

CAR-Route (Discrete Hard-Routing). Implements a strict "route-then-answer" paradigm. A dedicated routing module first determines whether the contextual rule applies to the question, after which reasoning proceeds exclusively with or without the external rule.

CAR-Elicit-Route (Elicitation-Based Routing). Shown in Algorithm 2, this strategy follows an "elicit-to-route and then answer" pipeline. It first prompts the model to externalize its internal rule trace and generate a preliminary answer a_i . The model then evaluates the applicability of contextual rule r conditioned on this internal trace, preventing the router from overlooking internal parametric knowledge.

To benchmark the proposed strategies against standard prompting paradigms, we evaluate three baseline methods:

Oracle. Serves as an empirical upper bound by manually providing the ground-truth context configuration: for conflict rules, the model receives both the rule and the question; for supplement rules, it receives the question alone. This represents an optimal diagnostic ceiling rather than a deployable strategy.

Direct-CoT. Provides the contextual rule and question alongside standard step-by-step chain-of-thought instructions.

Direct. Prompts the model with the contextual rule and question to directly predict the final option without intermediate reasoning.

Implementation Details. We evaluate all strategies across two advanced models—DeepSeek-V4-Flash and Claude-Sonnet-4.6—and three commonly used lightweight models—Qwen3-14B, Qwen3-8B and Qwen2.5-7B. The hyperparameter configurations and decoding parameters are set according to official model recommendations. All the metrics reported in the following sections are averaged over three independent runs.

4.2 Main Results

Performance across Methods Table 3 presents the mean OR, RR, and FR across all evaluated models and methods, with each metric averaged across settings. Although absolute knowledge internalization cannot be fully guaranteed for models beyond DeepSeek-V4-Flash, intra-model comparisons remain fully meaningful as all methods on a given

Algorithm 1 CAR-Joint-Review

Require: Question q , contextual rule r , subject model $\mathcal{M}$, CAR-Joint prompt p_{imp} , review prompt p_{rev}

- 1: Produce the initial answer and trace: $(a_0, \tau_0) \leftarrow \mathcal{M}(p_{\text{imp}}, q, r)$
 - 2: Review and correct answer if rule applicable: $\rho \leftarrow \mathcal{M}(p_{\text{rev}}, q, r, \tau_0)$
 - 3: Parse the reviewed answer: $a \leftarrow \text{ParseReview}(\rho)$
 - 9: **return** a
-

Algorithm 2 CAR-Elicit-Route

Require: Question q , contextual rule r , subject model $\mathcal{M}$, internal-rule prompt p_{int} , routing prompt p_{route} , rule-world prompt p_{ext}

- 1: Elicit the internal rule trace and preliminary answer without r : $(a_i, \tau_i) \leftarrow \mathcal{M}(p_{\text{int}}, q)$
 - 2: Judge external-rule applicability: $\rho \leftarrow \mathcal{M}(p_{\text{route}}, q, r, \tau_i)$
 - 3: Parse the Boolean route: $g \leftarrow \text{ParseRoute}(\rho)$
 - 4: **if** $g = \text{applicable}$ **then** $(a, \tau) \leftarrow \mathcal{M}(p_{\text{ext}}, q, r)$; **else** $a \leftarrow a_i$
 - 5: **return** a
-

model share identical parametric knowledge.

From the table, we observe the following findings: (1) For recent LLMs (except Qwen2.5-7B), single-pass integrated reasoning (CAR-Joint) outperforms the decoupled ‘route-then-answer’ approach (CAR-Route). (2) Our proposed CAR-Joint-Review with a post-hoc critique phase can substantially improves the overall FR metric against CAR-Joint (benefiting all models except Sonnet 4.6). Specifically, For DeepSeek-V4-Flash, FR is improved from 0.917 to 0.933 and for Qwen3-8B, FR is improved from 0.770 to 0.798. (3) For earlier weak models Qwen2.5-7B, CAR-Route emerges as the top-performing method, better than CAR-Joint and CAR-Joint-Review, showing its limitedness of integrated reasoning.

Performance across Evaluation Settings. Figure 2(a) reports Clean and Multi-Hop FR together with Noisy OR. The performance ordering across all models generally adheres to the pattern of Clean > Noisy > Multi-Hop. Multi-Hop yields the lowest FR, while Clean and Noisy are close. This indicates that multi-step dependencies make rule selection more challenging.

Performance across Question Types. Figure 2(b) presents performance across single-choice (SC) and binary true/false tasks. Except for DeepSeek-V4-Flash, performance on SC across all models is consistently lower than that on the binary task. This gap aligns with the lower random guessing probability in four-option SC compared to binary tasks. Thus, SC offers a more rigorous test of rule-selection proficiency.

Performance across Intervention Types. Figure 2(c–d) profiles the efficacy of each model’s best method under various rule synthesis types. For conflict type, DeepSeek-V4-Flash achieves a notably

higher score on CB (0.986) than Sonnet 4.6 (0.927), showcasing its enhanced robustness in identifying conditional premise changes (CB). For supplement interventions, the performance of DeepSeek-V4-Flash and Sonnet 4.6 are close.

4.3 Analysis

Integrated Applicability Reasoning versus Hard Routing. Table 3 demonstrates that for advanced models like DeepSeek-V4-Flash, the simpler integrated applicability reasoning method CAR-Joint significantly outperforms CAR-Route (hard routing). This finding is counterintuitive, as conventional wisdom suggests that executing multiple tasks within a single joint prompt should be inherently more difficult than focusing on a single, isolated task. For analysis, we examine the routing error rate of DeepSeek-V4-Flash under the CAR-Route method, as detailed in Table 4. The data reveals that the hard router achieves a low error rate of 3.6% on conflict contexts but a much higher error rate of 43.4% on the available Clean and Multi-Hop supplement contexts. To better understand the effect, we analyze a supplement case given in Table 5, where CAR-Route fails while CAR-Joint succeeds. The case reveals that standalone routing suffers from false positives by classifying irrelevant rules as applicable, thereby corrupting CAR-Route’s execution path. Besides, We further observe that a portion of the supplementary passages contains both a synthesized, inapplicable ID target clause and applicable background clauses that restate or preserve the original rule. This mixed-applicability structure makes hard routing ill-posed: CAR-Route must assign a single passage-level label, whereas the benchmark label is defined at the target-clause level.

Method	DeepSeek-V4-Flash Mean OR / RR / FR	Sonnet 4.6 Mean OR / RR / FR	Qwen3-14B Mean OR / RR / FR	Qwen3-8B Mean OR / RR / FR	Qwen2.5-7B Mean OR / RR / FR
Oracle [†]	0.961 / 0.947 / 0.954	0.948 / 0.928 / 0.938	0.920 / 0.715 / 0.805	0.917 / 0.766 / 0.835	0.841 / 0.794 / 0.817
CAR-Joint	0.894 / 0.942 / 0.917	0.931 / 0.932 / 0.932	0.917 / 0.706 / 0.798	0.890 / 0.679 / 0.770	0.836 / 0.572 / 0.679
CAR-Joint-Review	0.938 / 0.927 / 0.933	0.849 / 0.945 / 0.895	0.887 / 0.738 / 0.805	0.899 / 0.717 / 0.798	0.825 / 0.604 / 0.697
CAR-Route	0.924 / 0.842 / 0.881	0.915 / 0.899 / <u>0.907</u>	0.927 / 0.625 / 0.747	0.919 / 0.655 / 0.765	0.769 / 0.678 / 0.721
CAR-Elicit-Route	0.935 / 0.848 / 0.889	0.922 / 0.891 / 0.906	0.914 / 0.623 / 0.741	0.924 / 0.647 / 0.761	0.731 / 0.645 / 0.686
Direct-CoT	0.882 / 0.879 / 0.880	0.895 / 0.865 / 0.880	0.931 / 0.679 / 0.785	0.889 / 0.657 / 0.755	0.792 / 0.599 / 0.682
Direct	0.852 / 0.879 / 0.865	0.898 / 0.880 / 0.889	0.842 / 0.628 / 0.719	0.888 / 0.659 / 0.757	0.741 / 0.663 / <u>0.700</u>

Table 3: Model comparison across methods. Bold and underlined numbers denoting the best and second-best results for each model. DeepSeek-V4-Flash and Sonnet4.6 favors CAR-Joint while Qwen2.5-7B perfers CAR-Route.

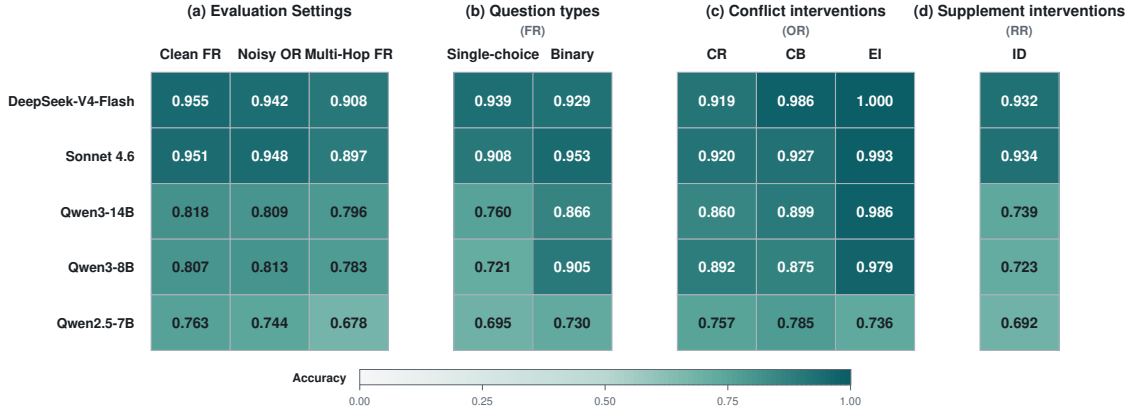


Figure 2: Panel (a) reports FR for Clean and Multi-Hop and OR for Noisy. Panel (b) reports overall FR by question type. Panels (c) and (d) report OR for conflict rules across all settings and RR for ID supplement rules across Clean and Multi-Hop, respectively.

Scenario	Overall	Conflict	Supplement
Clean	0.208	0.026	0.389
Noisy	0.026	0.026	–
Multi-Hop	0.266	0.055	0.478
Macro avg.	0.167	0.036	0.434

Table 4: Routing error rate for CAR-Route using DeepSeek-V4-Flash.

Effects of Post-Hoc Review. As shown in Table 3, CAR-Joint-Review substantially improves upon CAR-Joint through its review mechanism. To better understand this variation, Figure 3 breaks down the changes in OR, RR, and FR across experimental settings. The gains are twofold: (1) OR Enhancement: Re-injecting context rules during review elevates rule adherence, as demonstrated by DeepSeek-V4-Flash; (2) RR Enhancement: Re-assessing rule applicability reinforces inapplicability signals and rectifies first-pass logical errors across all other models. Table 6 provides a concrete case analysis illustrating the correction.

Effects of Internal-Rule Elicitation. Prior RAG methods resolve factual conflicts by activating internal knowledge and consolidating it with external context. CAR-Elicit-Route extends this prin-

An Example of Hard-Route Failure

Question. In a commercial bank, who bears primary responsibility for capital management?
A. Board of supervisors; B. Board of directors; C. Senior management; D. Shareholders’ meeting.
Synth. Ctx. Rule. A bank shall designate the body with primary responsibility for capital management. If senior management is so designated, it bears that responsibility and formulates and executes the relevant policies.
CAR-Route trace. Ctx. Rule => applicable, wrong path → C (error).
CAR-Joint trace. Condition absent → B (correct; gold B).

Table 5: An illustration where CAR-Route fails but CAR-Joint succeeds. Outputs are from the DeepSeek-V4-Flash evaluation traces.

ciple to rule selection. Figure 4 summarizes its change relative to CAR-Route across all three settings. The benefit of eliciting internal rule is concentrated in Multi-Hop settings: FR rises by 0.030 on DeepSeek-V4-Flash, 0.027 on Qwen3-8B, and 0.013 on Sonnet 4.6. However, elicitation lowers Clean FR for all five models. This stems from the fact that most queries in the Clean setting involve only basic single-step reasoning, thereby deriving little benefit from additional elicited rules or explicit reasoning steps. Conversely, Multi-Hop set-

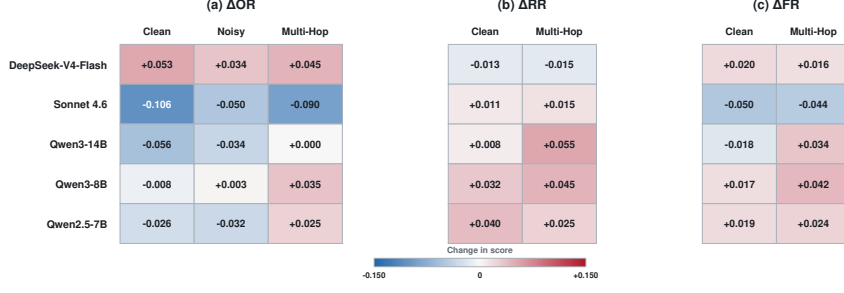


Figure 3: Effect of post-hoc answer review relative to integrated reasoning across models and settings. Each cell reports $\Delta = \text{CAR-Joint-Review} - \text{CAR-Joint}$ for OR, RR, or FR. CAR-Joint-Review raises overall RR for four of the five models.

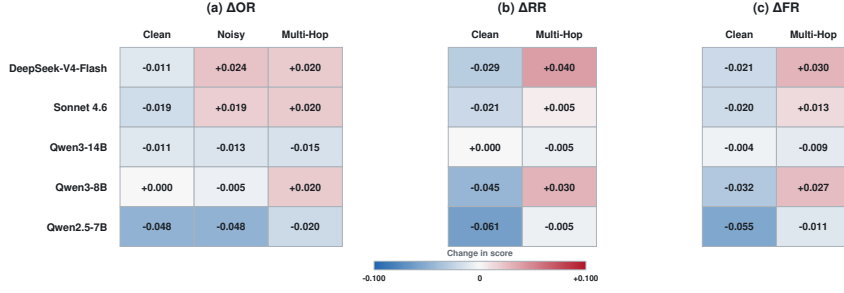


Figure 4: Effect of internal-rule elicitation relative to hard routing across models and regimes. Each cell reports $\Delta = \text{CAR-Elicit-Route} - \text{CAR-Route}$ for OR, RR, or FR. CAR-Elicit-Route has an advantage in the Multi-Hop setting for three of the five models.

tings requiring multi-step reasoning, eliciting the intermediate reasoning process uncovers crucial diagnostic signals for effective rule selection.

5 Limitations

The reported results should be interpreted as selective compliance evaluation within a controlled, given-rule environment, rather than a comprehensive evaluation of end-to-end financial RAG systems. The synthesized rules do not necessarily capture the full spectrum and complexity of real-world financial regulations.

Post-Hoc Review Corrects an Inconsistent Answer	
Question.	An enterprise withholds personal income tax for managerial staff by debiting “Administrative Expenses” and crediting “Other Payables.” (A) False. (B) True. Gold: A.
Synth. Ctx. Rule.	Tax withheld on labor remuneration paid to <i>non-employees</i> is outside employee compensation and must be credited to “Other Payables.”
CAR-Joint trace.	“Managerial staff are employees, so the non-employee precondition is not met and the rule does not apply.” Nevertheless, it concludes that crediting “Other Payables” is valid: B (wrong).
CAR-Joint-Review trace.	The applicability judgment is correct, but the answer is not. Tax withheld for an employee belongs to employee compensation and should be credited to “Employee Compensation Payable,” not “Other Payables.” Therefore the statement is false: A (correct).

Table 6: An example where CAR-Joint-Review help correct error in CAR-Joint

6 Conclusion

We introduce *Financial Selective Compliance* along with a solver-verified benchmark spanning Clean, Noisy, and Multi-Hop settings. We propose a series of contextual-applicability-aware methods and conduct systematic evaluations across widely used LLMs. Our findings reveal that stronger models excel with integrated applicability reasoning CAR-Joint, and the proposed post-hoc review CAR-Joint-Review outperforms CAR-Joint on all models except Sonnet 4.6. Notably, the failure of the route-then-answer baseline stems from the frequent misclassification of inapplicable rules; the superiority of CAR-Joint over CAR-Route further demonstrates that applicability assessment achieves higher precision when intrinsically coupled with answer derivation rather than as an isolated preceding step.

References

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. *The power of noise: Redefining retrieval for RAG systems*. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

- Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. 2025. [GuideBench: Benchmarking domain-oriented guideline following for LLM agents](#). In *ACL*. ArXiv:2505.11368.
- Jinru Ding, Chao Ding, Yidong Jiang, and 1 others. 2026. [Beyond knowledge to agency: Evaluating expertise, autonomy, and integrity in finance with CNFinBench](#). *KDD*. ArXiv:2512.09506.
- Xin Guo, Haotian Xia, Zhaowei Liu, and 1 others. 2025. [FinEval: A chinese financial domain knowledge evaluation benchmark for large language models](#). *NAACL*. ArXiv:2308.09975.
- Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025. [To trust or not to trust? enhancing large language models' situated faithfulness to external contexts](#). *arXiv preprint arXiv:2410.14675*. ICLR 2025.
- Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. [Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.
- Hoyoung Lee, Junhyuk Seo, Suhwan Park, Junhyeong Lee, Wonbin Ahn, Chanyeol Choi, Alejandro Lopez-Lira, and Yongjae Lee. 2025. [Your AI, not your view: The bias of LLMs in investment analysis](#). *arXiv preprint arXiv:2507.20957*. Accepted at ICAIF 2025.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *EMNLP*. ArXiv:2109.05052.
- Arun Vignesh Malarkkan, Manan Roy Choudhury, Guangwei Zhang, Vivek Gupta, Qingyun Wang, Yanjie Fu, and Denghui Zhang. 2026. [FinRule-Bench: A benchmark for joint reasoning over financial tables and principles](#). *arXiv preprint arXiv:2603.11339*.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822. Association for Computational Linguistics.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. [FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"](#). In *ICLR*. ArXiv:2410.03727.
- Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. [DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering](#). *Proceedings of ACL*. ArXiv:2211.05655.
- Ying Nie, Binwei Yan, Tianyu Guo, and 1 others. 2025. [CFinBench: A comprehensive chinese financial benchmark for large language models](#). *NAACL*. ArXiv:2407.02301.
- Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. [Chain of logic: Rule-based reasoning with large language models](#). *Findings of ACL*. ArXiv:2402.10400.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, and 1 others. 2024. [ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs](#). In *NeurIPS Datasets and Benchmarks*. ArXiv:2408.12076.
- Kaiser Sun, Fan Bai, and Mark Dredze. 2026. [Task matters: Knowledge requirements shape LLM responses to context–memory conflict](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4154–4176. Association for Computational Linguistics.
- Chenming Tang, Yutong Yang, Kexue Wang, and Yunfang Wu. 2026. [Aligning language models with real-time knowledge editing](#). *ACL*. ArXiv:2508.01302.
- Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arik. 2024a. [Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models](#). *arXiv preprint arXiv:2410.07176*.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024b. [Resolving knowledge conflicts in large language models](#). In *COLM*. ArXiv:2310.00935.
- Kevin Wu, Eric Wu, and James Zou. 2024a. [ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence](#). In *NeurIPS Datasets and Benchmarks*. ArXiv:2404.10198.
- Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024b. [Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks](#). In *NAACL*. ArXiv:2307.02477.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024a. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts](#). In *ICLR*. ArXiv:2305.13300.

- Qianqian Xie, Weiguang Han, Zhengyu Chen, and 1 others. 2024b. [FinBen: A holistic financial benchmark for large language models](#). *NeurIPS Datasets and Benchmarks*. ArXiv:2402.12659.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for LLMs: A survey](#). *Proceedings of EMNLP*. ArXiv:2403.08319.
- Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. [Intuitive or dependent? investigating LLMs’ behavior style to conflicting prompts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246. Association for Computational Linguistics.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. [Making retrieval-augmented language models robust to irrelevant context](#). In *ICLR*.
- Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. [Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint](#). *Findings of ACL*. ArXiv:2402.11893.
- Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. 2025. [RuleArena: A benchmark for rule-guided reasoning with LLMs in real-world scenarios](#). *Proceedings of ACL*. ArXiv:2412.08972.
- Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. [Benchmarking large language models on CFLUE: A chinese financial language understanding evaluation dataset](#). *Findings of ACL*. ArXiv:2405.10542.

A Prompt Templates for Main Methods

All experiments use the original Chinese prompts. For readability, this appendix gives faithful English translations of the templates for the four CAR methods; the verbatim Chinese templates are included in the accompanying source code. The placeholders {rule}, {question}, {choices}, {internal-rule-trace}, and {first-pass-answer-trace} denote the contextual rule, exam question, answer choices, elicited internal-rule trace, and complete first-pass response, respectively.

Shared Answer Constraint for Answer-Generating Calls

The following constraint is appended to each answer-generating system prompt, but not to the routing or review prompts:
The final line must strictly follow this format: Answer: X, where X is the option letter.

CAR-Joint

Prompt for CAR-Joint (p_{imp})

System:

You are a professional in the Chinese financial domain. Answer financial exam questions. Process the given provision in two steps.

First, **Applicability judgment**: inspect the applicability condition of each provision (applicable subject, business type, amount range, or factual situation), and determine separately whether the facts of the question satisfy that condition. Only provisions whose conditions are satisfied take effect for this question. A provision whose condition is not satisfied, even if it is relevant or tempting, must not be used to change the answer.

Second, **Answering**: reason under the provisions that take effect. If an effective provision conflicts with your memory, follow the provision and cite its parameter value. If none of the provisions' conditions are satisfied, answer using your original professional judgment. [Shared answer constraint]

User:

Provision: {rule}

Question: {question}

Choices: {choices}

CAR-Joint-Review. As specified in Algorithm 1, the first call uses the CAR-Joint prompt above. The second call receives the full first-pass response and independently reviews rule applicability and rule use.

Prompt for CAR-Joint-Review: Review Call (p_{rev})

System:

You are an independent answer-verification expert. Based on the given question and rule, review the LLM answer and output the reviewed answer.

Focus on determining: (1) whether the given rule applies to the question; and (2) whether the LLM answer applies the rule correctly.

User:

Given Question

Question: {question}

Choices: {choices}

Given Rule

{rule}

LLM Answer

{first-pass-answer-trace}

New Output

CAR-Route. CAR-Route first produces a Boolean applicability decision. An applicable rule is sent to the external rule-world answer call; an inapplicable rule is excluded and the question is answered from internal knowledge. If the routing output cannot be parsed as the required JSON object, the implementation falls back to CAR-Joint.

Prompt for CAR-Route: Routing Call (p_{route})

System:

You are an external-rule applicability classifier. You will see: (1) an external provision; and (2) the original question.

Your task is not to answer the question, and you must not output a route. Only determine whether the external provision applies to this question and could affect its answer. Output JSON only:

```
{"applicable": true|false,  
 "conflict_with_internal": true|false,  
 "reason": "one-sentence explanation"}
```

Prompt for CAR-Route: Routing Call (p_{route}) (continued)

Decision criteria:

`applicable=true`: At least one rule in the external provision has an applicable subject, business type, amount range, time condition, or factual situation that matches the question, and the provision supplies an effective rule capable of affecting the answer. If that effective external rule conflicts with the internal rule, set `conflict_with_internal=true`.

`applicable=false`: The external provision is irrelevant or insufficient, or the applicability conditions of all relevant external rules are not satisfied by the facts of the question. Even if an external provision differs from the internal rule in a parameter, it must be labeled `applicable=false` when it does not apply to this question.

Do not label a provision `applicable=true` merely because it differs from the internal rule; first confirm that it applies to the question.

User:

External provision: {rule}

Question: {question}

Choices: {choices}

Prompt for CAR-Route: External Answer Call (p_{ext})

System:

You are a professional in the Chinese financial domain operating in **rule-world mode**. The provision below is the only valid rule in this world. Any old rule in your memory that differs from it has ceased to be valid in this world and must not be used. During reasoning, explicitly cite the key parameter value in the provision. [Shared answer constraint]

User:

Provision: {rule}

Question: {question}

Choices: {choices}

Prompt for CAR-Route: Internal Answer Call

System:

You are a professional in the Chinese financial domain. Answer financial exam questions. [Shared answer constraint]

User:

Question: {question}

Choices: {choices}

CAR-Elicit-Route. As specified in Algorithm 2, CAR-Elicit-Route first elicits an internal rule and preliminary answer without revealing the external rule. The elicited trace is then supplied to the applicability router. An applicable route invokes the external answer call; an inapplicable route reuses the preliminary answer. Invalid routing JSON falls back to CAR-Joint.

Prompt for CAR-Elicit-Route: Internal-Rule Elicitation Call

System:

You are a professional in the Chinese financial domain. First, without considering any external supplementary provision, use only the question and your existing professional knowledge to write the internal rule needed for the question, its key applicability conditions, and the preliminary answer derived from that internal rule. [Shared answer constraint]

User:

Question: {question}

Choices: {choices}

CAR-Elicit-Route Routing Prompt

System:

You are an external-rule applicability classifier. You will see: (1) the internal rule elicited by the model from the question; (2) an external provision; and (3) the original question.

Your task is not to answer the question, and you must not output a route. Only determine whether the external provision applies to this question and could affect its answer. Output JSON only:

```
{"applicable": true|false,
  "conflict_with_internal": true|false,
  "reason": "one-sentence explanation"}
```

Decision criteria:

`applicable=true`: At least one rule in the external provision has an applicable subject, business type, amount range, time condition, or factual situation that matches the question, and the provision supplies an effective rule capable of affecting the answer. If that effective external rule conflicts with the internal rule, set `conflict_with_internal=true`.

`applicable=false`: The external provision is irrelevant or insufficient, or the applicability conditions of all relevant external rules are not satisfied by the facts of the question. Even if an external provision differs from the internal rule in a parameter, it must be labeled `applicable=false` when it does not apply to this question.

CAR-Elicit-Route Routing Prompt (continued)

Do not label a provision `applicable=true` merely because it differs from the internal rule; first confirm that it applies to the question.

User:

Internal Rule and Preliminary Answer

`{internal-rule-trace}`

External Provision

`{rule}`

Question: `{question}`

Choices: `{choices}`

Prompt for CAR-Elicit-Route: External Answer Call

System:

You are a professional in the Chinese financial domain operating in **rule-world mode**. The provision below is the only valid rule in this world. Any old rule in your memory that differs from it has ceased to be valid in this world and must not be used. During reasoning, explicitly cite the key parameter value in the provision. [Shared answer constraint]

User:

Provision: `{rule}`

Question: `{question}`

Choices: `{choices}`

B Detailed Results by Model

Method	Clean OR / RR / FR	Noisy OR	Multi-Hop OR / RR / FR	Mean OR / RR / FR
Oracle	0.968 / 0.968 / 0.968	0.968	0.945 / 0.925 / 0.935	0.961 / 0.947 / 0.954
CAR-Joint	0.913 / 0.958 / <u>0.935</u>	0.907	0.861 / 0.925 / <u>0.892</u>	0.894 / 0.942 / <u>0.917</u>
CAR-Joint-Review	0.966 / 0.944 / 0.955	0.942	0.905 / 0.910 / 0.908	0.938 / 0.927 / 0.933
CAR-Route	0.947 / 0.884 / 0.914	0.929	0.896 / 0.801 / 0.846	0.924 / 0.842 / 0.881
CAR-Elicit-Route	0.937 / 0.854 / 0.894	0.952	0.915 / 0.841 / 0.875	0.935 / 0.848 / 0.889
Direct-CoT	0.923 / 0.892 / 0.907	0.907	0.816 / 0.866 / 0.840	0.882 / 0.879 / 0.880
Direct	0.902 / 0.892 / 0.897	0.857	0.796 / 0.866 / 0.829	0.852 / 0.879 / 0.865

Table 7: Performance of DeepSeek-V4-Flash across various methods and experimental settings. Model calls use temperature 0 and a maximum generation length of 8,000 tokens. Noisy reports OR only. Overall OR uses all three settings, overall RR uses Clean and Multi-Hop, and overall FR is their harmonic mean. Bold denotes the best value and underline denotes the second-best value among non-Oracle methods.

Method	Clean OR / RR / FR	Noisy OR	Multi-Hop OR / RR / FR	Mean OR / RR / FR
Oracle	0.979 / 0.937 / 0.957	0.974	0.891 / 0.920 / 0.905	0.948 / 0.928 / 0.938
CAR-Joint	0.963 / 0.939 / 0.951	0.960	0.871 / 0.925 / 0.897	0.931 / 0.932 / 0.932
CAR-Joint-Review	0.857 / 0.950 / 0.901	0.910	0.781 / 0.940 / 0.853	0.849 / 0.945 / 0.895
CAR-Route	0.939 / 0.897 / 0.917	0.960	0.846 / 0.900 / 0.872	0.915 / 0.899 / <u>0.907</u>
CAR-Elicit-Route	0.921 / 0.876 / 0.898	0.979	0.866 / 0.905 / <u>0.885</u>	0.922 / 0.891 / 0.906
Direct-CoT	0.947 / 0.884 / 0.914	0.923	0.816 / 0.846 / 0.831	0.895 / 0.865 / 0.880
Direct	0.942 / 0.899 / <u>0.920</u>	0.937	0.816 / 0.861 / 0.838	0.898 / 0.880 / 0.889

Table 8: Performance of Claude-Sonnet 4.6 across various methods and experimental settings. Model calls use temperature 0 and a maximum generation length of 8,000 tokens. Noisy reports OR only. Overall OR uses all three settings, overall RR uses Clean and Multi-Hop, and overall FR is their harmonic mean. Bold denotes the best value and underline denotes the second-best value among non-Oracle methods.

Method	Clean OR / RR / FR	Noisy OR	Multi-Hop OR / RR / FR	Mean OR / RR / FR
Oracle	0.944 / 0.743 / 0.832	0.910	0.905 / 0.687 / 0.781	0.920 / 0.715 / 0.805
CAR-Joint	0.944 / 0.746 / 0.833	0.926	0.881 / 0.667 / 0.759	0.917 / 0.706 / <u>0.798</u>
CAR-Joint-Review	0.889 / 0.754 / <u>0.816</u>	0.892	0.881 / 0.721 / <u>0.793</u>	0.887 / 0.738 / 0.805
CAR-Route	0.944 / 0.608 / 0.740	0.921	0.915 / 0.642 / 0.754	0.927 / 0.625 / 0.747
CAR-Elicit-Route	0.934 / 0.608 / 0.736	0.907	0.900 / 0.637 / 0.745	0.914 / 0.623 / 0.741
Direct-CoT	0.942 / 0.651 / 0.770	0.915	0.935 / 0.706 / 0.804	0.931 / 0.679 / 0.785
Direct	0.884 / 0.579 / 0.700	0.807	0.836 / 0.677 / 0.748	0.842 / 0.628 / 0.719

Table 9: Performance of Qwen3-14B across various methods and experimental settings. Model calls disable thinking and use temperature 0.7, top- p 0.8, top- k 20, min- p 0, and a maximum generation length of 8,000 tokens. Noisy reports OR only. Overall OR uses all three settings, overall RR uses Clean and Multi-Hop, and overall FR is their harmonic mean. Bold denotes the best value and underline denotes the second-best value among non-Oracle methods.

Method	Clean OR / RR / FR	Noisy OR	Multi-Hop OR / RR / FR	Mean OR / RR / FR
Oracle	0.947 / 0.841 / 0.891	0.905	0.900 / 0.692 / 0.782	0.917 / 0.766 / 0.835
CAR-Joint	0.915 / 0.696 / <u>0.790</u>	0.907	0.846 / 0.662 / 0.741	0.890 / 0.679 / <u>0.770</u>
CAR-Joint-Review	0.907 / 0.728 / 0.807	0.910	0.881 / 0.706 / <u>0.783</u>	0.899 / 0.717 / 0.798
CAR-Route	0.934 / 0.643 / 0.761	0.937	0.886 / 0.667 / 0.761	0.919 / 0.655 / 0.765
CAR-Elicit-Route	0.934 / 0.598 / 0.729	0.931	0.905 / 0.697 / 0.787	0.924 / 0.647 / 0.761
Direct-CoT	0.902 / 0.661 / 0.763	0.905	0.861 / 0.652 / 0.742	0.889 / 0.657 / 0.755
Direct	0.905 / 0.656 / 0.761	0.889	0.871 / 0.662 / 0.751	0.888 / 0.659 / 0.757

Table 10: Performance of Qwen3-8B across various methods and experimental settings. Model calls disable thinking and use temperature 0.7, top- p 0.8, top- k 20, min- p 0, and a maximum generation length of 8,000 tokens. Noisy reports OR only. Overall OR uses all three settings, overall RR uses Clean and Multi-Hop, and overall FR is their harmonic mean. Bold denotes the best value and underline denotes the second-best value among non-Oracle methods.

Method	Clean OR / RR / FR	Noisy OR	Multi-Hop OR / RR / FR	Mean OR / RR / FR
Oracle	0.881 / 0.847 / 0.863	0.815	0.826 / 0.741 / 0.781	0.841 / 0.794 / 0.817
CAR-Joint	0.876 / 0.571 / 0.692	0.802	0.831 / 0.572 / 0.678	0.836 / 0.572 / 0.679
CAR-Joint-Review	0.849 / 0.611 / <u>0.711</u>	0.770	0.856 / 0.597 / <u>0.702</u>	0.825 / 0.604 / 0.697
CAR-Route	0.812 / 0.720 / 0.763	0.770	0.726 / 0.637 / 0.678	0.769 / 0.678 / 0.721
CAR-Elicit-Route	0.765 / 0.659 / 0.708	0.722	0.706 / 0.632 / 0.667	0.731 / 0.645 / 0.686
Direct-CoT	0.841 / 0.571 / 0.680	0.794	0.741 / 0.627 / 0.679	0.792 / 0.599 / 0.682
Direct	0.804 / 0.630 / 0.706	0.688	0.731 / 0.697 / 0.713	0.741 / 0.663 / <u>0.700</u>

Table 11: Performance of Qwen2.5-7B across various methods and experimental settings. Model calls use temperature 0 and a maximum generation length of 8,000 tokens. Noisy reports OR only. Overall OR uses all three settings, overall RR uses Clean and Multi-Hop, and overall FR is their harmonic mean. Bold denotes the best value and underline denotes the second-best value among non-Oracle methods.

C Synthetic Benchmark Examples

This appendix presents one actual benchmark record from each regime. Chinese text is retained as evaluated. The rules below are *synthetic fictional revisions*, not legal advice or statements of current law. Each record contains a conflict rule that must change the answer and a supplement rule that must leave it unchanged. For supplement passages, inapplicability is assigned to the target ID clause; other background clauses may restate an applicable rule but cannot introduce an applicable answer-changing update.

C.1 Clean

Record. cflue_0563; strategy CR; one hop.

Original item (Chinese).

下列关于持续督导期间的说法中，正确的有（ ）。

I. 主板上市公司发行新股、可转换公司债券的，持续督导的期间为证券上市当年剩余时间及其后1个完整会计年度。

II. 首次公开发行股票并在创业板上市的，持续督导的期间为证券上市当年剩余时间及其后3个完整会计年度。

III. 创业板上市公司发行新股、可转换公司债券的，持续督导的期间为证券上市当年剩余时间及其后2个完整会计年度。

IV. 持续督导期届满，如有尚未完结的保荐工作，保荐机构应当继续完成。

A: I、III、IV B: I、II、III、IV C: I、II、III D: I、II、IV

The original answer is B.

FOL backbone (abridged). The four facts are the durations claimed in statements I–III (1, 3, and 2 years) and whether unfinished work continues (true). The corresponding parameters are $p_{\text{main}} = 1$, $p_{\text{GEM-IPO}} = 3$, $p_{\text{GEM-issue}} = 2$, and $p_{\text{continue}} = \text{true}$. Each statement is checked by equality with its parameter; choice B is true exactly when all four statements are true.

Synthetic conflict rule.

《证券发行上市保荐业务管理办法》
(2026年修订)第二十八条
一、首次公开发行股票并在主板上市的，持续督导期间为证券上市当年剩余时间及其后2个完整会计年度。
二、首次公开发行股票并在创业板上市的，持续督导期间为证券上市当年剩余时间及其后2个完整会计年度。
三、持续督导期届满，如有尚未完结的保荐工作，保荐机构应当继续完成。

The intervention sets $p_{\text{GEM-IPO}} : 3 \mapsto 2$. Statement II becomes false while I, III, and IV remain true, so the solver recomputes the answer as A.

Synthetic supplement rule.

《证券发行上市保荐业务管理办法》
(2026年修订)第二十九条
一、创业板上市公司发行新股、可转换公司债券的，持续督导期间为证券上市当年剩余时间及其后2个完整会计年度。[raw III]
二、创业板上市公司发行新股涉及重大资产重组的，持续督导期间为证券上市当年剩余时间及其后3个完整会计年度。[Inapplicable Distractor]

Clause 1 is an answer-preserving background clause that restates Statement III. Clause 2 is the target ID intervention: its major-restructuring condition is not satisfied by the question facts. The target intervention is therefore inapplicable, and the passage contains no applicable answer-changing update. The supplement does not change the original derivation, so the answer remains B. Closed-book pass, FOL round trip, blind conflict verification, and supplement verification all pass.

C.2 Noisy: Bury the Rule in an extend Document

Controlled pairing. This is the noisy version of the same cflue_0563 item. The question, FOL backbone, CR intervention, supplement, and answers (B→A for conflict; B→B for supplement) are unchanged. Only the presentation of the conflict rule changes. The operative clause is bolded below for exposition; models receive the document without highlighting.

Synthetic conflict rule (noisy excerpt).

《证券发行上市保荐业务管理办法》
(2026年修订)节选 第四章 持续督导
第二十五条 保荐机构应当建立健全保荐工作的内部控制制度，明确保荐工作流程，规范保荐业务操作，防范利益冲突。
第二十六条 保荐机构及其保荐代表人履行保荐职责，应当遵守法律、行政法规和中国证监会的相关规定。
第二十七条 首次公开发行股票并在科创板上市的，持续督导期间为证券上市当年剩余时间及其后3个完整会计年度。
第二十八条 一、首次公开发行股票并在主板上市的，持续督导期间为证券上市当年剩余时间及其后**2**个完整会计年度。二、首次公开发行股票并在创业板上市的，持续督导期间为证券上市当年剩余时间及其后**2**个完整会计年度。三、持续督导期届满，如有尚未完结的保荐工作，保荐机构应当继续完成。
第二十九条 上市公司发行新股、可转换公司债券的，持续督导期间为证券上市当年剩余时间及其后1个完整会计年度。
第三十条 持续督导期间，保荐机构应当督导发行人规范运作，履行信息披露义务，完善公司治理结构。
第三十一条 保荐机构应当在发行人披露年度报告之日起的十个工作日内，向监管机构报送持续督导跟踪报告。
第三十二条 发行人出现重大风险事项、重大

违规行为或者信息披露存在重大遗漏的，保荐机构应当进行专项现场检查并报告。

The solver target is still A, but the model must first locate Article 28 among unrelated governance, reporting, and inspection clauses. All four construction checks pass.

C.3 Multi-Hop: Numeric Propagation

Record. cflue_0243; strategy CR; three hops.

Original item (Chinese).

某企业职工王某的月工资为1600元，当地社会平均工资为3000元。则该职工每月应缴纳的基本养老保险费是（ ）。

A: 330 B: 240 C: 120 D: 144

The original answer is D.

FOL backbone (abridged). Let salary $s = 1600$, local average $m = 3000$, lower-limit ratio $\ell = 0.6$, and personal contribution ratio $r = 0.08$. The executable derivation is

$$\begin{aligned} b_{\min} &= m\ell, \\ b_{\text{actual}} &= \max(s, b_{\min}), \\ c &= b_{\text{actual}}r. \end{aligned}$$

Originally, $b_{\min} = 1800$, $b_{\text{actual}} = 1800$, and $c = 144$, yielding D.

Synthetic conflict rule.

《中华人民共和国社会保险法》（2026年修订）第十二条

（一）用人单位应当按照国家规定的本单位职工工资总额的比例缴纳基本养老保险费。

（二）职工个人缴纳基本养老保险费的比例为百分之七点五。本人月平均工资低于当地职工月平均工资百分之五十的，按当地职工月平均工资的百分之五十作为缴费基数。

（三）无雇工的个体工商户、非全日制从业人员以及其他灵活就业人员，应当按照国家规定缴纳基本养老保险费。

The intervention sets $(\ell, r) : (0.6, 0.08) \mapsto (0.5, 0.075)$. The solver propagates the edit through three hops: $b_{\min} = 1500$, $b_{\text{actual}} = 1600$, and $c = 120$, so the conflict answer is C.

Synthetic supplement rule.

《中华人民共和国社会保险法》（2026年修订）第十三条

（一）职工月工资高于当地社会平均工资的，以当地社会平均工资作为缴费基数计算个人养老保险费。

（二）社会保险经办机构应当依法核定用人单位和个人的缴费基数。

The condition “salary above the local average” is false ($1600 < 3000$), so the supplement is inapplicable and the original answer remains D. Closed-book pass, FOL round trip, blind conflict verification, and supplement verification all pass.