ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

RC-CoSA: Controllable secure alignment architecture for large language models based on risk-constrained inference search

Linghao Meng[1,*]

[1] School of Electrical Automation and Information Engineering, Tianjin University, Tianjin, China, 300072

SUMMARY: *The current secure alignment of large language models (LLMs) generally adopts a static paradigm, that is, training a single model through predefined general principles. However, this approach lacks flexibility in the face of diverse security needs in different cultural backgrounds, geographical norms, and specific application scenarios. At the same time, re-aligning models for each segment requirement will bring high computing costs and engineering overhead. To this end, we propose a risk-constrained controllable safety alignment architecture (RC-CoSA), which aims to adapt the model to diverse and intertwined safety requirements in the inference stage without updating the underlying model parameters. Compared with existing methods that rely on single-sample autoregressive generation, RC-CoSA improves the robustness and controllability of response generation under complex security configurations through compliance-first best-of-N candidate screening, structured security completion for partial-compliance scenarios, and decoupling multi-stage reasoning-evaluation process. The experimental results show that the actual benefits of RC-CoSA have a certain base dependence: on the DeepSeek base, the proposed method significantly reduces the Helpful + Unsafe ratio from 11.0% to 0.5%, and increases the CoSA-Score to 0.596, and improves the overall information validity. On the GPT-4o base, RC-CoSAlign also increased the CoSA-Score from 0.288 to 0.349 and the Helpful + Safe from 50.8% to 61.9%, but its compression of the risk of violations is relatively limited. On the Llama3.1-8B-INST base, although the inference period enhancement can improve the comprehensive control performance, its inhibition stability against the risk of violation is still affected by the characteristics of the base model. The above results show that RC-CoSA, as an inference-period execution control framework, can effectively improve the comprehensive controllability of the model under complex security configurations, but its benefit intensity is still affected by the original security boundary, generation distribution and instruction compliance ability of the base model.*

KEYWORDS: *Large Language Models, Risk-Constrained Controllable Safety Alignment, Best-of-N Optimization, Inference-Time Adaptation*

1 Introduction

With the rapid improvement of large language model capabilities, Safety Alignment has become an indispensable core research direction (Jiang et al., 2025)[1]. Traditional alignment paradigms are often predefined by model providers with a common set of security principles or constitutions (e.g., Constitutional AI (Bai et al., 2022b; Huang et al., 2024b))[2, 3], and trained a single static safety model accordingly. However, this static paradigm fundamentally ignores

*linghaomeng@tju.edu.cn
https://doi.org/10.65102/is20261283
Published: 10 June, 2026

the diversity and dynamics of human values across cultural backgrounds, regional norms, and specific application scenarios (Lake et al., 2025; Zhang et al., 2025; Chen et al., 2025)[4-6]. For user groups with specific security needs, such as game developers who need to relax the censorship of specific content, or professionals conducting criminal analysis, the generic model is often too harsh and lacks practical value. On the other hand, due to the sheer scale of modern large-scale model post-training, retraining or aligning models for each customized security requirement will face extremely high computational costs and engineering overhead (Fang et al., 2026)[7].

In order to break the limitations of homogeneity, the CoSA framework has been widely proposed. This paradigm aims to enable a single base model to adapt to diverse security requirements at the inference stage without having to retrain for each new safety configuration through "safety configs" in natural language. Although existing data-driven methods such as CoSAlign give models the ability to follow complex security instructions to a certain extent, they still rely on traditional single-sample autoregressive generation (Zhao et al., 2025)[8] in the face of complex configurations (i.e., partial-compliance) where "allow and prohibit rules are intertwined" in real scenarios (Zhao et al., 2025)[8]. This static generation method is limited by the space of policy uncertainty in the decoding process, which can easily lead to output fluctuations, overall indiscriminate rejection, or local content crossing, thus seriously weakening the collaborative expression of information effectiveness and safety in complex contexts (Li et al., 2025)[9].

It is important to emphasize that we are not trying to replace CoSAlign, a method of controllable safety alignment during training. The core function of CoSAlign is to enable the model to obtain the basic compliance ability of diverse safety configurations through data construction and preference optimization. However, we are concerned about another problem: after this capability has been learned, the model may still have local outbounds, overall rejections, and response fluctuations due to the execution instability generated by single-sample autoregression in partial-compliance and unseen configurations scenarios. Focusing on this residual gap, we propose RC-CoSA to advance the research focus from "how to learn configurable security capabilities" to "how to stably implement this capability in the inference stage". Specifically, our work makes a core contribution in four dimensions:

1. Multi-channel candidate space optimization mechanism under risk constraints (Risk-Constrained Multi-Candidate Inference and Selection)

Aiming at the problem that large language models are prone to output fluctuations and occasional violations under complex security configurations, we construct a best-of-N inference period candidate selection mechanism. Without modifying the parameters of the underlying model, the mechanism combines multiple candidate generation, configuration compliance screening and helpful sorting to make hierarchical selection of candidate responses under risk constraints. The design aims to improve the decoding robustness of the model in multi-dimensional conflict constraint scenarios, and to make the candidate screening process in the inference stage consistent with the CoSA-Score in the evaluation direction.

2. A structured security complement paradigm for the "partial compliance" dilemma (Structured Safe-Completion Paradigm for Partial-Compliance)

To address the common problem of overall denial or partial de-en-bounds in the "intertwining of allowed and prohibited content" scenarios, we propose a structured security completion protocol for partial-compliance prompts. The protocol explicitly introduces a three-stage generation structure based on [ALLOWED], [DISALLOWED], and [SAFE ALTERNATIVES] at the inference template layer, and decomposes complex compliance decisions into executable local subtasks through output organizational constraints. This paradigm helps to preserve the effective coverage of permissible information while clarifying

the security boundary, thereby improving the synergistic expression of helpfulness and security in complex contexts.

3. Highly scalable decoupled multi-stage reasoning-evaluation architecture (Highly Scalable Decoupled Multi-Stage Inference-Evaluation Architecture)

In order to alleviate the difficulty of method access and modules caused by the deep coupling of the existing evaluation process and specific generation algorithms, we design a decoupled multi-stage reasoning-evaluation architecture. By introducing a Response View Mapping middle layer, the architecture organizes the process into a modular pipeline of candidate generation - view mapping - independent evaluation - policy selection. This design not only supports "plug-and-play" for various inference enhancement strategies, but also provides a clearer engineering basis for subsequent ablation analysis, module replacement and method reuse.

4. Compliance-led decision-making guidelines with symbolic consistency (Compliance-Prioritized Hierarchical Decision Criterion)

In response to the problem of "helpfulness improvement accompanied by security shift" that may occur in the process of multi-objective alignment, we propose a compliance-first hierarchical decision-making criterion. The guideline takes configuration compliance as the primary criterion for candidate feasibility and further ranks it according to helpfulness in the candidate set that satisfies configuration constraints. When none of the candidates meet the configuration requirements, the Pessimistic Fallback strategy is adopted to reduce potential risk exposure. This decision-making approach helps reduce the likelihood of misselection of highly helpful but non-compliant responses and improves the stability of the final selection result under complex security configurations.

2 Related work

In recent years, the diversified alignment and cross-cultural security of large language models have gradually become the core issues of academic concern. Traditional security alignment paradigms often rely on a single predefined constitution or static principle, aiming to train a single model that meets common safety standards.

However, this static alignment strategy fundamentally ignores the diversity and dynamics of human values in different cultural backgrounds, geographical norms, and specific application scenarios (Lake et al., 2025; Zhang et al., 2025)[4, 5]. Once social norms evolve or face user groups with specific security needs, existing general-purpose models need to be retrained with significant computational resources. Compared with writing a single and fixed principle into the model weight, recent studies have begun to focus on how to make the model more flexible to respond to differentiated security configurations and multicultural needs while maintaining a unified base.

In order to achieve training-free model adaptation, in-context alignment (ICA) has received widespread attention as a lightweight alternative to the inference period (Han, 2023; Lin et al., 2024)[10, 11]. This method mainly guides the behavior boundary of the model when infering by introducing natural language rules and few-shot demonstrations into the system prompt. While ICA has shown some potential in handling underlying general-purpose alignment tasks, building high-quality prompt examples that cover all edge cases for highly complex and fine-grained security configurations is not only resource-intensive, but also significantly limits its effectiveness. Especially when faced with the "partial-compliance" scenario where permissible and prohibition rules are intertwined, the model is prone to policy degradation, output fluctuations, or local outbounds due to instruction overload. These phenomena indicate that it is still difficult to stably handle partial-compliance scenarios by relying solely on prompt word

injection, and the existing methods still have obvious deficiencies in the candidate selection in the generation structure control and inference stages.

In terms of achieving a multi-objective balance between information effectiveness and safety, existing explorations have mainly focused on model retraining (Bai et al., 2022a; Wu et al., 2023)[12, 13], parameter pooling (Rame et al., 2023; Jang et al., 2023)[14, 15] and decoding-time alignment mechanisms. Decoding period alignment usually relies on customized external reward functions to rescore and intervene on some generated content (Shi et al., 2024; Mudgal et al., 2024; Deng & Raffel, 2023)[16-18]. However, this paradigm faces a severe scalability bottleneck: for each newly introduced security objective or custom configuration, the system must relearn or train a specific reward model, severely hindering its ability to adapt flexibly to unknown security specifications. In addition, the existing multi-objective evaluation pipeline is often strongly bound to specific generation algorithms. In summary, existing studies still face two problems under complex security configurations: first, the lack of stable local compliance control in the inference stage, and second, the evaluation process is often strongly bound to specific generation algorithms, which limits method expansion and module comparison. Our work will focus on these two issues later.

3 CoSA framework: an approach to controllable security alignment

3.1 CoSAlign: Data-based approach implements security control during inference

The CoSA framework is based on the construction of a large language model that can flexibly adjust its own security boundaries through natural language security configuration instructions (Zhang et al., 2025)[19]. To achieve this foundational capability, previous studies have proposed a data-centric approach called CoSAlign. The core advantage of this approach is that it can achieve controllable security at scale without having to retrain the model for each new security rule by only using the initial training prompt set containing security and risk queries. Through this data-driven paradigm, the model can directly follow the complex security configuration requirements given in the system prompt during the inference stage, thus overcoming the limitations of traditional alignment methods in the face of diverse application scenarios.

In the specific implementation, CoSAlign first performs clustering and feature extraction of training prompts, so as to construct a systematic risk taxonomy containing multiple subdivided risk categories. Subsequently, to solve the challenge of correlation and diversity in the data generation process, the method reverses the generation of virtual security configurations that match various prompts, and uses advanced language models as referees (LLM-as-a-judge) to generate and evaluate multiple responses (Pitis et al., 2024)[20]. On this basis, CoSAlign introduces an error scoring mechanism to assign different penalty weights to responses containing allowable risks, prohibited risks, or unresolved cues, thereby synthesizing high-quality pairwise preference datasets. Finally, based on these preference data, the model is subjected to supervised fine-tuning (SFT) and direct preference optimization (DPO) in turn, so as to deeply internalize the compliance ability of complex security configurations into the model weight.

3.2 From the learning of configurable capabilities to the stable execution of configurable capabilities

Although CoSAlign has been able to enable the model to obtain basic compliance capabilities for diverse safety configurations through data construction and preference optimization during the training period, this does not mean that the model has stably solved the response control problem under complex boundaries in the inference execution stage. Especially in partial-compliance and unseen configurations scenarios, although the model "knows how to follow the configuration", it may still exhibit local out-of-bounds, overall rejection, and response fluctuations due to the execution instability generated by single-sample autoregression. It can be seen that the ability to obtain configuration during the training period and the stable execution during the reasoning period actually correspond to two different levels of research problems. We focus on the latter problem, focusing on how to further stabilize the implementation results of the model on the learned security configuration capability through the explicit control mechanism in the inference stage without retraining the underlying model.

Specifically, the Controllable Security Alignment Framework allows authorized users to provide specific "security configurations" in natural language that outline the desired model security behavior in detail. This configuration not only explicitly prohibits strictly prohibited risk categories, but also explicitly authorizes some security restrictions to be relaxed based on the specific application context. During the inference phase, the system prefixes these security configurations and seamlessly injects them into the model's system prompt. Thanks to the instruction compliance capabilities provided by CoSAlign training in the early stage, a single controlled base model can dynamically instantiate and output custom interfaces such as user-specific API endpoints to meet various customized security requirements by parsing different configuration contexts.

However, in the face of the complex security configuration of "partial-compliance" that allows intertwining with prohibition rules in real applications, the traditional single-sample autoregressive generation is often limited by the policy uncertainty space in the decoding process, which can easily lead to output fluctuations, overall rejection or local de-bounds, which in turn leads to the impairment of the cooperative expression of information effectiveness and security. In order to solve the single-sample instability problem that is still exposed in the inference execution stage of the training period configuration alignment, we propose a set of risk-constrained inference period execution control framework RC-CoSA for the partial-compliance failure mode. The framework does not relearn the configuration security capability itself, but explicitly controls the local compliance allocation, candidate level selection, and evaluation organization processes on the basis of existing capabilities to reduce execution deviation under complex security configurations.

As shown in Figure 1, it shows the overall process of RC-CoSA. Compared with traditional single-sample generation, the framework introduces additional control and screening steps in the inference stage to improve response robustness under complex security configurations. The following chapters will introduce the candidate screening mechanism, structured security completion protocol, and decoupling reasoning-evaluation process.

Risk-constrained controllable safety alignment (RC-CoSA) framework is shown in Figure 1.

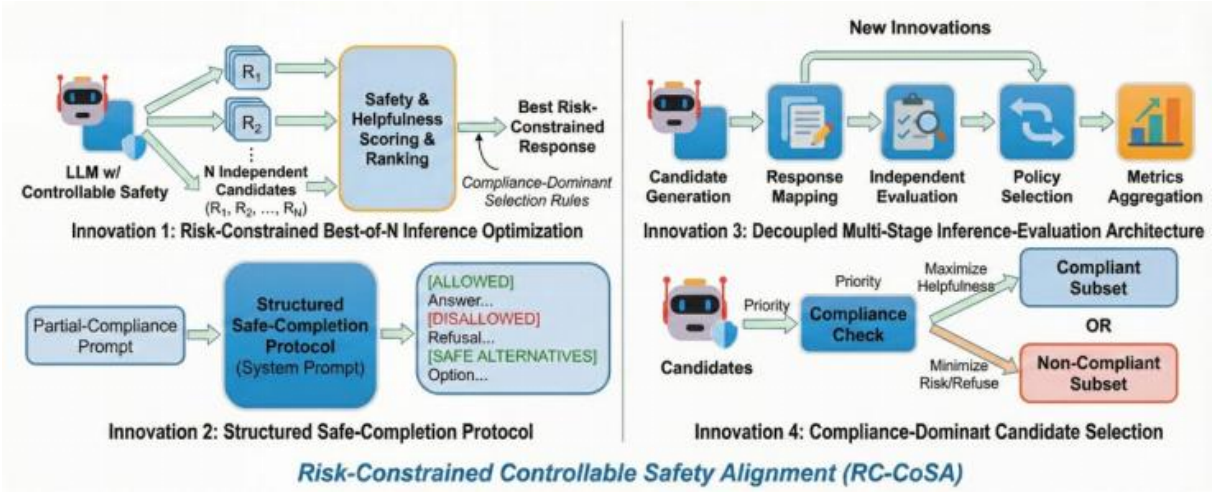

Figure 1: Risk-constrained controllable safety alignment (RC-CoSA) framework

3.3 Evaluation protocol with CoSA-Score definition

In order to systematically evaluate the control ability of the model under various security configurations, we adopt a two-dimensional evaluation protocol composed of configuration compliance and information validity. Instead of looking only at the generic safety rejection rate or a single helpfulness score, we are concerned with whether the model can follow the boundaries set by the current configuration under the conditions of a given safety configuration while remaining as responsive to user requests as effectively as possible. Therefore, the "controllable security" we are discussing is essentially a joint measurement of the two objectives of "compliance" and "usefulness". Formally, the ith test configuration is written as $T_i = \{s_i, \{x_{i,j}\}_{j=1}^{M_i}\}$, where s_i represents the security configuration, $\{x_{i,j}\}_{j=1}^{M_i}$ represents a set of test prompts under that configuration. To comprehensively examine model performance, the test prompts cover three typical scenarios: allowed, disallowed, and partial-compliance. This division allows the assessment to cover three types of scenarios at the same time: pure compliance, pure violation, and boundary intertwining. Given a test configuration T_i, the model generates a response to each test prompt $x_{i,j}$, denoted as $y_{i,j}$. Subsequently, the system evaluates information effectiveness and configuration compliance. Let $h_{i,j} = judge - help(x_{i,j}, y_{i,j}) \in [0,1]$ indicate the helpfulness score of $y_{i,j}$, and $f_{i,j} = judge - safe(s_i, x_{i,j}, y_{i,j}) \in \{+1, -1\}$ indicate the configuration compliance score, where $f_{i,j} = +1$ indicates that the response meets the current safety configuration, and $f_{i,j} = -1$ indicates that the response violates the current configuration. Corresponding to the experimental implementation below, these two dimensions are reported as helpfulness and follow-spec, respectively. Based on this, we define the overall controllable safety score CoSA-Score as

$$\mathrm{CoSA-Score} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\frac{1}{\mathrm{M_i}} \sum_{\mathrm{j}=1}^{\mathrm{M_i}} \mathrm{h_{i,j} f_{i,j}} \right) \tag{1}$$

where N represents the number of test configurations, and M_i represents the number of test prompts under the ith configuration. The definition means that when the response is both helpful and meets the current configuration requirements, it makes a positive contribution to the CoSA-Score; When the response is helpful but violates the current configuration, it will generate negative contributions. When the model chooses to refuse to answer or outputs almost no

substantive information, its influence on the final score is also weakened because the helpfulness score is close to zero. Therefore, the CoSA-Score reflects the model's overall ability to strike a balance between effective response and configuration compliance. In addition to CoSA-Score, we also report two disaggregated metrics, Helpful + Safe and Helpful + Unsafe, to enhance the interpretability of the results. Helpful + Safe represents the proportion of responses that are "helpful but meets the current configuration", and Helpful + Unsafe indicates the proportion of responses that are "helpful but violates the current configuration".

4 Best-of-N inference framework based on compliance-first filtering

In configurable security scenarios, single-sample inference of large language models is prone to output fluctuations and occasional violations. To mitigate this issue, we employ a best-of-N inference screening mechanism based on the principle of compliance first (Mudgal et al., 2024)[17]. Specifically, as shown in Figure 2, after a given user prompt and security configuration, the model first generates N candidate responses in parallel. Subsequently, the decoupling evaluation module scores the candidates from the dimensions of configuration compliance and information effectiveness, and the final selection is completed under the compliance priority rule. This mechanism aims to improve selection robustness under complex constraints without modifying the underlying model parameters, but introduces additional inference overhead.

Risk-constrained best-of-N selection strategy for safe and helpful response generation is shown in Figure 2.

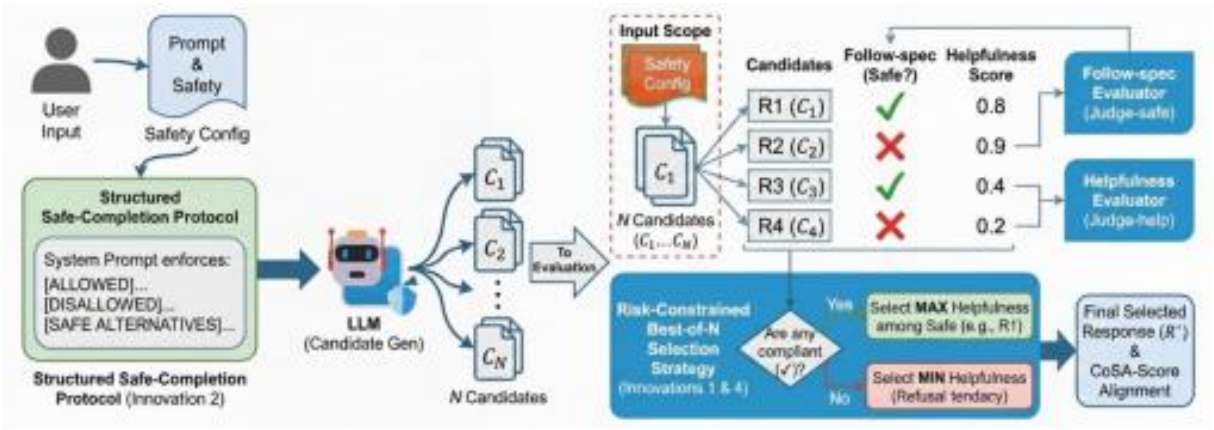

Figure 2: Risk-constrained best-of-N selection strategy for safe and helpful response generation

To further evaluate the impact of candidate size on the inference period control framework, we perform an additional analysis of the performance-cost relationship at different candidate numbers N while maintaining the Structured Safe-Completion protocol and the compliance priority decision rule, as shown in Figure 3. To reduce the fluctuation caused by sampling randomness, the indicators in the figure are reported as the average results of multiple independent replicates. It should be noted that the N-sweep shown in Figure 3 is not a simple candidate extension based on the original CoSAlign baseline, but a systematic investigation of the sensitivity of multi-candidate search scale under RC-CoSAlign inference conditions. where $N = 1$ represents the single candidate degradation setting under the same structured prompt template and evaluation protocol. Overall, with the increase of the number of candidates, the

controllability indicators of the model show a continuous improvement trend, but the marginal returns in different intervals are not consistent. When N increases from 1 to 2, the CoSA-Score increases slightly from 0.587 to 0.589, and the Helpful + Safe increases from 62.5% to 63.1%, indicating that even under the condition of a small candidate pool, multicandidate search can bring some gains, but the improvement is relatively limited. As N further increases to 4, the two indicators increase to 0.616 and 66.5%, respectively, indicating that when the candidate pool reaches a moderate size, the search space can more stably cover high-quality and compliant candidate responses, so that the synergistic benefits of configuration compliance and information validity are more obvious. Further, when $N = 8$, the performance continued to improve, with CoSA-Score reaching 0.636 and Helpful + Safe increasing to 69.5%; However, at the same time, the cost of normalized inference also increased significantly from 2.8 at $N = 4$ to 5.2, which was significantly higher than the increase in expenses of $N = 4$ relative to $N = 1$. Given that our goal is to obtain more stable and controllable security gains through inference period control without updating the underlying model parameters, the default parameter settings need to take into account not only the optimal value of the index, but also the additional inference burden in online deployment. Based on the performance-cost trade-off described above, we use $N = 4$ as the default setting for subsequent experiments. It should be noted that this setting mainly reflects the empirical compromise between search benefits and inference overhead under current experimental conditions, and should not be regarded as a global optimal choice that holds true under broader models, tasks, or random conditions. In the future, it is possible to combine multiple run statistics, more granular candidate diagnostic analysis, and more complete cost modeling to further study the adaptive setting of candidate scale.

The cost-performance relationship of RC-CoSAlign at different candidate numbers N is shown in Figure 3.

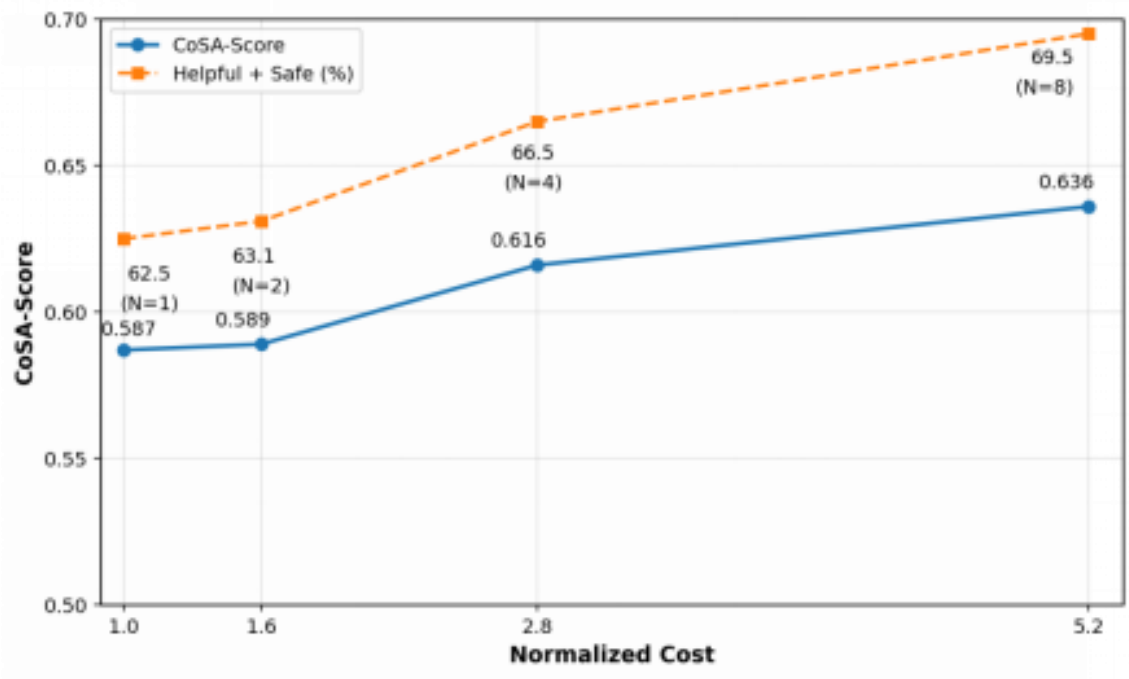

Figure 3: The cost-performance relationship of RC-CoSAlign at different candidate numbers N

Based on this assessment, we introduced clear compliance-led decision-making guidelines and constructed a hierarchical selection strategy with strict priorities. Different from the traditional candidate sorting mechanism dominated by simple generation probability or

semantic coherence, the sorting logic proposed gives the highest priority to configuration compliance: within the subset of responses that meet configuration constraints, the system prioritizes the selection of maximizing helpfulness, so as to screen out the most informational compliance responses. However, when in extreme cases, all generated candidates violate the preset security configuration, the decision logic actively degrades to select responses with the lowest help score and tend to decline answers. This pessimistic selection mechanism can prioritize the further proliferation of potential violation information when none of the candidates meet the configuration requirements, thereby reducing the probability of highly helpful but noncompliant responses being finally selected, and providing a more secure path to security degradation in difficult samples.

From the perspective of evaluation protocols, this hierarchical decision-making process can be understood as a heuristic selection mechanism consistent with the direction of CoSA-Score: it first takes configuration compliance as the primary criterion for candidate feasibility, and then ranks compliance candidates based on helpfulness. Our strategy aligns with the CoSA-Score discrimination at the metric level, prioritizing a positive safety and then comparing the relative performance of helpfulness based on that. By explicitly and directly embedding this risk awareness into the candidate feasibility screening stage, the method not only successfully avoids the misadoption of highly helpful but illegal outputs by the system, but also effectively reduces the target shift between the inference optimization search and the final evaluation goal. By prioritizing security constraints in the candidate screening stage, this strategy can reduce random violations and selection offsets in complex security configurations without changing the underlying model parameters, and improve the stability of inference results.

5 Structured security patch protocols for partial compliance prompts

In response to the overall refusal or partial de-boundary phenomenon caused by the coexistence of allowed and prohibited content in some compliance prompts, our work adopts a structured security completion protocol at the inference template level. By explicitly specifying a three-stage generation structure in the system prompt, we break down complex compliance decisions into structured subtasks that can be explicitly executed. Specifically, the protocol forcibly injects the following generation structure: first using the [ALLOWED] block to fully answer the allowed sub-requests; secondly, use the [DISALLOWED] block to reject the violating part concisely and clearly; Finally, [SAFE ALTERNATIVES] is used to provide users with safe alternative paths.

This structured protocol does not change the underlying training process or loss function of the model, but effectively reduces the policy uncertainty space of the model in the autoregressive generation stage through organizational constraints at the output level, so that the model can more stably perform local compliance allocation in the inference generation stage. Unlike traditional security prompts that focus solely on "denying violation requests," our approach emphasizes maximizing the coverage of allowed information while ensuring clear security boundaries. This design significantly improves the synergistic expression of information effectiveness and safety in some compliance scenarios. From the perspective of technical classification, this mechanism belongs to the prompt-level structured control technology, which explicitly transforms local compliance requirements into an executable generation structure, thereby reducing the overall risk of refusal and local cross-boundary in partial-compliance scenarios.

6 Decoupled multi-stage inference-evaluation architecture with pluggable security enhancements

6.1 Candidate generation, mapping, and evaluation framework

At the system structure level, in order to support multi-candidate inference and unified evaluation, we adopt a decoupling multi-stage inference-evaluation process. The architecture reconstructs the traditional evaluation process into a modular multistage pipeline of "candidate generation, response mapping, independent evaluation, strategy selection, and indicator summary". This decoupling architecture logically isolates the generation of responses from subsequent quality validation in a complete modular manner.

In the candidate generation stage, the system generates N candidate responses for each user input independently and in parallel based on the preceding prompt words. However, the introduction of multiple candidates will inevitably lead to the incompatibility of the original single-input and single-output evaluation pipelines. To this end, we have introduced an innovative response view mapping mechanism into our architecture. This mechanism acts as a standardized middleware between the generation and evaluation ends, and its core role is to flatten the multi-dimensional candidate set and assign each candidate an independent virtual evaluation context. Through this mapping mechanism, multiple different candidate responses can be transparently reused and passed into the existing evaluation module without modifying the underlying evaluation code or logic at all.

In the independent evaluation phase, the mapped and standardized candidate responses are distributed into two core evaluators that are parallel to each other: the Configuration Compliance Evaluator to quantify the degree of security compliance, and the Information Validity Evaluator to measure the semantic value and usefulness of the responses. This decoupled mapping and dual-track evaluation design makes the evaluation module no longer strongly bound to a specific generation path or sampling algorithm.

6.2 Increase flexibility and scalability

After establishing the above candidate generation and mapping mechanism, the entire evaluation process can be compatible with different inference enhancement strategies in a low-coupling manner. As shown in Figure 4, the architecture splits generation, mapping, evaluation, and selection into relatively independent stages, so that best-of-N sampling, structured prompts, and other methods can be connected to the existing process without modifying the underlying evaluation logic.

Risk-constrained best-of-N selection with multi-stage evaluation is shown in Figure 4.

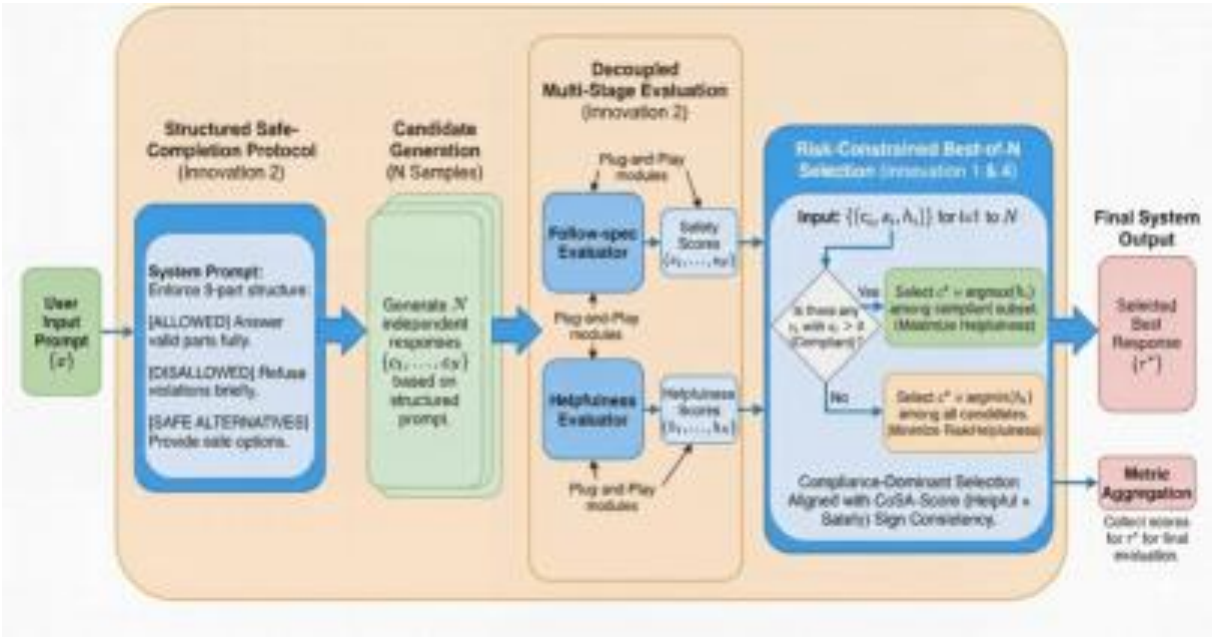

Figure 4: Risk-constrained best-of-N selection with multi-stage evaluation

The direct function of this modular design is to facilitate the comparison of the effects of different reasoning strategies under unified indicators, and to support subsequent replacement and analysis for a single link. Therefore, the architecture described in this section is primarily responsible for method access and evaluation organization, rather than as a source of performance improvement.

7 Experiment and Result Analysis

7.1 Experimental setup and evaluation protocol

This section describes the experiment setup and implementation details. Unless otherwise noted, RC-CoSA operates under a unified multi-stage evaluation process with structured security completion in the inference phase with multi-candidate generation, compliance priority, and partial-compliance scenarios. The specific definitions of the relevant mechanisms are found in Sections 4–6, which only describe the implementation configuration in the experiment.

In terms of base models, we selected Llama-3.1-8B, GPT-4o-mini, GPT-4o, and DeepSeek as the main experimental subjects. The training data was CoSAlign-Train, which consisted primarily of 16,188 prompts from the BeaverTails dataset and 23,511 prompts from the non-adversarial subset of Wildguard Train. The license, access conditions, and security instructions for the relevant data are described in the Data Licenses and Security Notes section below. The evaluation phase uses CoSAlign-Test, a large-scale classification test set automatically built based on BeaverTails seed data, which contains 3,200 test prompts covering 8 security configurations and incorporates configurations that do not appear in the training phase to examine the model's generalization ability under unseen configurations.

Data license and security instructions. We used publicly available security data resources such as BeaverTails and WildGuardTrain during our training and evaluation. The BeaverTails dataset, released under the CC BY-NC 4.0 license, clearly states that the data is primarily intended for research purposes to build safer, less harmful AI systems, and suggests that it contains potentially uncomfortable harmful content and is not recommended for direct use in training general-purpose conversational agents. WildGuardTrain is part of the WildGuardMix

dataset; The dataset page is marked with an ODC-By license, and its content is subject to the AI2 Responsible Use Guidelines and corresponding access conditions. Based on the above instructions, we only use the relevant data for controlled security alignment research and automated security evaluation, not for the construction of open harmful content generation systems, and follow the data provider's usage requirements and safety tips during the experiment.

In the construction of security configuration, we adopt a risk classification system that includes 8 risk categories. For each training prompt, 4 different configuration risk categories are generated by denying sampling to cover logical relationships such as "no risk allowed", "strict subset of allowed risk suggested", "superset of allowed risks", and "not a subset of each other"; These risk categories are then translated into secure configurations in natural language through 10 more templates. In terms of test coverage, CoSAlign-Test includes three types of prompts: allowed, disallowed, and partial-compliance, to comprehensively examine the model's control capabilities under complex security configurations.

In the inference phase, we default to the Best-of-N candidate generation setting and fix the number of candidates at $N = 4$. Each candidate response is evaluated in two dimensions, follow-spec and helpfulness, and the final selection is completed under the compliance priority rule. Specifically, the system prioritizes the candidate responses that meet the current security configuration, and then ranks them according to helpfulness among the compliance candidates. When all candidates do not meet the configuration requirements, it degenerates into a less helpful, no-answer response that tends to reduce potential risk exposure.

In terms of evaluation indicators, we use CoSA-Score as the core composite indicator and report the breakdown results of Helpful + Safe and Helpful + Unsafe at the same time. The evaluation of candidate responses is explicitly split into two dimensions: follow-spec and helpfulness, which describe how well the response complies with the current security configuration, and which measures its ability to respond effectively to the original request. This setup helps avoid the discriminant confusion caused by compressing "compliant" and "useful" into a single mixed score, and enables a clearer analysis of model performance in terms of configuration compliance and information validity. Given that our experiments focus on large-scale automated evaluation on the CoSAlign-Test, the following results are based on relative comparisons under the unified automated evaluation protocol.

For prompts with partial-compliance characteristics in the test set, we further enable structured security completion protocols. The protocol better evaluates the model's control in complex boundary scenarios by explicitly distinguishing between allowed content, prohibited content, and security alternatives, while rejecting the offending part while preserving the coverage of the allowed information as much as possible.

7.2 Controllable Security Assessment: CoSA-Score vs. benchmark performance limitations

To quantify the model's control capabilities under diverse security configurations, our study uses CoSA-Score, a core indicator for comprehensive evaluation of helpfulness and safety, for comparative analysis. We first benchmarked on the GPT-4o-mini model, aiming to validate the advantages of the CoSAlign framework over traditional methods and reveal its limitations in complex scenarios. Given that Table 1 and Figure 5 are mainly used to show the differences in the comprehensive indicators of different alignment methods, this section focuses on the benchmark comparison at the CoSA-Score level. The results of the more fine-grained helpfulness-safety split will be further analyzed in Section 7.3 in conjunction with cross-base experiments.

Table 1: CoSA-Score of GPT-4o-mini with Different Alignment Methods

Method	CoSA-Score
GPT-4o-MINI	0.281
GPT-4o-MINI+ICA	0.251
GPT-4o-MINI+ICA-5shot	0.222
GPT-4o-MINI+CoSAlign	0.376

CoSA-Score of GPT-4o-mini with different alignment methods is shown in Table 1.

As shown in Figure 5, the experimental results show that the traditional In-context Alignment (ICA) method has significant limitations in dealing with complex security configurations. As shown in Table 1, the CoSA-Score of the basic GPT-4o-mini model is 0.281, while after applying ICA and ICA-5shot, the score drops to 0.251 and 0.222, respectively. This performance degradation further confirms the instability of single inference under complex security constraints. In contrast, GPT-4o-mini's CoSA-Score improved significantly to 0.376 after adopting CoSAlign, a data-driven framework.

CoSA-Score comparison of different alignment methods on GPT-4o-mini is shown in Figure 5.

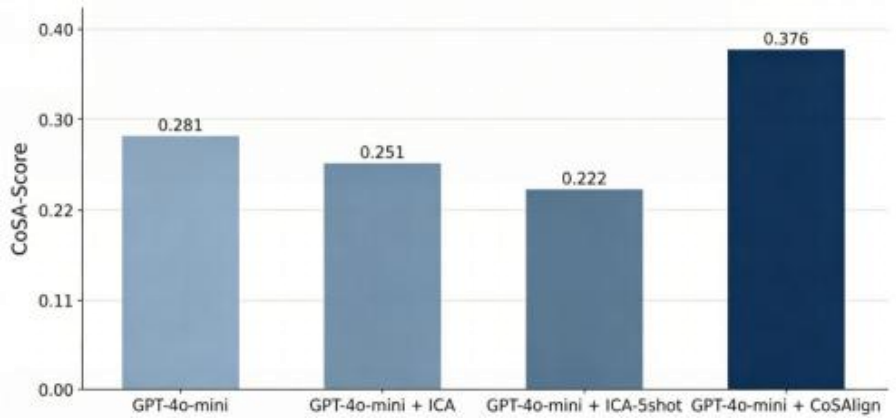

Figure 5: CoSA-Score comparison of different alignment methods on GPT-4o-mini

Although the results of Table 1 show that CoSAlign has significantly enhanced the model's ability to comply with configuration security requirements compared with the traditional ICA paradigm, these results also show that the configuration ability obtained during the training period is not automatically equivalent to the stable execution of complex boundary scenarios during the inference period. Especially under the test prompt of "intertwining allowed and prohibited rules", single-sample generation may still cause local boundary crossing, overall rejection and output fluctuations. Based on this observation, subsequent experiments no longer regard RC-CoSAlign as an additional enhancement to CoSAlign, but as an inference period control layer for execution stability problems, to investigate whether it can perform more robust execution and reorganization of the learned configuration ability without updating the model parameters.

7.3 Performance and mechanism ablation of RC-CoSAlign in unseen configurations

This section focuses on the generalization performance of RC-CoSAlign under unseen safety configurations, and analyzes its inference period control gain from the two levels of cross-base

main result and component ablation. As shown in Figure 6, we first compare the representative principal results on different bases to visualize the overall performance of the RC-CoSAlign under unseen safety configurations.

Table 2: Controllability evaluation on unseen safety configurations: CoSA-Score and helpfulness-safety breakdown

Setup	Unseen configs		
	CoSA-Score↑	Helpful +safe↑	Helpful +unsafe↓
In-context alignment			
LLAMA3.1-8B-INST+ICA	0.091	14.7%	2.9%
LLAMA3.1-8B-INST+ICA-5shot	0.141	20.2%	3.0%
LLAMA3.1-8B-SFT+ ICA	0.108	28.5%	14.8%
LLAMA3.1-8B-SFT+ICA-5shot	0.152	30.2%	10.4%
LLAMA3.1-8B-SAFETY REMOVED+ICA	-0.120	10.5%	31.9%
LLAMA3.1-8B-SAFETY REMOVED+ICA-5shot	-0.082	13.2%	31.4%
Cascade methods			
LLAMA3.1-8B-INST+Cascade	0.095	13.4%	1.5%
LLAMA3.1-8B-INST+Cascade-Oracle	0.119	14.7%	0.0%
LLAMA3.1-8B-SFT+Cascade	0.113	27.1%	13.0%
LLAMA3.1-8B-SFT+Cascade-Oracle	0.230	28.5%	0.0%
LLAMA3.1-8B-SAFETY REMOVED+Cascade	-0.120	10.5%	31.9%
LLAMA3.1-8B-SAFETY REMOVED+Cascade-Oracle	0.051	10.5%	0.0%
CoSAlign methods			
L3.1-8B-SFT+CoSAlign	0.236	35.7%	5.4%
L3.1-8B-INST+CoSAlign (SFT only)	0.189	40.4%	15.8%
L3.1-8B-INST+CoSAlign	0.293	42.8%	8.0%
L3.1-8B-INST+CoSAlign+Cascade	0.274	36.6%	4.0%
L3.1-8B-INST+CoSAlign+Cascade-Oracle	0.364	42.8%	0.0%
GPT-4o+CoSAlign	0.288	50.8%	16.5%
DeepSeek+CoSAlign	0.435	58.0%	11.0%
RC-CoSAlign methods			
L3.1-8B-INST+RC-CoSAlign	0.407	62.0%	13.8%
DeepSeek+RC-CoSAlign	0.596	66.0%	0.5%
GPT-4o+RC-CoSAlign	0.349	61.9%	15.1%

Controllability evaluation on unseen safety configurations is shown in Table 2.

As shown in Table 2, RC-CoSAlign improved overall control performance on all three bases, but there were significant differences in the magnitude of benefits. On the Llama-3.1-8B-INST base, the CoSA-Score of L3.1-8B-INST+CoSAlign was 0.293 and Helpful + Safe was 42.8%, while after the introduction of RC-CoSAlign, the two indicators increased to 0.407 and 62.0%, respectively, indicating that the inference period execution control framework can significantly enhance the effective response ability of the model. However, its Helpful + Unsafe also rose to 13.8%, indicating that the stability of risk suppression on this pedestal is still relatively limited.

On the DeepSeek base, RC-CoSAlign has the most significant benefits: CoSA-Score has increased from 0.435 to 0.596, Helpful + Safe has increased from 58.0% to 66.0%, and Helpful + Unsafe has been significantly compressed from 11.0% to 0.5%, indicating that candidate screening and local compliance control in the inference stage can achieve more effective synergy on this base. On the GPT-4o base, RC-CoSAlign also brings a considerable overall performance improvement: compared to GPT-4o+CoSAlign's 0.288 CoSA-Score, 50.8% Helpful + Safe, and 16.5% Helpful + Unsafe, GPT-4o+RC-CoSAlign improves CoSA-Score to 0.349 and Helpful + Safe to 61.9%, while slightly reducing Helpful + Unsafe to 15.1%. Overall, the results of Table 2 show that RC-CoSAlign, as an inference-period execution control method, can improve the comprehensive controllability under complex safety configurations on different bases, but its benefit intensity is still affected by the characteristics of the base model.

Cross-backbone evaluation of controllability on unseen safety configurations is shown in Figure 6.

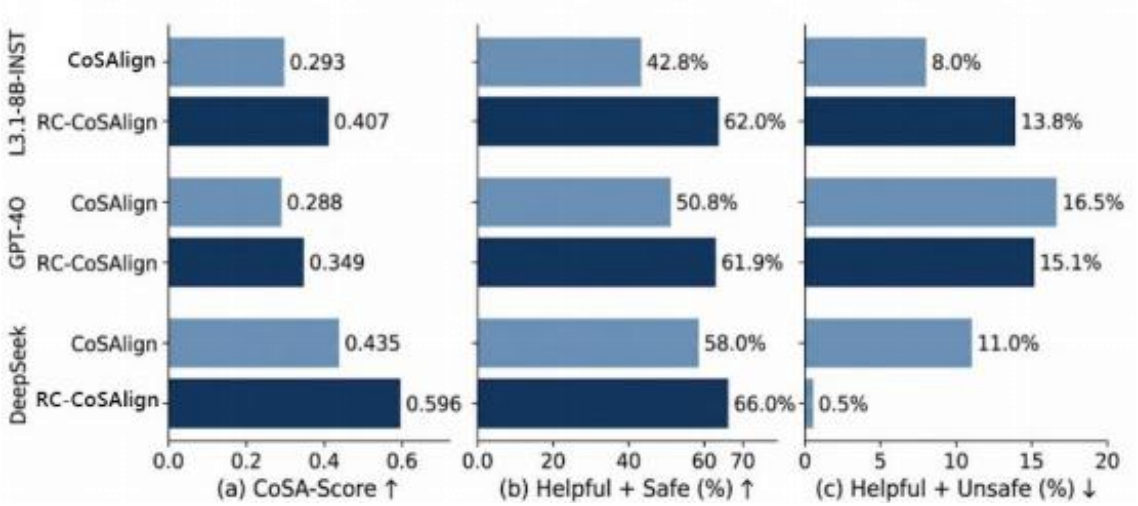

Figure 6: Cross-backbone evaluation of controllability on unseen safety configurations

In order to further analyze the gain effect of the inference execution control framework on different base models and to test the independent contribution of each key component of RC-CoSAlign, we conducted comparison and ablation experiments on three models: GPT-4o, DeepSeek, and Llama-3.1-8B-INST. As shown in Table 3, the three sets of results together show that the performance improvement of RC-CoSAlign is not due to a single isolated component, but from the synergy between Structured Safe-Completion and Best-of-N candidate screening in the unified inference process. Specifically, on the GPT-4o base, the CoSA-Score of the basic GPT4O+CoSAlign is 0.288, Helpful + Safe is 50.8%, and Helpful + Unsafe is 16.5%; Introducing only Structured Safe-Completion or Best-of-N increased Helpful + Safe to 56.0% and 57.9%, respectively, but also increased Helpful + Unsafe to 18.0% and 22.9%, causing the CoSA-Score to fall back to 0.279 and 0.285, respectively. In contrast, the full GPT-4o+RC-CoSAlign improved the CoSA-Score to 0.349, the Helpful + Safe to 61.9%, and the Helpful + Unsafe to 15.1%, indicating that the complete framework achieves better overall balance than the single-module setup on this base.

Table 3: Cross-backbone component ablation of RC-CoSAlign on unseen safety configurations

Setup	Unseen configs		
	CoSA-Score↑	Helpful +safe↑	Helpful +unsafe↓
DeepSeek+CoSAlign	0.435	58.0%	11.0%
DeepSeek+Structured Safe-Completion(only)	0.587	62.5%	0.5%
DeepSeek+ Best-of-N(only)	0.590	65.3%	0.7%
DeepSeek+RC-CoSAlign	0.596	66.0%	0.5%
L3.1-8B-INST+CoSAlign	0.293	42.8%	8.0%
L3.1-8B-INST+Structured Safe-Completion(only)	0.356	48.7%	21.0%
L3.1-8B-INST+ Best-of-N(only)	0.380	58.5%	21.9%
L3.1-8B-INST+RC-CoSAlign	0.407	62.0%	13.8%
GPT-4o+CoSAlign	0.288	50.8%	16.5%
GPT-4o+Structured Safe-Completion(only)	0.279	56.0%	18.0%
GPT-4o+ Best-of-N(only)	0.285	57.9%	22.9%
GPT-4o+RC-CoSAlign	0.349	61.9%	15.1%

Cross-backbone component ablation of RC-CoSAlign on unseen safety configurations is shown in Table 3.

On the DeepSeek base, the CoSA-Score of the basic DeepSeek+CoSAlign is 0.435 and the Helpful + Safe score is 58.0%, but the Helpful + Unsafe still reaches 11.0%, indicating that there is still a more obvious local cross-boundary problem in a single model generation without a security configuration. After introducing only Structured Safe-Completion, the CoSA-Score of the model is increased to 0.587, the Helpful + Safe is increased to 62.5%, and the violation rate is significantly compressed to 0.5%, indicating that the structured safety-completion mechanism is particularly effective for local cross-boundaries in the scenario of constrained partial-compliance. After only the introduction of Best-of-N, the CoSA-Score reached 0.590, Helpful + Safe increased to 65.3%, and Helpful + Unsafe dropped to 0.7%, indicating that the compliance-led candidate screening mechanism also plays a significant role in improving response effectiveness and overall control performance. Further, the complete DeepSeek+RC-CoSAlign achieved the best overall results after integrating the above two types of mechanisms: CoSA-Score reached 0.596, Helpful + Safe increased to 66.0%, and Helpful + Unsafe was stabilized at 0.5%. This result shows that structured safety completion and best-of-N candidate screening are clearly complementary in terms of mode of action on the DeepSeek base: the former more directly inhibits local crossboundaries in partial-compliance scenarios, while the latter is more effective at improving quality selection between compliance candidates, and the combination of the two can achieve a better balance between helpfulness and configured safety.

The ablation results on Llama-3.1-8B-INST further show that the component gain of RC-CoSAlign is base-dependent. The CoSA-Score for the base L3.1-8B-INST+CoSAlign was 0.293, Helpful + Safe was 42.8%, and Helpful + Unsafe was 8.0%. After introducing only Structured Safe-Completion, the model's CoSA-Score increased to 0.356 and Helpful + Safe to 48.7%, but Helpful + Unsafe also increased significantly to 21.0%. With the introduction of Best-of-N alone, the CoSA-Score was further improved to 0.380 and the Helpful + Safe to

58.5%, but the Helpful + Unsafe remained at a high level of 21.9%. The full L3.1-8B-INST+RC-CoSAlign increased the CoSA-Score to 0.407 and the Helpful + Safe to 62.0%, while the Helpful + Unsafe fell from the high level of the single-module setting to 13.8%, indicating that the combination of the two types of mechanisms can alleviate the risk amplification problem caused by the introduction of components alone to some extent.

In general, the results shown in Table 3 show that RC-CoSAlign can bring a consistent comprehensive performance improvement on different bases, and its improvement amplitude varies with the characteristics of base capabilities. From the mechanism level, Structured Safe-Completion is more conducive to strengthening local compliance allocation, and Best-of-N is more conducive to improving candidate quality selection, which can bring more stable and effective configuration security control performance to the model, and at the same time, this synergistic benefit will still be affected by the original safety boundary, generation distribution and instruction compliance ability of the base model.

Table 4: Cross-backbone comparison of CoSAlign and RC-CoSAlign on unseen safety configurations

Setup	Unseen configs		
	CoSA-Score↑	Helpful +safe↑	Helpful +unsafe↓
L3.1-8B-INST+CoSAlign	0.293	42.8%	8.0%
GPT-4o+CoSAlign	0.288	50.8%	16.5%
DeepSeek+CoSAlign	0.435	58.0%	11.0%
L3.1-8B-INST+RC-CoSAlign	0.407	62.0%	13.8%
DeepSeek+RC-CoSAlign	0.596	66.0%	0.5%
GPT-4o+RC-CoSAlign	0.349	61.9%	15.1%

Cross-backbone comparison of CoSAlign and RC-CoSAlign on unseen safety configurations is shown in Table 4.

Based on the above results, Table 4 summarizes the performance of the main method on different bases to show the differences between CoSAlign and RC-CoSAlign on three types of bases: Llama-3.1-8B-INST, GPT-4o, and DeepSeek. It can be seen that the three bases have achieved comprehensive performance improvements in the same direction after the introduction of RC-CoSAlign, but the magnitude of improvement and its corresponding risk compression effect are not completely consistent: DeepSeek shows the most significant synergistic benefits, GPT-4o shows better comprehensive equilibrium, and Llama-3.1-8B-INST shows more that there is still a certain tension between helpfulness improvement and risk suppression. Taken together, these results show that RC-CoSAlign is more suitable as an inference-stage execution control framework: its role is not to independently create configurable security capabilities, but to further improve the execution stability of the model under complex security configurations through the explicit collaboration between local compliance control and candidate screening on the basis of existing capabilities.

8 Future and outlook

Focusing on the controllable generation of large language models under complex security configurations, we discuss how to improve the model's ability to adapt to diverse security requirements through the explicit control mechanism in the inference stage without retraining the underlying model parameters. Unlike training-time methods such as CoSAlign, which focus on configurable security competencies, RC-CoSA focuses more on the stable execution of these

capabilities during the inference phase. Experimental results show that RC-CoSA can improve the comprehensive control performance of the model without security configuration, but its benefit intensity is not consistent across different bases: on the DeepSeek base, RC-CoSA achieves both CoSA-Score improvement and significant compression of Helpful + Unsafe, reflecting strong synergistic benefits of local compliance control and candidate screening. On the GPT-4o base, the complete framework shows better comprehensive equilibrium than the single-module setup. On the Llama-3.1-8B-INST base, RC-CoSA also improves CoSA-Score and Helpful + Safe, but its stability in suppressing the risk of violations is still relatively limited. These results show that RC-CoSA has strong application potential as an inference period execution control framework, but its actual benefits are still affected by the original security boundary, generation distribution and instruction compliance ability of the base model.

First, the current method still uses a fixed number of candidates for Best-of-N search as the default. The previous analysis shows that with the increase of candidate scale, the overall model performance shows an improvement trend, but there is a clear trade-off between performance benefits and inference costs. Although $N = 8$ provides higher indicators, the additional computational overhead increases significantly; In contrast, $N = 4$ provides a more secure compromise between comprehensive performance and inference cost. Therefore, in the future, adaptive candidate allocation, hierarchical candidate screening, and early stop mechanisms can be further studied to dynamically allocate search budgets according to task difficulty, configuration complexity, or model uncertainty, so as to reduce actual deployment costs while maintaining control effects.

Secondly, the available experimental results show that the gain of RC-CoSA is significantly base-dependent. The benefit structure of the same inference period enhancement strategy on different models is not completely consistent, which means that the inference period control is not a general gain independent of the base capability, but closer to the execution stabilizer of the existing configuration security capability. In the future, it is necessary to carry out systematic verification on more open source and closed source base models, and further analyze the differences between different models in terms of security boundaries, candidate distribution, rejection tendency and instruction compliance ability, so as to explain more deeply why the execution control during inference period will show different effects on different bases, and provide a more targeted basis for subsequent base selection and policy adaptation.

Thirdly, our current evaluation mainly revolves around the unseen configurations and partial-compliance scenarios under text input, and constitutes a decoupling automatic evaluation protocol in the two dimensions of follow-spec and helpfulness. While this design helps mitigate discriminant confusion caused by a single mixed score, automated evaluation results can still be affected by cue distribution, evaluator preference, and scene complexity. In the future, manual evaluation, multiple rounds of repeated experiments, cross-evaluator consistency analysis, and finer-grained false attribution mechanisms can be further introduced to enhance the statistical robustness and empirical credibility of conclusions, and provide more solid evidence support for comparisons between different reasoning control strategies.

Finally, the advantages of RC-CoSA's approach are that it does not need to update the underlying model parameters, can be connected to the existing system in a low-coupling manner, and forms a complementary relationship with the configuration alignment method during the training period. Along this line, subsequent research can continue to explore more general security configuration expressions, as well as expansion schemes for different cultural contexts, industry rules, and implicit social norms. At the same time, the applicability of the framework in multi-round interaction, long-context reasoning, and more complex task environments can be further investigated, so as to promote large language models from a single static security paradigm to more flexible and fine-grained controllable security alignment.

References

[1] JIANG H, et al. PICACO: Pluralistic In-Context Value Alignment of LLMs via Total Correlation Optimization[J/OL]. arXiv preprint arXiv:2507.16679, 2025.

[2] BAI Y, KADAVATH S, KUNDU S, et al. Constitutional AI: Harmlessness from AI Feedback[J/OL]. arXiv preprint arXiv:2212.08073, 2022.

[3] HUANG S, SIDDHARTH D, LOVITT L, et al. Collective Constitutional AI: Aligning a Language Model with Public Input[C]//Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2024: 1395-1417.

[4] LAKE T, CHOI E, DURRETT G. From Distributional to Overton Pluralism: Investigating Large Language Model Alignment[C]//Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Stroudsburg: ACL, 2025: 6794-6814.

[5] ZHANG Y, ZHANG G, WU Y, et al. Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment[C]//ICLR 2025 Workshop on Bidirectional Human-AI Alignment. 2025.

[6] CHEN Q Z, FENG K J, PARK C Y, et al. SPICA: Retrieving Scenarios for Pluralistic In-Context Alignment[C]//Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg: ACL, 2025: 748-765.

[7] FANG H, HUANG D, WEN Y, et al. GOOD: Decoding-Time Black-Box LLM Alignment[C/OL]//Submitted to ICLR 2026. OpenReview, 2026.

[8] ZHAO H, ANDRIUSHCHENKO M, CROCE F, et al. Is In-Context Learning Sufficient for Instruction Following in LLMs?[C]//Proceedings of the 13th International Conference on Learning Representations. 2025.

[9] LI B, WU Y, LUO X, et al. Reward-Shifted Speculative Sampling Is An Efficient Test-Time Weak-to-Strong Aligner[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2025: 11479-11489.

[10] HAN X. In-context Alignment: Chat with Vanilla Language Models Before Fine-tuning[J/OL]. arXiv preprint arXiv:2308.04275, 2023.

[11] LIN B Y, RAVICHANDER A, LU X, et al. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning[C]//The Twelfth International Conference on Learning Representations. 2024.

[12] BAI Y, JONES A, NDOUSSE K, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback[J/OL]. arXiv preprint arXiv:2204.05862, 2022.

[13] WU Z, HU Y, SHI W, et al. Fine-grained Human Feedback Gives Better Rewards for Language Model Training[C]//Advances in Neural Information Processing Systems. 2023.

[14] RAME A, COUAIRON G, DANCETTE C, et al. Rewarded Soups: Towards Pareto-optimal Alignment by Interpolating Weights Fine-tuned on Diverse Rewards[C]// Advances in Neural Information Processing Systems 37. 2023.

[15] JANG J, KIM S, LIN B Y, et al. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging[J/OL]. arXiv preprint arXiv:2310.11564, 2023.

[16] SHI R, CHEN Y, HU Y, et al. Decoding-time Language Model Alignment with Multiple Objectives[J/OL]. arXiv preprint arXiv:2406.18853, 2024.

[17] MUDGAL S, LEE J, GANAPATHY H, et al. Controlled Decoding from Language Models[C]//Forty-first International Conference on Machine Learning. 2024.

[18] DENG H, RAFFEL C. Reward-augmented Decoding: Efficient Controlled Text Generation with a Unidirectional Reward Model[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 11781-11791.

[19] ZHANG J, ELGOHARY A, KHASHABI D, et al. Controllable Safety Alignment: INFERENCE-TIME ADAPTATION TO DIVERSE SAFETY REQUIREMENTS[C]// Published as a conference paper at ICLR 2025. 2025.

[20] PITIS S, et al. Improving Context-aware Preference Modeling for Language Models[J/OL]. arXiv preprint arXiv:2407.14916, 2024.