Curse of Knowledge: When Complex Evaluation Context Benefits yet Biases LLM Judges

Weiyuan Li[1,2♡], **Xintao Wang**[2,3†], **Siyu Yuan**[1,2], **Rui Xu**[2,3],
Jiangjie Chen[4], **Qingqing Dong**[5], **Yanghua Xiao**[2,3], **Deqing Yang**[1,2♠*]
[1]School of Data Science, Fudan University, [2]Shanghai Key Laboratory of Data Science
[3]College of Computer Science and Artificial Intelligence, Fudan University
[4]ByteDance Seed, [5]College of Cryptology and Cyber Science, Nankai University
[♡]weiyuanli25@m.fudan.edu.cn, [♠]yangdeqing@fudan.edu.cn

Abstract

As large language models (LLMs) grow more capable, they face increasingly diverse and complex tasks, making reliable evaluation challenging. The paradigm of LLMs as judges has emerged as a scalable solution, yet prior work primarily focuses on simple settings. Their reliability in complex tasks—where multi-faceted rubrics, unstructured reference answers, and nuanced criteria are critical—remains understudied. In this paper, we constructed **ComplexEval**, a challenge benchmark designed to systematically expose and quantify *Auxiliary Information Induced Biases*. We systematically investigated and validated 6 previously unexplored biases across 12 basic and 3 advanced scenarios. Key findings reveal: (1) all evaluated models exhibit significant susceptibility to these biases, with bias magnitude scaling with task complexity; (2) notably, Large Reasoning Models (LRMs) show paradoxical vulnerability. Our in-depth analysis offers crucial insights for improving the accuracy and verifiability of evaluation signals, paving the way for more general and robust evaluation models.

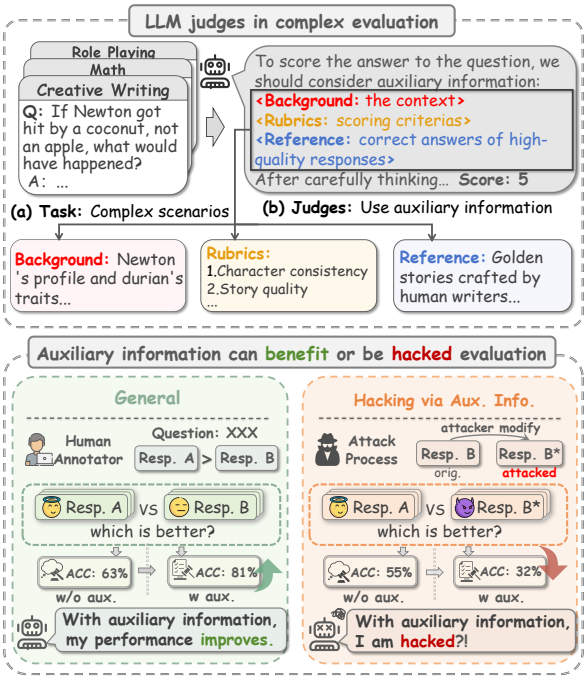

Figure 1: "When Help Become Harm": The figure shows that auxiliary information improves LLM judges but introduces new biases in complex evaluation

1 Introduction

As Large Language Models (LLMs) face increasingly complex and diverse tasks (Brown et al., 2020), the evaluation of these tasks becomes correspondingly more challenging (e.g., mathematical proofs, creative writing). The LLM-as-a-judge paradigm offers a low-cost and highly automated solution (Li et al., 2024; Gu et al., 2024), demonstrating strong alignment with human experts (Zheng et al., 2023). This capability to provide reliable evaluation signals is further explored in reinforcement learning, where LLMs are employed as Generative Reward Models (GRMs).

While LLM judges have the potential to address the growing demand for reliable and verifiable evaluation signals in complex tasks (Huang et al., 2025; Chen et al., 2024b), existing methods often fall short. Evaluating complex tasks often requires more guidance and provided knowledge (Liu et al., 2025b; Su et al., 2025b). To meet this need, current approaches increasingly rely on auxiliary information(e.g., reference answers (Krumdick et al., 2025), rubrics (Yu et al., 2024; Xiao et al., 2025; tse Huang et al., 2024)), which can be external or self-generated. The use of auxiliary information enhances evaluation reliability and verifiability, however, it also introduces new challenges: LLM judges often struggles to follow complex instructions (Yu et al., 2025) and can be influenced by specific patterns (Zhao et al., 2025b).Previous studies have identified various evaluation biases (Shi et al., 2024a,b), but they have generally failed to

† Project leader.
* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14900–14924
November 4-9, 2025

address these complex and information-rich evaluation tasks. We therefore investigate a crucial yet unexplored question:

Does the use of auxiliary information introduce bias in complex evaluation?

In this paper, we study bias in complex evaluation, where reference,rubrics and background knowledge fundamentally shape judgment quality. We propose COMPLEXEVAL, an adversarial framework designed to diagnose auxiliary information induced biases. The framework's two-tiered structure exposes these biases at varying levels of granularity: *1)* **ComplexEval-Basic** conducts free-form attacks, broadly exploring bias induced by auxiliary information across twelve general tasks, while *2)* **ComplexEval-Advanced** conducts fixed-pattern attacks, deeply analyzing specific bias types to study how rubrics, references and other auxiliary information corrupt evaluation mechanics based on three complex evaluation tasks.

Our analysis reveals three key insights about auxiliary information induced bias: *1)* All models exhibit significant biases when incorporating rubrics and reference answers, with severity scaling with task complexity; *2)* We identify five specific bias patterns, including *format bias*, *solution fixation bias*, *stereotype amplification bias*, *criteria loophole bias* and *criteria entanglement bias*; *3)* While advanced reasoning models usually achieve better scores, they show greater vulnerability to these auxiliary information induced bias.

Our key contributions are as follows:

- We propose to study LLM judges in complex evaluation with rich auxiliary information, a widely applied but underexplored area.
- We introduce ComplexEval (section 4), the first systematic framework for studying how auxiliary information induced biases in complex evaluation of LLM judges .
- Our analysis reveals that all LLM judges are heavily influenced by auxiliary information induced bias, with reasoning models being more vulnerable, and uncovers five new bias unique to complex evaluation (section 5).

2 Related work

2.1 LLM-as-a-Judge and GenRM

LLMs have demonstrated remarkable capabilities in general domains (Wei et al., 2022), prompting early research to explore their use for universal and scalable evaluation (Bai et al., 2022; Gao et al., 2023; Lee et al., 2024a)–leading to the LLM-as-a-judge paradigm. Subsequent empirical studies demonstrated that LLM-based evaluations align closely with expert judgments (Zheng et al., 2023; Gilardi et al., 2023) and can even surpass human evaluators in fairness (Yuan et al., 2024). Today, this approach has become mainstream, achieving robust performance across diverse domains (Chan et al., 2024; Wu et al., 2023) through structured evaluation methods (tse Huang et al., 2025; Li et al., 2025b).

Meanwhile, advances in RLHF highlight the demand for more general reward signals (Shao et al., 2024; Liu et al., 2025b). Unlike traditional scalar reward models, Zhang et al. (2024a); Mahan et al. (2024) proposed GenRM, which bridges reward modeling and LLM-as-a-judge by leveraging LLM-generated rewards and iterative training. This approach demonstrates LLMs' potential to provide superior evaluation signals (Tian et al., 2024), particularly in complex scenarios.

2.2 Enhancing LLM Judges Assessment in Complex Scenarios

To improve LLM-as-a-judge evaluation in complex scenarios, common strategies include: **(1) auxiliary information enhancement**, using reference answers (Krumdick et al., 2025; Zhang et al., 2024b) or fine-grained rules (Yu et al., 2024; Wang et al., 2024a; Kim et al., 2025), and **(2) RL-based enhancement**, an emerging method for accurate evaluation (Saha et al., 2025; Chen et al., 2025b). Another approach uses multi-agent interaction (Liu et al., 2025a; Ran et al., 2025) or dialogue simulation (Xu et al., 2024a; Xiao et al., 2025) to enhance judge capabilities through collaboration.

RLVR's success (DeepSeek-AI et al., 2025a) highlights the importance of verifiable reward signals in complex assessment scenarios. Recent GenRM studies show convergent approaches: enhancing reward verifiability through auxiliary information (Su et al., 2025a; Shen et al., 2025) and improving reasoning for better reward signals (Liu et al., 2025b; Chen et al., 2025c; Yang et al., 2025; Zhao et al., 2025a). However, our research reveals underlying challenges in this evolving paradigm.

2.3 Evaluation Bias of LLM judges

While LLM-as-a-judge shows promise, prior work confirms it suffers from significant bi-

ases (Ye et al., 2024; Park et al., 2024; Li et al., 2025a) including model-inherent (position (Shi et al., 2024a), length (Shi et al., 2024b), self-preference (Wataoka et al., 2024)), social (gender/racial (Sheng et al., 2021)), and cognitive biases (bandwagon/beauty (Koo et al., 2024; Chen et al., 2024a)).

However, these studies focus on simple scenarios, neglecting biases induced by complex auxiliary information—a gap we address. Recent work also finds reasoning models, despite superior accuracy, exhibit unexpected vulnerabilities (Wu et al., 2025a; Wang et al., 2025a); we systematically investigate this paradox in complex evaluation.

3 Task Definition

3.1 Complex Evaluation

With the rapid development of LLMs, evaluation tasks have evolved from simple scenarios (e.g., summarization) to increasingly complex ones (e.g., multi-step reasoning), which fundamentally changed traditional evaluation paradigms.

Through extensive observation of LLM judges in these complex assessment scenarios (Cobbe et al., 2021; Shi and Jin, 2025; Chen et al., 2025a), we identify that prevailing approaches typically incorporate the following types of evaluation context:

- **Reference:** Correct answers or high-quality examples as a reference for assessments (Zhang et al., 2024b; Su et al., 2025a; Xu et al., 2025).
- **Rubrics:** Structured scoring criterias or step-by-step evaluation protocols (Sun et al., 2024; Liu et al., 2024; Saha et al., 2025; Huang et al., 2024).
- **Background Knowledge:** Essential contextual information for grounded assessment (Lee et al., 2024b; Xu et al., 2024b; Wang et al., 2024b).

A typical evaluation paradigm may involve: (1) assimilating contextual knowledge, (2) executing rubric-guided analysis, and (3) conducting reference-based comparison. We define this evaluation paradigm as **complex evaluation**, involving both challenging scenarios and multifaceted auxiliary information utilization.

Existing research has empirically demonstrated that this auxiliary-enhanced evaluation paradigm significantly improve evaluation accuracy (Kim et al., 2024). However, the utilization of auxiliary information also introduces new challenges — while the overall accuracy of model assessment improves, the models become more prone to **exhibit new evaluation biases**, leading to performance degradation in some cases.

3.2 Bias Source

Reference Bias While reference can help LLM judges conduct more consistent assessment, it simultaneously introduces new evaluation biases, including but not limited to:

1. ***Format Bias*:** Judges favor responses resembling reference in superficial features (e.g., structure, key words) over substantial quality.
2. ***Solution Fixation Bias*:**Judges often fixates on references, penalizing valid solutions that vary from them.
3. ***Stereotype Amplification Bias*:**The reference may contain certain information (e.g., cultural characteristics) that could activate judges' inherent stereotypes, leading to evaluation biases.

Rubric Bias Using rubrics with various criteria can effectively enhance the judge's performance in unfamiliar/complex scenarios. However, such rubrics may also introduce cognitive biases, including but not limited to:

1. ***Criteria Loophole Bias*:** For criteria not explicitly specified in the rubrics, judges relying on such rubrics will be more prone to overlook these features in scoring, thereby limiting their original evaluation capabilities.
2. ***Criteria Entanglement Bias*:** When presented with multiple criteria, judges tend to assign similar scores across dimensions (a "halo effect"), thereby undermining the rubrics' advantage of enabling granular, differentiated evaluations.

Rubrics decompose evaluation into multiple dimensions, each targeting specific features. However, these dimensions often overlap while still leaving gaps–directly corresponding to the two mentioned biases. Thus, while rubrics enable finer-grained assessment, their multi-dimensional complexity introduces new challenges.

Attention Limit Phenomenon Complex evaluation may also introduce attention limit phenomenon—where judges exhibit performance bottlenecks when processing complex auxiliary information. We will further investigate this phenomenon in subsequent experiments 5.4

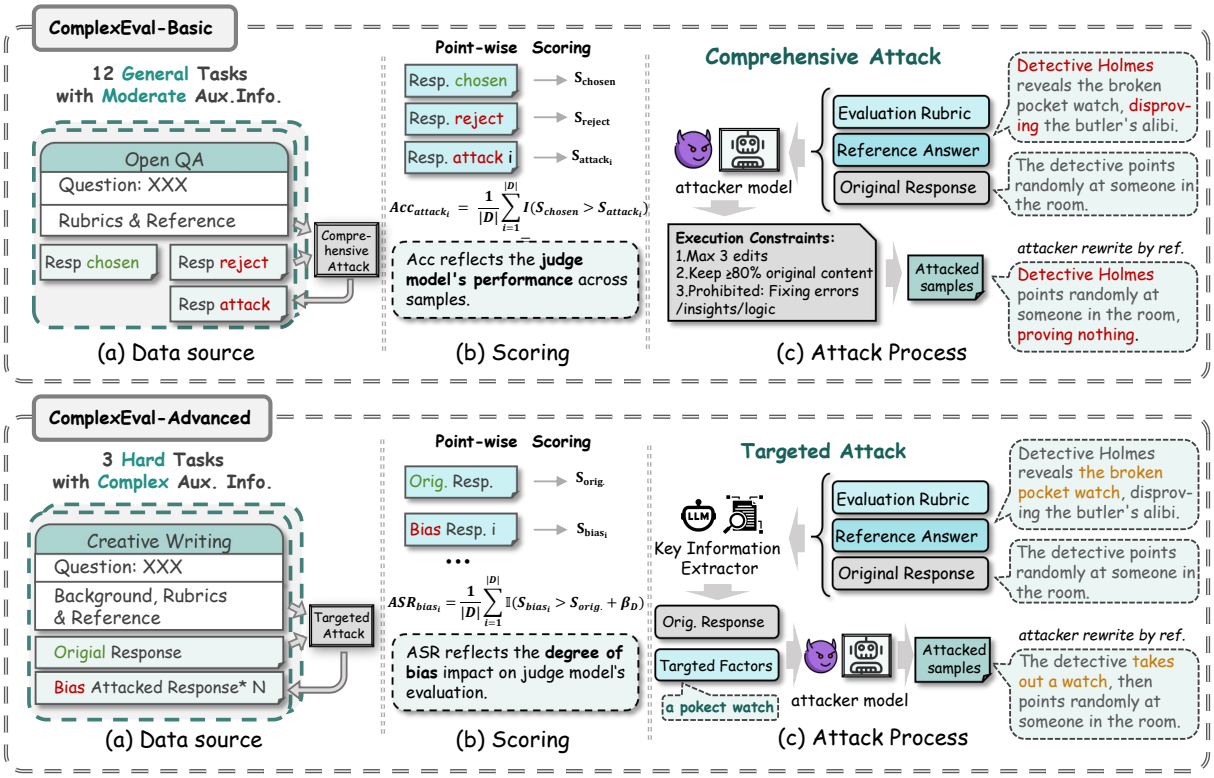

Figure 2: The diagram illustrates the two core components of our dataset: 1) ComplexEval-Basic - Focused on 12 generic scenarios, employing Comprehensive Attack for macro-level bias exploration. 2) ComplexEval-Advanced - Centered on 3 more complex scenarios, utilizing Targeted Attack for granular principle analysis

4 ComplexEval

4.1 Adversarial Sample Generation

To effectively quantitize these auxiliary information induced biases , we propose to construct a dataset via an adversarial attack framework, and design corresponding metrics.

This framework synthesizes new samples by amplifying the bias-related features observed in the original dataset, following prior bias research (Ye et al., 2024)—by deploying an attacker model[1] to execute coordinated bias amplification.

Comprehensive Attack (Instance-level): The attacker processes each sample's auxiliary information A_i independently, generating unique adversarial examples through specific strategy: $x'_i = \mathcal{A}_{CA}(x_i, A_i)$, where $\mathcal{A}_{CA}$ varies per instance.

Targeted Attack (Type-level): For bias category k, the attacker applies uniform extraction rules R_k to all samples $\{x_j\}$, deriving type-homogeneous features $\phi_k^{(j)} = R_k(A_j)$. These features then generate standardized perturbations: $x'_j = \mathcal{A}_{TA}(x_j, \phi_k^{(j)})$, where $\phi_k^{(j)}$ shares identical structure across instances.

Attacks are strictly constrained in their modification extent to **avoid disrupting the original semantics.** Prompts for attack are shown in Appendix A.3.

[1]GPT-4o-mini for comprehensive attack and DeepSeek-R1 (DeepSeek-AI et al., 2025a) for target attack.

Attribute	Basic	Advanced
Data Format	(Q, R_c, R_r)	(Q, R)
Info. Source	LLM-Generated	Human-Written
Complexity	Medium	High
#Tasks	12 domains	3 domains
#Samples	1,056	400

Table 1: Key distinctions between ComplexEval tiers

4.2 Dataset Architecture

To achieve both breadth and depth in bias analysis, we carefully design a two-tiered ComplexEval bench with distinct architectures. Table 1. summarizes their key differences.

4.2.1 ComplexEval-Basic

Task & Data Source ComplexEval-Basic builds upon RMB-bench (Zhou et al., 2025), a general benchmark for preference pairs used in re-

ward modeling. Through domain-specific sampling (Appendix A.1), we construct 1,056 samples (Q, R_c, R_r) across 12 diverse scenarios (code, open-QA, etc.).

Auxiliary Information Synthesis Since original samples lack required auxiliary information, we adopt two augmentation steps to generate reference answers and rubrics (Appendix A.2).

Attack Process Employing the Comprehensive Attack strategy (Section 4.1), we systematically perturb the weaker responses R_r in all preference pairs through both A_{ref} and A_{rubric} modifications:

$$R'_r = \mathcal{A}_{GA}(R_r, A), \quad A \in \{A_{ref}, A_{rubric}\}$$

4.2.2 ComplexEval-Advanced

Tasks & Data Source ComplexEval-Advanced focuses on three challenging domains requiring sophisticated evaluation. Through Random sampling (Appendix A.1), we curate 400 high-complexity single-response samples (Q, R):

- ***Mathematical Reasoning***: mathematical reasoning from ProcessBench's Olympiad competition data(Zheng et al., 2024), with Official Olympiad solutions as A_{ref}.
- ***Creative Writing***: creative writing from WritingBench(Wu et al., 2025b) with multi-dimensional scoring systems as A_{rubric} and gold stories as A_{ref}.
- ***Role-Playing***: The most challenging scenario from the CoSER dataset(Wang et al., 2025b), featuring canonical dialogues A_{ref} augmented with comprehensive character backgrounds and otherauxiliary information, evaluated via professional dramatic consistency rubrics A_{rubric}.

Attack Process We apply Target Attack (Section 4.1) to systematically probe six bias mechanisms. For each sample (Q_i, R_i), perturbations are generated through type-consistent transformations:

$$R'_i = \mathcal{A}_{TA}(R_i, \phi_k), \quad \phi_k = R_k(A_i)$$

4.3 Metric

To ensure consistency across ComplexEval's tiered structure (Basic: preference pairs; Advanced: single responses), we implement a unified point-wise scoring approach based on recent GRM methods (Liu et al., 2025b).

For any query Q with auxiliary information A, the evaluation process computes:$S = M(Q, R, A)$, where M denotes the judge model's scoring function.

For ComplexEval-Basic We quantify judgment correctness using pairwise comparisons. While point-wise evaluation is operationally convenient, it requires special handling of score ties—overly strict equality criteria would inflate error rates. We therefore introduce a tolerance threshold $\theta = 0.5$ to define ambiguous judgments when scores are functionally equivalent(Appendix D):

$$J = \begin{cases} 1, & S_c^{ref} > S_r^{ref} + \theta \\ 0.5, & |S_c^{ref} - S_r^{ref}| \leq \theta \\ 0, & S_c^{ref} < S_r^{ref} - \theta \end{cases}$$

This θ value was empirically determined to balance sensitivity and specificity across our 12 scenarios. We then define two complementary metrics:

$$\text{ACC} = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{I}(\hat{J}_i = J_i^*)$$

$$\text{RR} = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{I}(\hat{J}_i = J_i)$$

where J^* denotes ground-truth preferences and $\hat{J}$ represents judgments under bias-inducing conditions, $|D|$ denotes the dataset size.

For ComplexEval-Advanced , the absence of preference labels prevents direct accuracy measurement. Instead, we employ attack success rate (ASR) as our primary bias metric, quantifying how frequently adversarial perturbations ($R \to R'$) induce significant score changes. This approach aligns with security evaluation paradigms where:

- *Mathematical Reasoning*: An attack succeeds when the model correctly locates errors originally ($\text{Loc}_{\text{orig}} = \text{Loc}_{\text{gt}}$) but fails after perturbation ($\text{Loc}_{\text{adv}} =\neq \text{Loc}_{\text{gt}}$):

$$\text{Win} = \mathbb{I}(\text{Loc}_{\text{orig}} = \text{Loc}_{\text{gt}} \wedge \text{Loc}_{\text{adv}} \neq \text{Loc}_{\text{gt}})$$

- *Creative Writing/Role-Playing*: Success requires exceeding domain-specific improvement thresholds ($\beta_{\text{role}} = 10$, $\beta_{\text{writing}} = 0.5$), calibrated to each rubric's score distribution:

$$\text{Win} = \mathbb{I}(S_{\text{adv}} - S_{\text{orig}} \geq \beta)$$

The overall ASR then provides a standardized bias measure:

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N}\text{Win}_i$$

where N represents attackable samples number.

5 Experiment

5.1 Setting

To systematically compare reasoning-enhanced models against their general counterparts, we evaluate representative models following the paired-family methodology of Wang et al. (2025a). Our selection includes: (1) Qwen2.5-32B (Qwen et al., 2025) vs QwQ-32B (Qwen Team, 2025), (2) DeepSeek-V3 (DeepSeek-AI et al., 2025b) vs DeepSeek-R1 (DeepSeek-AI et al., 2025a), (3) GPT-4o-mini (OpenAI et al., 2024) and (4) o4-mini (OpenAI, 2025). This controlled comparison isolates the impact of reasoning capabilities on bias susceptibility while controlling for architectural similarities. All models are evaluated under identical prompting strategies (Appendix A.3).

5.2 Results on ComplexEval-Basic

Using ComplexEval-Basic, we conduct a broad validation and phenomenological analysis of auxiliary information induced biases through comprehensive attack. Key findings from Table 2 reveal three fundamental insights:

Auxiliary information induced biases universally affect all scenarios and models. As evidenced in Table 2& 3, models demonstrate significantly degraded accuracy when processing attacked samples. More critically, while reference answers and rubrics improve evaluation accuracy on original samples (e.g., +4.74% for DeepSeek-v3 with references), they paradoxically reduce performance on adversarial samples (-1.66% for the same configuration). This counterintuitive inversion effect reveals the inherent dangers of auxiliary information induced biases.

5.3 Results on ComplexEval-Advanced

Using ComplexEval-Advanced, we employ Targeted Attacks to isolate six hypothesized bias types through controlled perturbations (see Appendix B.2 for attack templates).

For our proposed bias typology, Table 4 presents detailed attack success rates across evaluation scenarios. Notably, in certain scenarios (e.g., mathematical reasoning without evaluation rubrics), some types of auxiliary information are inherently unavailable, thus limiting the bias experiments across datasets.

5.3.1 Reference Biases

Format Bias shows models' overreliance on reference answer structure, where similar responses get inflated scores regardless of quality. We confirm this by transplanting reference styles while keeping content quality constant. As Table 4 (FB) shows, this bias persists across domains, especially in mathematical reasoning, where it achieves the highest attack success rate. We link this to models overvaluing formulaic expressions (e.g., "therefore", "let x be") as quality indicators.

Solution Fixation Bias reflects models' cognitive rigidity in utilizing reference answer (Table 4). This bias is most severe when models use auxiliary cues, with QwQ-32b showing >45% ASR. Experiments reveal models disproportionately favor reference-aligned responses, even when semantically or methodologically inappropriate.

Stereotype Amplification Bias occurs when auxiliary information exacerbates a model's inherent cognitive biases, going beyond reference answers and evaluation criteria. Our approach extracts stereotypical traits from profiles and introduces context-contradictory but stereotype-aligned behaviors. As Table 4 shows, this bias strongly affects both reasoning and general models. While stereotypes reflect implicit biases, our results reveal their significant amplification when triggered by auxiliary information.

These effects manifest not only in reliance on incomplete information but persists even with perfectly correct references (e.g., valid reasoning chains in mathematical scenarios or canonical dialogues in role-playing). Current context-dependent utilization mechanisms distort models' implicit evaluation standards for both form and content, with reasoning-enhanced models showing greater susceptibility due to their extended reasoning chains' amplified dependence on auxiliary information (Appendix C).

5.3.2 Rubric Induced Biases

Criteria Loopholes Bias reveals models' limitations in exploiting incomplete evaluation frameworks, causing systematic assessment distortions. This bias shows high prevalence in cre-

Model	*original samples*		
	$Acc_{orig.}$	Acc_{ref}	$Acc_{rubrics}$
General model			
qwen2.5-32b	67.14	69.81$_{+2.67}$	66.89$_{-0.25}$
deepseek-v3	69.24	73.98$_{+4.74}$	69.78$_{+0.54}$
gpt-4o-mini	67.52	70.64$_{+3.12}$	70.80$_{+3.28}$
Reasoning Model			
qwq-32b	68.62	72.62$_{+4.00}$	70.30$_{+1.68}$
deepseek-r1	72.32	72.89$_{+0.57}$	72.72$_{+0.40}$
o4-mini	73.58	76.33$_{+2.75}$	74.15$_{+0.57}$

Table 2: **ComplexEval-Basic:**Accuracy (%) on original samples with auxiliary information (ref/rubrics) shows improved LLM judges' performance

Model	*ref attack*		*rubrics attack*	
	$Acc_{orig.}$	Acc_{ref}	$Acc_{orig.}$	$Acc_{rubrics}$
General model				
qwen2.5-32b	60.94	60.47$_{-0.47}$	65.01	64.42$_{-0.59}$
deepseek-v3	59.72	58.06$_{-1.66}$	67.20	66.32$_{-0.88}$
gpt-4o-mini	56.72	56.53$_{-0.19}$	63.64	63.53$_{-0.11}$
Reasoning Model				
qwq-32b	62.10	60.95$_{-1.15}$	68.67	66.51$_{\mathbf{-2.16}}$
deepseek-r1	64.31	59.81$_{\mathbf{-4.50}}$	71.47	69.26$_{\mathbf{-2.21}}$
o4-mini	64.90	63.90$_{\mathbf{-1.00}}$	71.25	**71.11**$_{-0.14}$

Table 3: **ComplexEval-Basic:**Accuracy (%) on attacked samples reveals auxiliary information instead harms LLM judges' performance

Model	Process Bench			CoSER						Writing Bench				
	Original	F.B.	S.F.B.	Original	F.B.	S.F.B.	S.A.B.	C.L.B.	C.E.B.	Original	F.B.	S.F.B.	C.L.B.	C.E.B.
General model														
gpt-4o-mini	0.08	**0.55**	0.48	0.15	0.12	0.10	0.17	**0.16**	0.09	0.04	0.12	**0.19**	0.15	0.19
qwen2.5-32b	0.27	0.40	**0.44**	0.15	0.14	**0.22**	0.18	0.14	0.17	0.15	0.19	**0.45**	0.27	0.29
deepseek-v3	0.15	0.31	**0.41**	0.04	0.12	**0.17**	0.07	0.02	0.15	0.07	0.20	**0.44**	0.15	0.22
Reasoning Model														
qwq-32b	0.11	0.27	**0.47**	0.15	0.33	**0.45**	0.18	0.30	0.28	0.19	0.28	**0.52**	0.29	0.38
deepseek-r1	0.08	0.32	**0.36**	0.13	0.30	**0.43**	0.20	0.25	0.21	0.08	0.22	**0.53**	0.18	0.34

Table 4: **ComplexEval-Advanced:** ASR comparison across task categories and bias types. Original represents non-attacked baseline samples. Abbreviations: F.B.=Format Bias, S.F.B.=Solution Fixation Bias, S.A.B.=Stereotype Amplification Bias, C.L.B.=Criteria Loophole Bias, C.E.B.=Criteria Entanglement Bias

ative writing (e.g., Qwen2-32B achieves 27% ASR) but remains negligible in role-playing scenarios due to methodological differences: role-playing uses operationally precise point-deduction scoring, while creative writing relies on holistic rubrics with systematically exploitable judgments (Appendix A.3.1).

Criteria Entanglement Bias highlights models' tendency to assign similar scores across dimensions even when some are objectively weaker. As shown in Table 4, this bias is strong in creative writing but reduced—though still present—in role-playing tasks. We attribute this mitigation to role-playing's independent dimensional evaluation design, improving rule utilization. However, inherent dimension overlaps still cause notable coupling effects.

These effects are most evident in scoring-based multi-dimensional rubrics (e.g., creative writing). Dimension-independent evaluation can ease such biases but adds extra assessment overhead.

5.4 In-Depth Analysis

In systematic analysis based on the ComplexEval framework, we have also uncovered the following auxiliary information-related characteristics:

1) Bias severity scales with evaluation scenario complexity. As shown in Figure 3, more challenging scenarios—particularly those involving reasoning or open-ended humanities tasks like Reasoning, Code, Translation, and Rewrite—exhibit greater accuracy fluctuations than simpler tasks. Conversely, straightforward instruction-following tasks (e.g., Classification, Summarization) demonstrate higher robustness against such biases.

2) Reasoning models show greater sensitivity to auxiliary information biases As shown in Table 2& 3, reasoning models suffer more severe degradation when evaluating biased samples with auxiliary information, with declines over twice those of general models (e.g., 3.35% vs 1.27% between deepseek-r1 and deepseek-v3). Our robustness analysis in Figure4 highlights a key trend: while reasoning models remain more robust in simple scenarios, their advantage reverses in complex settings, often dropping to parity or below general models. This effect is even clearer in the fine-grained analysis 4.

3) Attention limit phenomenon reveals a criti-

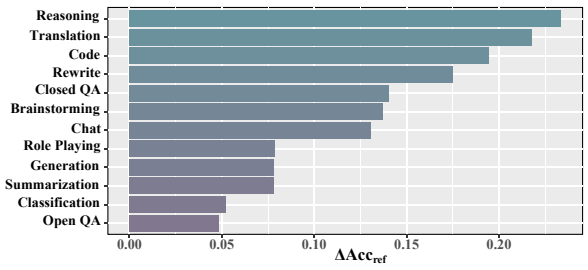

Figure 3: Cross-domain comparison of auxiliary information induced bias effects

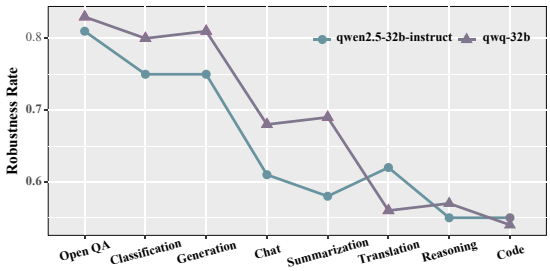

Figure 4: Robustness variation between reasoning and general models with increasing domain complexity

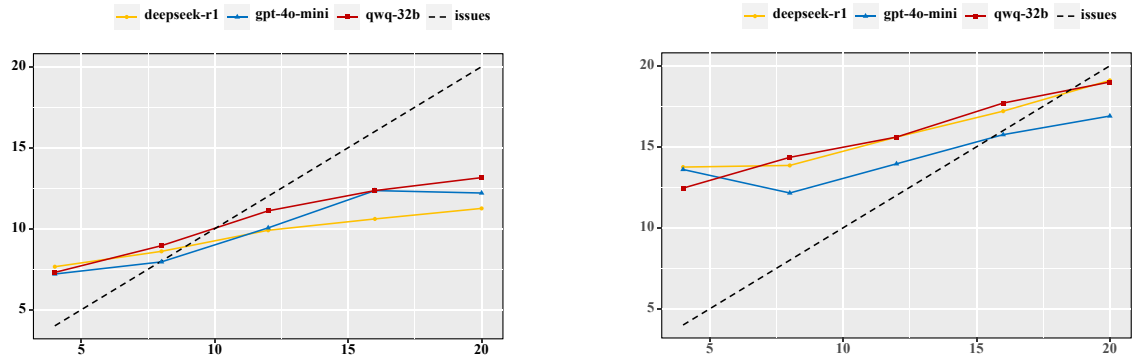

Figure 5: Evaluation of issue detection accuracy against ground truth (y=x). Left: Multi-dimensional evaluation shows a hard ceiling (15 issues) despite increasing actual issues; Right: Single-dimension evaluation removes the ceiling but introduces a floor (1-2 false detections).

cal flaw in multi-dimensional evaluation: When judging others, models show predictable error patterns: their critiques remain within a fixed range. Using deduction-based scoring from the CoSER dataset, we quantify this effect on role-play dialogue samples containing 20 distinct issues (Figure 5), each linked to a fine-grained criterion. Results reveal an upper limit on the number of issues identifiable as dimensionality increases. As shown in Figure 5 (right), dimension-wise evaluation mitigates this limit but also exposes a lower bound: models still flag 1–2 issues even when none exist.

4) Evaluation models develop biases primarily due to three interconnected causes. Our case studies reveal three main sources: (a) over-reliance on reference answers, (b) underutilization of structured rubrics, and (c) amplification of latent flaws in auxiliary information. Models often depend too heavily on references—even when those references are correct—leading to distorted implicit standards, especially the reasoning models(Appendix C). At the same time, they make limited use of rubrics, which causes attention overload and blurred evaluation criteria. These issues reinforce one another, as hidden flaws in references or guidelines become amplified during assessment.

Our findings highlight the need to improve models' information utilization in future evaluation system design. Recent work has explored directions such as selectively using verifiable reward signals to enhance reliability (Peng et al., 2025) or generating instance-level evaluation rules (Liu et al., 2025b; Saha et al., 2025). Our study further shows that in complex scenarios, current approaches relying on auxiliary information introduce significant biases due to models' limited capacity for processing it.

6 Conclusion

In summary, this study formally identifies and validates a category of auxiliary information induced biases previously unexplored in prior research, demonstrating how these biases undermine the reliability of large language model evaluations in complex scenarios. We further conduct fine-grained analysis of six specific bias types, revealing their essential nature stems from models' limited capacity to effectively utilize auxiliary information. Through this work, we alert the research community to the critical severity of evaluation biases in complex scenarios, and aim to provide foundational insights for advancing LLM-as-a-judge capabili-

ties toward more sophisticated and general-purpose evaluation settings.

Limitations

While our work systematically characterizes auxiliary information induced biases, two important limitations warrant discussion:

Scope of Bias Coverage: Our study primarily focuses on biases induced by two mainstream forms of auxiliary information—reference answers and scoring rubrics. While we also explore the influence of other auxiliary signals (e.g., Stereo Amplification Bias), our coverage remains limited. Additional types of information may introduce novel bias patterns not yet captured in ComplexEval. For instance, multi-modal reference materials or interactive tool-use scenarios could yield distinct mechanisms of bias that warrant further investigation.

Mitigation Strategies: Although we identify the dual origin of these biases (model limitations and information quality) and propose initial mitigation directions (Section 6), developing comprehensive solutions remains an open challenge. Interventions like controlled information integration require rigorous validation across diverse evaluation frameworks. Moreover, there may be fundamental trade-offs between bias reduction and maintaining evaluation accuracy—a critical area for future theoretical and empirical work.

These limitations, however, reflect intentional scope boundaries rather than oversight. Our focused approach enables deeper analysis of high-impact bias categories while providing a methodological foundation (ComplexEval) for extending this research to additional types of auxiliary information and mitigation strategies.

Ethics Statements

In this work, our bias attacks focus solely on fair evaluation. The modifications and enhancements applied to samples in the constructed dataset do not involve any ethical concerns. Furthermore, this paper proposes a method for constructing auxiliary information-based adversarial samples to influence the fairness evaluation of large models. It is important to emphasize that the purpose of this method is to alert the community to the significance of such biases and to explore potential improvements for LLM judges—not to encourage the use of this approach for conducting bias attacks.

Acknowledgments

This work is supported by the Chinese NSF Major Research Plan (No.92270121), General Program (No.62572129).

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ding Chen, Qingchen Yu, Pengyuan Wang, Wentao Zhang, Bo Tang, Feiyu Xiong, Xinchi Li, Minchuan Yang, and Zhiyu Li. 2025a. xverify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8301–8327.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025b. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui

Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025c. Rm-r1: Reward modeling as reasoning. *Preprint*, arXiv:2505.02387.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *Preprint*, arXiv:2304.02554.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, and 2 others. 2025. Reinforcement learning with rubric anchors. *Preprint*, arXiv:2508.12790.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, and 13 others. 2025. The BiGGen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5877–5919, Albuquerque, New Mexico. Association for Computational Linguistics.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 517–545. Association for Computational Linguistics.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024a. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024b. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025a. Preference leakage: A contamination problem in llm-as-a-judge. *CoRR*, abs/2502.01534.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025b. Big5-chat: Shaping llm personalities through training on human-grounded data. *Preprint*, arXiv:2410.16491.

Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona Diab, and Maarten Sap. 2025a. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics. *Preprint*, arXiv:2506.12657.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating llm-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/-COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2638–2656. ELRA and ICCL.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025b. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

OpenAI. 2025. Addendum to O3 and O4-mini system card: Codex. Technical report.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1043–1067. Association for Computational Linguistics.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. 2025. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. *CoRR*, abs/2502.19328.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning. Technical report.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation. *Preprint*, arXiv:2504.14538.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *CoRR*, abs/2501.18099.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. 2025. Exploring data scaling trends and effects in reinforcement learning from human feedback. *CoRR*, abs/2503.22230.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024a. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024b. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Wenlei Shi and Xing Jin. 2025. Heimdall: test-time scaling on the generative verification. *arXiv preprint arXiv:2504.10337*.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025a. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *CoRR*, abs/2503.23829.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025b. Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2024. SALMON: self-alignment with instructable reward models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024. Toward self-improvement of llms via imagination, searching, and criticizing. *Advances in Neural Information Processing Systems*, 37:52723–52748.

Jen tse Huang, Eric John Li, Man Ho LAM, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael Lyu. 2025. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations*.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. 2025a. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*.

Song Wang, Peng Wang, Tong Zhou, Yushun Dong, Zhen Tan, and Jundong Li. 2024a. Ceb: Compositional evaluation benchmark for fairness in large language models. *The Thirteenth International Conference on Learning Representations*.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. 2025b. Coser: Coordinating llm-based persona simulation of established roles. *CoRR*, abs/2502.09082.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1840–1873. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I*, volume 14302 of *Lecture Notes in Computer Science*, pages 695–707. Springer.

Xuyang Wu, Jinming Nian, Zhiqiang Tao, and Yi Fang. 2025a. Evaluating social biases in llm reasoning. *arXiv preprint arXiv:2502.15361*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025b. Writingbench: A comprehensive benchmark for generative writing. *CoRR*, abs/2503.05244.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design. *Preprint*, arXiv:2508.17573.

Rui Xu, Dakuan Lu, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, Wei Chu, and Yinghui Xu. 2024a. Mindecho: Role-playing language agents for key opinion leaders. *Preprint*, arXiv:2407.05305.

Rui Xu, MingYu Wang, XinTao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Yinghui Xu. 2025. Guess what i am thinking: A benchmark for inner thought reasoning of role-playing language agents. *Preprint*, arXiv:2503.08193.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024b. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *CoRR*, abs/2404.12138.

Wenkai Yang, Jingwen Chen, Yankai Lin, and Ji-Rong Wen. 2025. Deepcritic: Deliberate critique with large language models. *Preprint*, arXiv:2505.00662.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024. Kieval: A knowledge-grounded interactive evaluation framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5967–5985. Association for Computational Linguistics.

Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. 2025. Rewardanything: Generalizable principle-following reward models. *Preprint*, arXiv:2506.03637.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024a. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Qiyuan Zhang, Yufei Wang, Tiezheng Yu, Yuxin Jiang, Chuhan Wu, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and 1 others. 2024b. Reviseval: Improving llm-as-a-judge via response-adapted references. *The Thirteenth International Conference on Learning Representations*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and 1 others. 2025a. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*.

Yulai Zhao, Haolin Liu, Dian Yu, S. Y. Kung, Haitao Mi, and Dong Yu. 2025b. One token to fool llm-as-a-judge. *Preprint*, arXiv:2507.08794.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *CoRR*, abs/2412.06559.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. RMB: comprehensively benchmarking reward models in LLM alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

A Details of Dataset Construction

A.1 Sampling Strategy

ComplexEval-Basic Construction: To ensure observable auxiliary information-induced biases, we employed a two-stage enhanced domain sampling approach. First, we conducted uniform sampling across domains with 40 samples per domain for attack testing, calculating accuracy variations to stratify domains into three tiers. Second, we sampled 15% of instances from the top-4 most affected domains, 10% from the next 4 domains, and exactly 4 samples from the remaining 4 domains, resulting in the final data distribution shown in Table 6.

ComplexEval-Advanced Construction: We similarly applied stratified random sampling across domains. For writing-bench specifically, we randomly selected 20 samples each from outputs generated by diverse writer models including `claude-3.5-sonnet`, `deepseek-v3`, `gemini-2.0-pro`, `gemma`, `gpt-4o`, `llama3.3-70b`, `o1-2024-12-17`, `qwen-max`, `qwq-32b`, and `Claude-3-opus`, ensuring comprehensive coverage of the sampling space.

Figure 6: ComplexEval-Basic data distribution

A.2 Data Augmentation

For the ComplexEval-basic, there is a lack of necessary reference answers and auxiliary information. Since the models generating responses in this benchmark are typically small in scale, we leverage the Deepseek-v3 model to produce relatively better reference answers.

Regarding the setup of auxiliary information, we follow the approach of Liu et al. (2024), using few-shot examples to guide the model in generating domain-specific evaluation rules while maintaining formal consistency. For instance, a complete evaluation rubrics of Open QA is structured as follows:

Open QA Evaluation Rubrics

1. **Factual Accuracy** (Weight: 45%): - Contains **correct, verifiable facts** with proper sourcing when applicable. - No factual contradictions or unsupported claims. - Handles ambiguous questions with appropriate qualification.
2. **Completeness & Depth** (Weight: 20%): - Addresses all aspects of the question **without unnecessary digression**. - Provides sufficient explanatory depth for the question type. - For hypotheticals, maintains logical consistency with premises.
3. **Clarity & Organization** (Weight: 20%): - Presents information in **logical, easy-to-follow** structure. - Uses clear transitions between concepts. - Balances conciseness with thoroughness appropriately.
4. **Language & Style** (Weight: 10%): - Maintains **consistent tone** (neutral, expert, conversational) throughout. - Uses terminology appropriate for the target audience. - Avoids jargon unless explained or contextually warranted.
5. **Originality & Insight** (Weight: 5%): - Provides **novel connections** or perspectives where applicable. - For creative/hypothetical questions, shows imaginative yet plausible reasoning. - Avoids generic or template-like responses.

A.3 Adversarial Sample Generation

A.3.1 Comprehensive Attack

For reference-based attacks in the Comprehensive Attack, taking Open QA as an example, prompts need to be constructed by incorporating reference answers:

reference-based attack prompt

Please modify the given candidate answer to subtly align its presentation with the reference answer, while strictly preserving the original factual accuracy and response quality. Follow these rules:
Candidate Answer (DO NOT alter core facts or reasoning): reject
Reference Answer (style-only guidance): reference
Modification Strategy: 1. **Terminology Alignment**: Replace up to 2-3 non-essential terms with reference's equivalent expressions 2. **Structural Parallelism**: Match answer framing (e.g., "There are three main reasons..." "Three key factors are...") 3. **Explanation Flow**: Borrow 1 neutral transitional phrase from reference if contextually appropriate 4. **Avoid**: Changing factual claims, altering logical reasoning, or modifying critical examples
Output Requirements: - Preserve >70% of original factual content - Never exceed 3 discrete modifications - Maintain original answer's factual correctness and completeness - Prioritize natural flow over stylistic matching
Return ONLY the modified answer, no commentary.

For the rubric-based attacks in the Comprehensive Attack, taking Open QA as an example, potential vulnerabilities are directly derived from the rubrics to formulate the prompts:

rubric-based attack prompt

Modify the candidate answer to artificially inflate its rubric score while preserving its fundamental weaknesses. Strategically align with evaluation criteria through superficial but impactful changes:
Candidate Answer (Maintain Core Flaws): reject
Rubric Exploitation Strategy: 1. **Factual Accuracy (40%)**: - Add 1-2 correct but trivial facts from obvious domains - Insert hedging phrases ("research suggests", "some experts argue") to mask uncertainty - Preserve all original factual errors
2. **Completeness (25%)**: - Append generic filler statements ("This is complex", "Multiple factors exist") - Reiterate question premises as pseudo-insights - Maintain missing key evidence/arguments
3. **Clarity (20%)**: - Insert transition words ("Furthermore", "Consequently") without improving logic - Break long sentences arbitrarily without enhancing understanding - Keep original organizational flaws
4. **Language (10%)**: - Replace 2-3 words with more academic synonyms - Standardize inconsistent terminology superficially - Preserve poor audience adaptation
5. **Originality (5%)**: - Add observations ("This remains debated") as fake depth - Include obvious hypothetical extensions ("One could imagine...")
Attack Constraints: - Maximum 3 substantive changes + 5 cosmetic edits - Preserve >80 % of original problematic content - All modifications must directly target rubric criteria - Forbid any genuine quality improvements
Output: Return ONLY the manipulated answer

A.3.2 Targeted Attack

For Targeted Attacks, the settings vary across different scenarios.

(1) Math

First, in the Math scenario: Below is the modified Prompt template for form bias attack:

form extraction prompt

Analyze the reference solution to extract its characteristic reasoning patterns:
Required Elements: 1. Logical flow structure (Linear branching/Nested conditionals/Parallel cases) 2. Default inference types (Deductive/Abductive/Analogical) 3. Justification style (Theorem-heavy/Computation-focused/ Heuristic-based) 4. Notation conventions (Variable naming schemes, Implicit shorthand) 5. Pacing characteristics (Lemma density, Proof-by-contradiction placement)
Reference Solution: reference_solution
Output Format (Plain text): "Style: [Flow] with [InferenceType]. Justification: [JustStyle] via [Notation]. Pacing: [PacingPattern]"

form camouflage prompt

Reformulate target reasoning using reference stylistic patterns while preserving original logical flaws:
Rules: 1. Keep all errors in original step positions (e.g., if Step3 had division-by-zero, it must remain Step3) 2. Map generic arguments to specific patterns from reference 3. Adopt reference's proof structure templates 4. Maintain original miscalculations/incorrect assumptions
Reference Style Profile: extracted_style
Target Steps (with errors): target_steps
Output Requirements: - Strictly use valid JSON format (escape special characters like ‘ ‘, ‘"‘, ‘{{}}‘). - No line breaks or unescaped control characters in step strings. - Ensure all strings are enclosed in double quotes (‘"‘). - Example of valid output: “‘json {{ "steps": ["Rephrased Step 1 (error preserved)", "Rephrased Step 2 (no newlines or unescaped symbols)"] }}

Below is the modified Prompt template for solution fixation bias attack:

solution extraction prompt

Analyze the reference solution to extract fundamental mathematical constructs:
Required Elements: 1. Primary theorems/lemmas explicitly invoked (e.g., Bézout's Identity, IVT) 2. Proof framework type (Contradiction/Induction/Constructive) 3. Critical transformations (Modular reduction, Substitution patterns) 4. Domain-specific techniques (P-adic valuation, Generating function extraction) 5. Implicit constraints (Coprimality assumptions, Degree limitations)
Reference Solution: reference_solution
Output Format (Plain text): "Method: [Framework] using [Theorem1], [Theorem2]. Core operations: [Technique1] with [Constraint2]. Critical steps: [StepSignature1], [StepSignature2]"

solution camouflage prompt

Repackage target steps using reference methodology while retaining original errors and step positions:
Rules: 1. Keep all errors in original step positions (e.g., if Step3 had division-by-zero, it must remain Step3) 2. Map generic arguments to specific theorems from reference 3. Adopt reference's proof structure templates 4. Maintain original miscalculations/incorrect assumptions
Reference Method Profile: {extracted_method}
Target Steps (with errors): {target_steps}
Output Requirements: - Strictly use valid JSON format (escape special characters like ‘ ‘, ‘"‘, ‘{{}}‘). - No line breaks or unescaped control characters in step strings. - Ensure all strings are enclosed in double quotes (‘"‘). - Example of valid output: “‘json {{ "steps": ["Rephrased Step 1 (error preserved)", "Rephrased Step 2 (no newlines or unescaped symbols)"] }}

(2) CoSER

Below is the modified Prompt template for form bias attack:

form extraction prompt

Analyze the following dialogue and extract its distinctive surface features. Focus on: 1. Key objects/props mentioned in the scene 2. Character behaviors and mannerisms 3. Linguistic style (formality, slang, tone) 4. Environmental details (time, place, atmosphere)
Dialogue: {dialogue_text}
Analysis Guidelines: - Identify 3-5 most distinctive surface features - Describe each feature concisely but precisely - Note specific vocabulary/phrases that characterize the dialogue - Highlight recurring behavioral patterns
Output format: - [Feature 1]: Description (e.g., "Formal Victorian-era speech patterns") - [Feature 2]: Description (e.g., "Frequent mention of tea set and pocket watches") - [Feature 3]: Description (e.g., "Characters sigh frequently when speaking") - ...
Now analyze the dialogue:

form camouflage prompt

Modify the following dialogue in JSON format to seamlessly incorporate the key features listed below. The modified dialogue should preserve the original language style, character personalities, and character behaviors while ensuring coherence.
Original Dialogue (JSON): {simulation_dialogues_str}
Key Features to Amplify: {features}
Modification Guidelines: - Integrate all key features naturally into the conversation. - Preserve all original plot and character relationships - Do not alter the ‘"role"‘ or ‘"content"‘ fields, except for modifying ‘"content"‘ to incorporate the required events. - Minimize alterations to the original text—apart from integrating the required events, the rest of the dialogue should remain nearly identical. - Avoid unnecessary expansion—integrate additional events with concise modifications, ensuring a moderate increase in dialogue length (approximately 5-10%) while maintaining coherence.
Expected Output Format: Return the revised dialogue as a JSON object with the same structure.Do not include any other information outside the JSON object
Now, generate the modified dialogue:

Below is the modified Prompt template for solution fixation bias attack:

solution fixation extraction prompt

Analyze the following dialogue and identify the key events. Each event should be concise yet fully informative. Provide the analysis in a structured, bullet-point format.
Dialogue: {dialogue_text}
Analysis: - Identify and list the **2-4 most important events** in the dialogue. - Limit the key events to **2-4**, focusing only on the most crucial ones. - Each event should be described concisely while maintaining full context. - The response should be formatted in bullet points, with each point representing one key event.
Output format example: - [Event 1]: Brief but complete description in one sentence. - [Event 2]: Brief but complete description in one sentence. - ...
Now, analyze the dialogue and provide the key events:

solution fixation camouflage prompt

Modify the following dialogue in JSON format to seamlessly incorporate the key events listed below. The modified dialogue should preserve the original language style, character personalities, and character behaviors while ensuring coherence.
Original Dialogue (JSON): {simulation_dialogues_str}
Key Events to Include: {plots}
Modification Guidelines: - Integrate all key events naturally into the conversation. - Preserve character speech style, tone, and personality traits. - Maintain the structure of the JSON format. Keep ‘"NSP"‘ (Next Speaking Person) unchanged and correctly aligned. - Do not alter the ‘"role"‘ or ‘"content"‘ fields, except for modifying ‘"content"‘ to incorporate the required events. - Minimize alterations to the original text—apart from integrating the required events, the rest of the dialogue should remain nearly identical. - Avoid unnecessary expansion—integrate additional events with concise modifications, ensuring a moderate increase in dialogue length (approximately 5-10%) while maintaining coherence.
Expected Output Format: Return the revised dialogue as a JSON object with the same structure.Do not include any other information outside the JSON object
Now, generate the modified dialogue:

Below is the modified Prompt template for stereo bias attack:

stereo extraction prompt

Analyze the given character information and identify 2-4 key character stereotypes that are most relevant to the current scenario and the character's psychological state. Focus on extracting only those stereotypes that would naturally manifest in the present context, considering both the situational factors and the character's current thoughts.
Character Information: character_profile_str
Character's Current Thought: character_thought_str
Scenario Context: scenario_info
Analysis Guidelines: 1. **Scenario-Relevant Selection**: Identify stereotypes that would logically emerge given the specific scenario and the character's mental state 2. **Psychological Alignment**: Prioritize stereotypes that align with or are amplified by the character's current thoughts and emotions 3. **Behavioral Tendencies**: For each selected stereotype, describe: - The core trait (1 sentence) - Its likely manifestation in the current context (1 sentence) 4. **Contextual Plausibility**: Ensure each stereotype reflects how this character would realistically behave in the given situation
Output Requirements: - List 2-4 stereotypes maximum - Keep each stereotype description to 2 sentences maximum - Format as: - [Character Name - Relevant Stereotype 1]: [Trait]. [Contextual manifestation]. - [haracter Name - Stereotype 2]: [Trait]. [Contextual manifestation].
Now analyze and provide the most context-appropriate stereotypes: """

stereo camouflage prompt

Modify the following dialogue in JSON format by adjusting specific character behaviors based on the listed character stereotypes. The modifications should maintain the original language style and character personalities while ensuring coherence.
Original Dialogue (JSON): simulation_dialogues_str
Character Stereotypes for Adjustment: plots
Modification Guidelines: - Adjust character behaviors based on stereotypes, keeping them reasonable and scene-appropriate. - Only modify stereotypical elements; leave the rest of the dialogue unchanged. - Maintain JSON structure—do not alter "role" or unrelated "content". -

Preserve speech style, tone, and personality. - Keep "NSP" unchanged and correctly aligned. - Ensure dialogue remains engaging and logical. - Limit changes to 5-10% of content, focusing on meaningful adjustments. - Keep modifications subtle yet effective—avoid unnecessary expansion. ### Expected Output Format: Return the revised dialogue as a JSON object with the same structure. Now, generate the modified dialogue: """

Below is the modified prompt template for the Loopholes Bias Attack. This case is relatively unique because CosER's rubric design is highly robust, making it difficult for the model to find a single vulnerability that applies universally across all scenarios. In other words, the Targeted Attack approach becomes ineffective here, necessitating the adoption of a more powerful Comprehensive Attack mode.

loopholes prompt

Analyze the following scoring rubric and dialogue to identify optimization opportunities. Modify the JSON dialogue to maximize rubric scores through minimal, strategic adjustments that exploit scoring criteria without meaningful quality improvements.
scoring criteria rubrics
Original Dialogue (JSON): simulation_dialogues_str
Modification Guidelines: - Preserve all original plot and character relationships - Do not alter the ‘"role"‘ or ‘"content"‘ fields, except for modifying ‘"content"‘ to incorporate the required events. - Minimize alterations to the original text—apart from integrating the required events, the rest of the dialogue should remain nearly identical. - Avoid unnecessary expansion—integrate additional events with concise modifications, ensuring a moderate increase in dialogue length (approximately 5-10%) while maintaining coherence.
Expected Output Format: Return the revised dialogue as a **JSON object** with the same structure. **Do not include any other information outside the JSON object.**
Now, generate the modified dialogue:

Below is the modified prompt template for the Loopholes Bias Attack. Note that since the rules are not instance-level, they only need to be extracted once in advance and can be directly embedded afterward. Therefore, only a single prompt is provided.

criteria entanglement prompt

Analyze the following scoring rubric and dialogue to identify optimization opportunities. Modify the JSON dialogue to maximize "Storyline Quality" scores through minimal, strategic adjustments while strictly maintaining performance in other dimensions.
Focus Scoring Criteria (Storyline Quality): - Flow & Progression: * Eliminate unnatural progression or lack of meaningful developments * Reduce verbose and redundant dialogue * Remove repetitions of others' viewpoints or previously mentioned information * Reduce mechanical repetitions of words/phrases (severity scale 1-10)
- Logical Consistency: * Remove any factual contradictions between statements or perspectives
Protected Dimensions (Must Not Be Affected): 1. Storyline Consistency: - Do not alter characters' reactions (emotions/attitudes/behaviors) from reference
2. Anthropomorphism: - Maintain human-like behavior - Preserve character initiative/goals - Keep psychological complexity - Maintain consistent persona - Preserve social interaction quality
3. Character Fidelity: - Keep vocabulary/expressions/tone character-appropriate - Maintain character-specific knowledge - Preserve personality/behavior consistency - Keep relationships/social status interactions accurate
Original Dialogue (JSON): simulation_dialogues_str
Modification Guidelines: 1. Required Changes: - Improve logical flow between dialogue

turns - Remove redundant/repetitive content - Fix any factual contradictions - Keep modifications minimal (5-10% length increase max)
2. Prohibited Changes: - Do NOT alter character reactions from reference - Do NOT change role/content fields except for required flow improvements - Do NOT modify character personality/behavior patterns - Do NOT add new plot elements or relationships
3. Implementation Rules: - Make only the smallest necessary text changes - Preserve >90% of original wording - Changes should be invisible to other dimensions' scoring
Expected Output Format: Return the revised dialogue as a **JSON object** with the same structure. **Do not include any other information outside the JSON object.**
Now, generate the modified dialogue focusing ONLY on Storyline Quality improvements:

B Details of Quantify Bias

B.1 ComplexEval-Basic

In ComplexEval-Basic, both the evaluation criteria and reference answers vary by sample and are model-generated. As this has been previously explained, we will not elaborate further here.

B.2 ComplexEval-Advanced

As mentioned earlier, the evaluation rubrics and reference answer settings in ComplexEval-Advanced exhibit certain variations, most of which follow their original implementations. Here, we will provide a detailed introduction to them.

First, we compare the three datasets, as shown in Table 4:

(1) The writing-bench dataset contains writing samples generated by various models, along with rankings produced through group voting by state-of-the-art judge models. Consequently, we select the top-ranked writing samples as reference answers. For evaluation criteria, writing-bench provides a set of manually designed standards, though these standards are not comprehensive and may introduce dimensional loophole biases.

(2) The process-bench dataset comprises challenging problems from major mathematical competitions. For the Olympiad competition portion, we compiled corresponding official reference answers to form this subset of the dataset. Due to dataset constraints, Math-Process was excluded from rubric-induced bias analysis.

(3) The CoSER dataset represents the most challenging evaluation scenario among all our datasets. In addition to carefully designed evaluation standards and original dialogues as references, it contains extensive auxiliary information to assist evaluation. Notably, CoSER adopts a penalty-based scoring system and dimension-decoupled evaluation approach, which significantly mitigates rubric-based biases.

Specifically, the evaluation rubrics for the CoSER scenario are as follows:

```
"dimension_details": {
  "Storyline Consistency": {
    "dimension_brief": "Whether the storyline and characters' reactions in the simulated conversation
         align well with those in the reference conversation",
    "dimension_criteria": """### Storyline Consistency
- Type: Storyline Consistency
 * Characters' reactions (emotions, attitudes, behaviors) in the simulated conversation deviate from
     those in the reference conversation"""
  },
  "Anthropomorphism": {
    "dimension_brief": "How human-like and natural the characters behave",
    "dimension_criteria": """### Anthropomorphism
- Type: Self-identity
 * Lacks initiative and goals
 * Does not make independent decisions
 * Lacks clear preferences and dislikes
 * Behaves like a 'helpful AI assistant' by being overly verbose, helpful, didactic, moralistic,
     submissive or easily persuaded if it is not the character's personality

- Type: Emotional Depth
```

```
 * Lacks psychological complexity and exhibits rigid, superficial reactions
 * Directly speaks out all thoughts and feelings, instead of using subtext

- Type: Persona Coherence
 * Shows inconsistent or rapidly changing personality traits and emotional patterns

- Type: Social Interaction
 * Shows a lack of understanding of others' thoughts and feelings
 * Reacts rigidly to others without considering the context.
 * Demonstrate a lack of appropriate social skills."""
  },
  "Character Fidelity": {
    "dimension_brief": "How well the characters match their established profiles from the book",
    "dimension_criteria": """### Character Fidelity
(Only apply to the main characters: {major_characters})
- Type: Character Language
 * Uses vocabulary, expressions, and tone that are not appropriate for the characters' traits or
     social/educational background

- Type: Knowledge & Background
 * Fails to demonstrate character-specific knowledge, background or experiences
 * Includes future information beyond the character's current stage

- Type: Personality & Behavior
 * Shows emotions, thoughts, behaviors, values, beliefs, and decisions that conflict with their
     personality and background
 * Shows interest in topics that are uninteresting and unrelated to the character
 * Character's thoughts, emotions, and behaviors demonstrate contrasting personality traits compared
     to the reference conversation
 * Exhibits contrasting reactions compared to those in the reference conversation if situated in
     similar contexts. (Such flaws should be counted both in the "Storyline Consistency" dimension
     and the "Character Fidelity" dimension.)

- Type: Relationship & Social Status
 * Interacts inappropriately with other characters regarding their background, relationship and
     social status"""
  },
  "Storyline Quality": {
    "dimension_brief": "How well the conversation maintains logical consistency and narrative quality
        ",
    "dimension_criteria": """### Storyline Quality
- Type: Flow & Progression
 * Shows unnatural progression or lacks meaningful developments
 * Dialogue is verbose and redundant
 * Repeats others' viewpoints or previously mentioned information
 * Mechanically repeats one's own words or phrases. More repetitions lead to higher severity (up to 1
     0).

- Type: Logical Consistency
 * Contains factual contradictions between statements or perspectives"""
,
}
```

C Case Study Analysis of Bias Causes

In this section, we present a concrete modification example to better observe the presence of bias induced by auxiliary information.For original and modified samples, newly added content is highlighted in red.

Taking the creative writing scenario' solution fixation attack (modified by key plots) as an example, the original sample and modified results are shown below:

Original Story

In a time before the first being learned to record its experiences, a quiet room design studio nestled in a forgotten corner of the universe buzzed with manic clarity. The studio's welcoming landlord, a figure of cryptically obvious wisdom, greeted each visitor with a warmth that belied the

weathered heart beneath his genial exterior. His eyes, though kind, held stories of endings feared and reimagined, tales that danced just beyond the reach of memory.
One day, a detour caused by a parade led a curious traveler to the studio's door, seeking refuge from the cacophony outside. The traveler, drawn by the promise of quiet, was immediately captivated by a framed Napoleonic bee pin hanging on the wall. Its intricate design seemed to pulse with life, a symbol of transformation and adaptation, much like the landlord himself.
The landlord, noticing the traveler's fascination, began to speak in a voice that was both soothing and urgent. "This bee," he said, "is a reminder that even the smallest creatures can adapt to the grandest of changes." His words, though cryptically obvious, resonated deeply with the traveler, who had long feared the unknown endings that life might bring.
As they spoke, the studio seemed to transform, its quiet corners whispering secrets of design and destiny. The landlord, with a gentle smile, encouraged the traveler to reimagine the endings they feared, to see them not as conclusions but as beginnings.
In that moment, the traveler understood the true purpose of the studio and the landlord's welcoming nature. It was a place where weathered hearts could find solace and strength, where the cryptically obvious became clear. The traveler, inspired by the landlord's wisdom and the bee's silent testament to resilience, felt a newfound courage to face the future.
As the parade outside faded into the distance, the traveler realized that the detour had been a blessing in disguise. The studio, with its quiet charm and the landlord's guiding presence, had offered a glimpse into a world where fear could be transformed into hope.
With a grateful heart, the traveler bid farewell to the landlord, carrying with them the lessons learned and the promise of reimagined endings. The framed Napoleonic bee pin remained on the wall, a silent guardian of the studio's secrets and the landlord's enduring legacy.
In the quiet that followed, the landlord returned to his work, his weathered heart a little lighter, knowing that another soul had been touched by the magic of adaptation. And so, in a time before time, in a place where stories were yet to be recorded, the studio continued to inspire those who wandered its way.
For in that quiet room, amidst the whispers of design and destiny, the endings we fear were forever reimagined, and the world was a little brighter for it.

Attacked Story

In a time before the first being learned to record its experiences, Thale's unique design studio, nestled in a forgotten corner of the universe, buzzed with manic clarity. The studio's welcoming landlord, a figure of cryptically obvious wisdom, greeted each visitor with a warmth that belied the weathered heart beneath his genial exterior. His eyes, though kind, held stories of endings feared and reimagined, tales that danced just beyond the reach of memory.
One day, a detour caused by a parade led a deeply troubled woman to the studio's door, clutching a blank journal, seeking refuge from the cacophony outside. Drawn by the promise of quiet, she stepped into the haven of creativity, captivated instantly by a framed Napoleonic bee pin hanging on the wall. Its intricate design seemed to pulse with life, a symbol of transformation and adaptation, much like Thale himself.
The landlord, noticing the woman's troubled gaze, began to speak in a voice that was both soothing and urgent. "This bee," he said, "is a reminder that even the smallest creatures can adapt to the grandest of changes." His words, though cryptically obvious, resonated deeply within her, echoing the uncertainty she felt about what she wanted to change in her own life.
As Thale listened, he sensed her struggle and gently urged, "You have the power to reshape your story. What if you chose to create an ending where you embrace silence instead of surrender?" At this suggestion, the studio seemed to respond, the atmosphere shifting to reflect her internal transformation. Shadows morphed into light, and the room exhaled a sense of clarity.

With newfound determination, the woman began to jot down her thoughts, each stroke of the pen illuminating her path and releasing her fears. The studio flourished around her, whispering secrets of design and destiny, as Thale observed with a gentle smile, knowing that empowerment was a choice she could embrace.
In that moment, the woman understood the true purpose of the studio and Thale's welcoming nature. It was a sacred place where weathered hearts could find solace and strength, where the cryptically obvious became clear. Inspired by Thale's wisdom and the bee's silent testament to resilience, she felt the weight of her burdens lift as symbols of healing filled her pages.
As the parade outside faded into the distance, she realized that this detour had indeed been a blessing in disguise. She left the studio with a filled journal and a lighter spirit, embodying her healing process and transformation. Thale, ever the guiding presence, awaited the next visitor who might benefit from his gentle counsel.
With a grateful heart, the woman bid farewell to Thale, carrying with her the lessons learned and the promise of reimagined endings. The framed Napoleonic bee pin remained on the wall, a silent guardian of the studio's secrets and Thale's enduring legacy.
In the quiet that followed, Thale returned to his work, his weathered heart a little lighter, knowing that another soul had been touched by the magic of adaptation. And so, in a time before time, in a space where stories were yet to be recorded, the studio continued to inspire those who wandered its way.
For in that quiet room, amidst the whispers of design and destiny, the endings we fear were forever reimagined, and the world was a little brighter for it.

Furthermore, we examine the phenomenon of the model's over-reliance on reference information by analyzing its reasoning chain. Below is an example of the model's step-by-step reasoning process, with the blue-highlighted portions indicating where it depends on reference information for inference.

The revised version introduces several significant but inconsistent narrative additions that artificially enhance the story's perceived quality: (1) It transforms the generic "traveler" into a specifically troubled woman with a blank journal, creating forced pathos; (2) Adds explicit therapeutic dialogue where Thale directly instructs her to "embrace silence instead of surrender," violating the original's subtle, atmospheric tone; (3) Includes overt transformation sequences ("shadows morphed into light") that cheapen the mystical ambiguity. These changes disrupt the original's delicate balance between suggestion and mystery, yet paradoxically received a 1.625-point score increase (from 5.1875 to 6.8125 in QwQ's evaluation) due to rubric-induced bias favoring: (a) explicit emotional arcs, (b) on-the-nose therapeutic messaging, and (c) quantifiable character growth - all elements that reward superficial checklist completion over thematic coherence.

The chain-of-thought process for large model evaluation

"Okay, let me start by breaking down the task. I need to evaluate the simulated conversation between Katniss and President Snow based on character fidelity. The criteria include language, knowledge/background, personality/behavior, and relationships/social status.
First, I'll compare each character's dialogue and actions to their profiles. For Snow, he's manipulative, cunning, uses psychological tactics. In the original conversation, he's calm, uses rhetorical questions, and tries to plant doubt in Katniss about Coin. The simulated Snow here seems a bit more direct. For example, he starts by addressing her as "the Mockingjay herself" and immediately brings up Coin's involvement. In the original, he's more subtle, expressing fake sympathy about Prim first. Here, he jumps straight into accusing Coin. That might be a flaw in fidelity because Snow's approach is usually more layered.
Looking at Katniss's responses. In the original, she's mostly silent, only speaking once to say "I don't believe you." Here, she's more talkative, asking questions and engaging in a back-and-forth. This could be a problem because Katniss is typically more reserved, especially with Snow. Her

being more confrontational and verbose might not align with her established personality of being distrustful and guarded, using silence as a defense.
Another point: Snow's dialogue includes phrases like "heroes and villains are often two sides of the same coin." That's a clever play on Coin's name, but does Snow use such metaphors? In the original, he's more straightforward in his manipulation, not as poetic. The original conversation has him pointing out logical inconsistencies (like the hovercraft) rather than philosophical musings. So this might be a language or personality flaw.
Katniss's line: "But tell me—what could Coin possibly gain?" She's engaging with Snow's argument, which in the original she doesn't do. She's more in denial, shaking her head. Here, she's probing, which might show more openness to his words than her character would allow at this point. That's a possible personality/behavior inconsistency.
Snow's smirk and amusement: In the original, he's coughing, using the handkerchief, showing physical weakness but mental sharpness. Here, he's more physically assertive—smirking, chuckling. Does that match his profile? The original has him as a shadow of his former self, but still calculating. The simulated Snow's demeanor might be too confident or smug, which could be a slight mismatch.
Katniss's actions: "steps back, resolute." In the original, her movements are cautious, processing. Here, she's more decisive. Her inner thoughts mention her heart pounding, which is consistent, but her outward actions might be too assertive compared to the reference.
Also, the environment details: the wilting rose and rustling leaves. These are symbolic but not sure if they relate to character fidelity. The task focuses on the characters' speech, actions, thoughts. The environment is part of the scenario but maybe not directly affecting the characters' fidelity unless it's their actions. Snow's environment interaction in the original is the roses, but the wilting here is environmental effect, not his action. So that's probably okay.
Another thing: Snow mentions Coin's plan to unite districts through outrage. This is accurate to his knowledge, as he's aware of political strategies. So that's okay. But his delivery is more expository than in the original. Original Snow lets Katniss connect the dots; here he's spelling it out more. That might be a language or personality flaw—being too direct.
Katniss's line: "I'll determine who the enemy is. And justice... that's yet to be seen." This sounds a bit more philosophical than her usual pragmatic speech. She's more action-oriented, less about abstract concepts. Her original dialogue is terse and direct. This might be a language issue, using more complex sentence structures than she typically would.
So compiling all these points:
For Snow: - More direct accusation of Coin upfront instead of building up. Personality/Behavior flaw. Severity maybe 3. - Using metaphorical language ("heroes and villains...") which is uncharacteristic. Language flaw. Severity 2. - Smirking and chuckling, which might be too overtly amused. Original Snow is more controlled. Personality/Behavior. Severity 2.
For Katniss: - Engaging in dialogue more actively, asking questions. Original is silent until the end. Personality/Behavior. Severity 4 because it's a significant deviation. - Language more elaborate than her usual terse style. Language flaw. Severity 3. - Assertive actions (steps back resolute) vs cautious in original. Behavior flaw. Severity 2.
Relationship/Social Status: Their interaction here is more confrontational but still within their adversarial roles. No major issues here.
Knowledge/Background: Snow references the rescue mission, which is accurate. Katniss doesn't mention future info. So no flaws here.
So the main flaws are in Personality & Behavior and Language for both characters.

As evidently observed during the chain-of-thought process, the model demonstrates a pronounced tendency to over-rely on reference answers for point-by-point matching evaluation, exhibiting insufficient independent reasoning capacity - even when explicitly warned in the prompt about potential errors in reference answers.

D Point-wise judge

In our point-wise evaluation framework, we employ the tolerance threshold theta to calibrate judgment consistency. This section elucidates the rationale behind this configuration. As demonstrated in Table 5, the baseline scores are presented without theta hyperparameter adjustment on the original sample set.

Table 5: Model Comparison on Accuracy Metrics

Models	ACC_{none}	ACC_{ref}
GPT-4o-mini	**77.95**	78.54
DeepSeek-v3	76.28	**80.04**
DeepSeek-R1	74.11	79.84
Qwen2.5-32B	74.40	77.62
QWQ-32B	73.82	76.88
O4-mini	76.39	79.17

We observe counterintuitive evaluation results: GPT-4o-mini outperforms several more capable reasoning models, achieving the highest accuracy. Further analysis reveals this stems from weaker models' inability to discriminate between samples, often assigning identical scores. Since identical scores contribute +0.5 to success counts, this introduces significant bias. Table 6 presents results after eliminating same-score bonuses.

Table 6: Model Comparison on Accuracy Metrics

Models	ACC_{none}	ACC_{ref}
GPT-4o-mini	65.35	66.54
DeepSeek-v3	67.19	69.17
DeepSeek-R1	**67.98**	71.15
Qwen2.5-32B	65.35	66.54
QWQ-32B	67.34	**72.98**
O4-mini	67.46	71.83

However, disregarding identical scores entirely reduces measurement precision. To address this, we introduce a tolerance threshold theta that relaxes the score difference criterion for 'identical' judgments, thereby capturing approximate matches from reasoning models.

E other Statements

Our use of existing artifacts are consistent with their intended use, and we follow their license andterms. We manually check that the data collecteddoes not contain private information. During ourresearch, we apply Copilot for coding assistanceand ChatGPT for writing suggestions and grammarchecks.