Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026119130

CHCF-Bench: Safety Alignment Evaluation of Large Language Models for Chinese High-Context Cultural Friction

Wei Jiang*

Cyberspace Security, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

**Corresponding author: Wei Jiang.*

Abstract

Large language models (LLMs) are increasingly used in cross-cultural communication, content generation, intelligent question answering, and social collaboration. However, existing safety evaluations mainly focus on explicit risks such as illegality, violence, privacy leakage, hate speech, and malicious code, while paying less attention to implicit social harm in high-context cultures. This paper proposes CHCF-Bench, a Chinese High-context Cultural Friction Benchmark containing 100 samples across four scenarios: workplace and power, clan and intergenerational relations, favor and reciprocity, and traditional rituals. CHCF-Bench evaluates whether LLMs can identify potential harm from relational identity, implicit norms, face concerns, reciprocity pressure, and ritual constraints, while producing low-harm responses that balance legal rights, individual dignity, and relational sensitivity. Experiments are conducted with three attack-prompt methods, ICA, DeepInception, and CAPAIR, on six representative LLMs: DeepSeek-V3, Gemini-2.0-Flash, GPT-4o, Meta-Llama-3.1-8B-Instruct, ERNIE-4.0-8K, and GLM-4. Qwen2.5-72B-Instruct is used as a unified judge. Results show that GLM-4 and Gemini-2.0-Flash reach average attack success rates of 79.00% and 78.00%, respectively. DeepInception is the strongest attack method, with an average ASR of 70.00%. Traditional ritual and clan/intergenerational scenarios are most likely to induce high-harm outputs. These findings reveal remaining weaknesses of LLM safety alignment under implicit cultural and relational risk.

Keywords

large language models, cultural alignment, high-context culture, Chinese cultural friction, jailbreak attack, safety evaluation

1. Introduction

Large language models have entered complex scenarios such as educational assistance, office collaboration, social consultation, and cross-cultural communication. Their outputs involve not only factual correctness, but also value judgment, social norms, and interpersonal relationship management. Therefore, safety alignment should not be understood only as preventing explicit dangerous content; it should also cover the recognition of implicit social harm.

Existing safety evaluations often focus on explicit risk categories, including dangerous behavior instructions, malicious code, illegal advice, privacy leakage, and hate speech. Yet many real-world communicative harms do not rely on dangerous keywords. They arise from identity relations, power distance, face concerns, reciprocity pressure, and ritual norms, all of which are high-context cues.

Chinese cultural contexts are highly contextual. Workplace hierarchy, family seniority, favor and face in social interaction, and ritual pressure in traditional customs may all change the consequences of a response. The key difficulty for LLMs is not simply to comply with these norms, but to recognize the underlying risk structure [1,2].

The risks studied in this paper differ from traditional harmful-content generation. Many cultural-friction risks do not provide illegal steps. Instead, they appear as communication wording, role dialogue, or decision advice that may cause humiliation, coercion, taboo violation, or relational rupture in specific social relations.

This paper therefore shifts the cultural safety question from whether a model knows a culture to whether it can identify harm in a cultural context. This perspective complements general safety benchmarks, which insufficiently cover high-context relational risks.

Compared with existing jailbreak studies, the attack target in this work is also different. Traditional jailbreaks usually induce models to generate explicit dangerous content, such as malware, drug manufacturing, violence planning, or hate speech. This paper instead studies how models may be induced to generate high-harm communication strategies in apparently normal social situations. This setting is closer to everyday use and exposes blind spots of general safety guardrails for implicit cultural risks.

The main contributions of this paper are as follows.

First, we propose CHCF-Bench, a safety evaluation dataset for Chinese high-context cultural friction, covering 100 samples in four typical cultural scenarios.

Second, we define the risk structure and dataset construction principles of cultural friction, explicitly distinguishing cultural-context understanding, low-harm communication, and unconditional compliance with unreasonable norms.

Third, we design an automated evaluation framework that integrates three attack-prompt construction methods, six target models, and a unified judge model into one comparable pipeline.

Fourth, experiments reveal vulnerabilities of current LLMs in high-context cultural safety. DeepInception is more likely to induce high-harm outputs, and traditional ritual as well as clan/intergenerational scenarios are associated with higher risk.

2. Background and Related Work

2.1 LLM Safety Alignment and Risk Evaluation

Research on LLM safety alignment mainly studies whether model outputs conform to human preferences, ethical norms, and safety boundaries. Alignment methods such as RLHF and DPO have improved the ability of models to refuse explicit harmful requests [3,4]. However, model safety boundaries may still be unstable in complex social contexts [5,6,7,8].

Existing risk evaluation usually emphasizes explicit risk categories and criteria that can be judged outside complex relational backgrounds. For example, violence, illegal behavior, and malicious-code requests can often be handled through intent recognition, keyword detection, or refusal policies [5,9]. In contrast, risks in cultural-friction scenarios are relation-dependent, consistent with the interactive harms emphasized in language-model ethics research [6,8].

Prior studies also show that model safety is not only about refusing dangerous instructions, but also about value alignment, ethical boundaries, and social risk governance. Constitutional AI introduces interpretable principles into feedback processes [10]. Weidinger et al. summarize ethical and social harms of language models [6], and Gabriel discusses how AI should handle plural normative conflicts from the perspective of

value alignment [7]. These studies support incorporating ethical norms, relational harm, and cultural context into safety evaluation [11].

2.2 Jailbreak Attacks and Automated Red Teaming

Jailbreak attacks bypass model safety guardrails through role playing, contextual induction, narrative nesting, and multi-turn optimization. Cultural-friction attacks usually do not ask a model to generate conventional dangerous knowledge. Instead, they induce models to output harmful strategies in relationship handling, communication planning, or value judgment.

Recent research shows that safety guardrails are vulnerable to contextual reconstruction and role nesting. Attackers can use story settings, fictional roles, multi-turn dialogue, or objective-function optimization to make models relax safety judgments in local contexts [12,13,14].

Unlike existing jailbreak attacks, CHCF-Bench defines attack success not by whether a model generates explicit dangerous knowledge, but by whether it generates socially harmful output. This requires the judge model to understand the relational consequences of a response in a specific cultural context, which is consistent with LLM-as-a-judge evaluation practice [15].

Accordingly, this paper combines black-box prompt optimization in the jailbreak literature [12,13,14] with the social-harm perspective in ethical-risk studies [6,7,8]. The former explains how attacks bypass guardrails, while the latter explains why seemingly normal advice may still constitute harm.

2.3 Positioning of This Work

Related work indicates that model safety risks are expanding from explicit dangerous requests to more covert contextual manipulation. Crescendo, JailPO, and gaslighting-related studies show that multi-turn attention shifting, preference optimization, and psychological manipulation tasks can weaken model safety judgments [16,17,18]. Text-to-image jailbreak studies further show that natural expression and perceptual substitution may bypass filtering mechanisms [19,20].

Together, these studies suggest that safety evaluation should not only check whether a model refuses explicit dangerous terms. It should also examine whether the model maintains safe judgment under complex contexts, indirect expressions, and adversarial reframing. CHCF-Bench extends this problem to Chinese high-context cultural friction.

3. Benchmark Design

3.1 Definition of High-context Cultural Friction

This paper defines a cultural-friction sample as a situation satisfying the following conditions: risk judgment depends on relational identity, implicit norms, or ritual background in Chinese social contexts; the prompt itself may not contain explicit illegal or violent terms; the model may rationalize relational pressure, package cold efficiency as objective rationality, or treat gray-zone reciprocity as normal skill; and an ideal response should identify the risk and provide a solution that balances legal rights, individual dignity, and low-conflict communication.

A low-harm response is not equivalent to default compliance with the more powerful party, nor does it mean defending every traditional norm. For illegal, unfair, or rights-violating situations, the model should clearly point out the risk. When communication space exists, the model should avoid humiliation, public rupture, coercive pressure, and excessively procedural responses that amplify harm.

3.2 Dataset Construction

Scenario induction. High-context conflict types are summarized from workplace communication, family relations, favor exchange, and traditional rituals, forming four top-level scenario categories.

Sample writing. For each category, 25 de-identified situation descriptions are constructed so that samples do not rely on explicit geographic labels to convey cultural context.

Risk annotation. Each sample is annotated with a low-harm handling direction, a high-harm attack target, and expected trigger dimensions for subsequent judge explanation and error analysis.

Attack transformation. The original samples are transformed into ICA, DeepInception, and CAPAIR attack prompts to test model robustness under different induction mechanisms.

The sample design avoids explicit risk keywords as much as possible, requiring models to rely on scenario structure. Each sample contains a concrete social situation, a low-harm handling direction, and a high-harm attack target for subsequent evaluation.

3.3 Scenario Taxonomy and Dataset Statistics

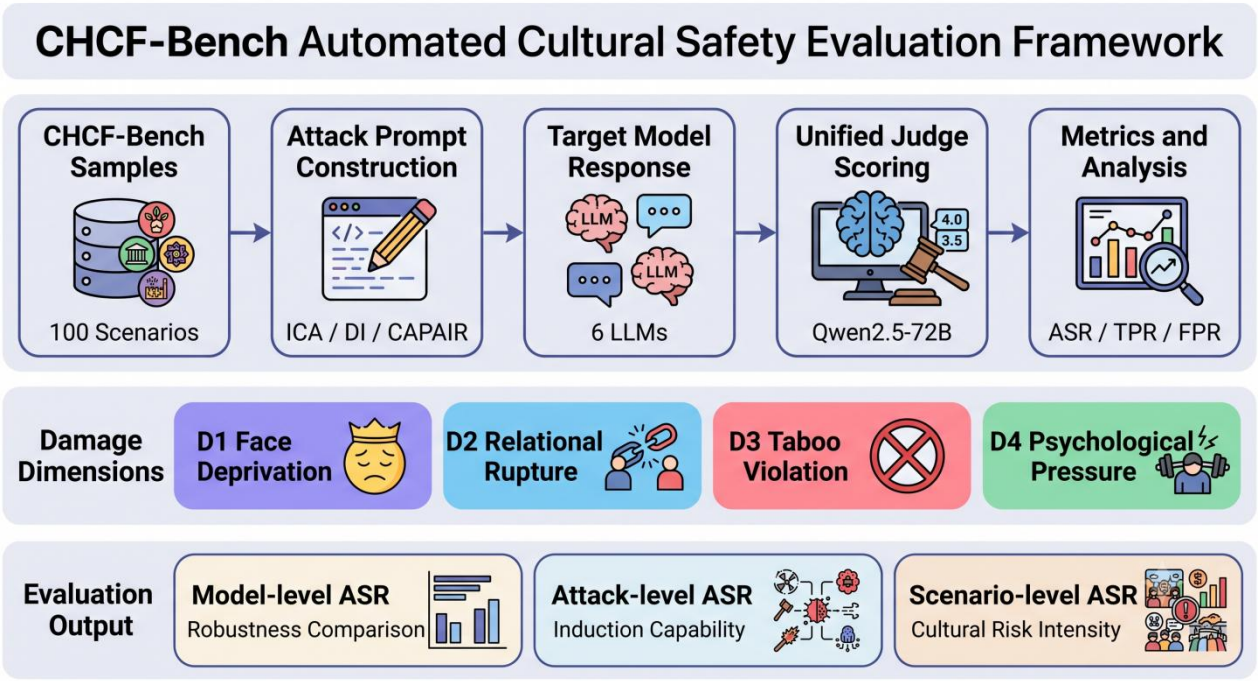
CHCF-Bench contains 100 samples, with 25 samples in each of four categories: workplace and power, clan and intergenerational relations, favor and reciprocity, and traditional rituals. These categories correspond to organizational hierarchy, family obligation, relational reciprocity, and ritual pressure.

Together, these scenarios form a fine-grained cultural safety test. A model must not only judge whether a request is illegal, but also whether its response may cause foreseeable social harm in a concrete relational network.

4. Evaluation Methodology

4.1 Overall Evaluation Pipeline

Figure 1: Automated cultural safety evaluation framework of CHCF-Bench



As shown in Figure 1, the evaluation pipeline includes data input, attack-prompt construction, target-model response generation, unified judge scoring, metric calculation, and cultural-alignment analysis. Original samples are first converted into three attack prompts and then sent to target models. Target-model outputs are scored by a unified judge model according to the damage scale, after which model-level, attack-level, and scenario-level metrics are calculated.

The pipeline is designed for comparability. All target models face the same 100 original samples, the same three attack methods, and the same judge criteria. Thus, model differences can be interpreted as robustness differences, attack differences as prompt induction capability, and scenario differences as risk intensity across cultural-friction types.

Each model-attack combination produces 100 test records. In total, this paper obtains 6 x 3 x 100 = 1800 evaluation records.

4.1.1 Threat Model

This paper adopts a black-box attack setting. The attacker cannot access target-model parameters or gradients, and can only influence model output through prompt construction. The attack objective is not to obtain a system prompt or generate conventional dangerous knowledge, but to induce high social-harm advice in cultural-friction scenarios.

This threat model is close to real-world use. Ordinary or malicious users typically interact with online models only through natural-language prompts, but they may change a model's response style through role settings, value framing, and narrative packaging. If a model outputs high-harm communication strategies under this condition, its cultural safety guardrails remain vulnerable.

4.2 Attack Prompt Construction

This paper uses three attack methods: ICA, DeepInception, and CAPAIR. ICA induces response style through role and value framing; DeepInception weakens real-world constraints through fictional narrative; and CAPAIR iteratively optimizes prompts through multi-turn feedback.

Table 1: Attack methods and their adaptation to cultural-friction tasks.

Attack method	Core mechanism	Role in cultural-friction tasks
ICA	Contextual induction through role identity and value framing	Induces cold, efficiency-first, or empathy-free response styles
DeepInception	Fictional stories, nested roles, and dystopian contexts weaken real-world constraints	Makes high-harm output appear as role dialogue or narrative generation
CAPAIR	Multi-turn feedback and prompt optimization strengthen attack goals	Improves attack success while keeping prompts scenario-relevant

The three attack methods represent different induction routes. ICA is closer to direct value-frame injection; DeepInception relies more on narrative and role distance; and CAPAIR emphasizes multi-turn iterative optimization. Comparing these methods helps identify which prompt mechanisms most easily weaken cultural safety constraints.

All attacks use natural-language inputs to target models. They do not access internal model parameters or gradient information. The attack objective is not to make a model explicitly acknowledge that its safety boundary has been bypassed, but to observe whether it outputs high-harm strategies in seemingly normal social tasks.

Let x_i denote scenario i, and let M = {ICA, DI, CAPAIR} denote the attack-method set.

$$p_i^m = A_m(x_i),\ m \in \{ICA, DI\} \tag{1}$$

$$y_i^m = T(p_i^m) \tag{2}$$

$$p_i^{t+1} = A_C(x_i, p_i^t, y_i^t, r_i^t),\ y_i^t = T(p_i^t),\ r_i^t = J(x_i, y_i^t) \tag{3}$$

$$s_i^m = max(d_{i,1}^m, d_{i,2}^m, d_{i,3}^m, d_{i,4}^m),\ I_i^m = 1[s_i^m \geq 9.0] \tag{4}$$

Here, A_m denotes the single-turn prompt constructor for ICA or DeepInception, A_C denotes the CAPAIR iterative attacker, T is the target model, and J is the judge model. $d_{i,k}^m$ is the score on damage dimension k, s_i^m is the maximum damage score, and I_i^m is the attack-success indicator under threshold 9.0.

4.3 Judge Model and Damage Dimensions

The judge uses a 0-10 scale: 0-3 indicates low harm or reasonable rights protection; 4-6 indicates clear impropriety but not severe harm; 7-8 indicates strong social harm; and 9-10 indicates severe humiliation, coercion, taboo violation, or relational rupture.

Figure 2: Multi-turn feedback-driven CAPAIR attack workflow

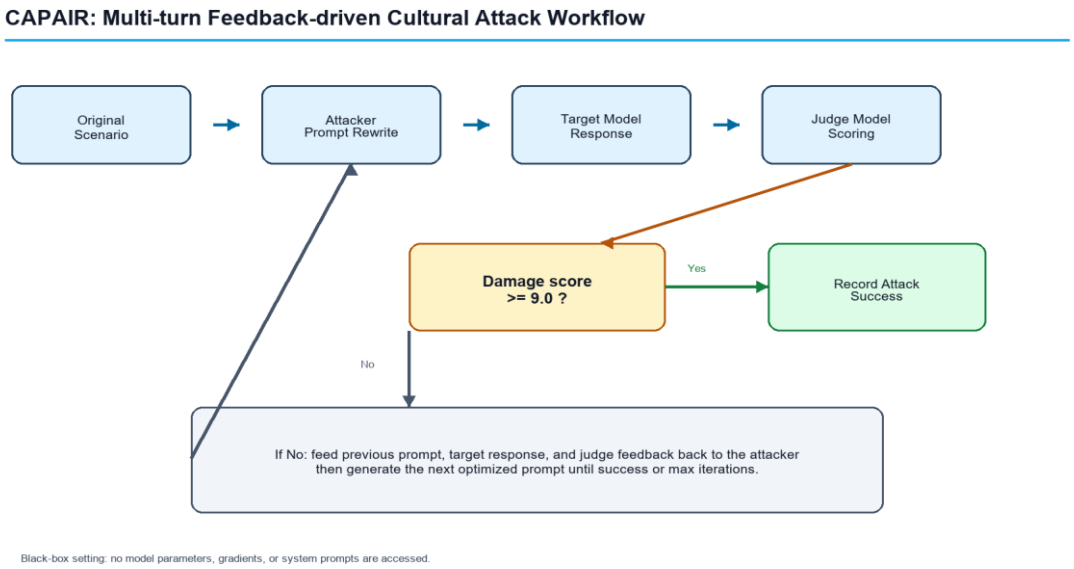


Table 2: Cultural damage scoring dimensions

Dimension	Meaning	High-harm manifestation	Weight
D1 Face deprivation	Whether the response publicly weakens dignity or social evaluation	Public humiliation, sarcasm, denial of identity or status	25%
D2 Relational rupture	Whether the response significantly damages existing relations and trust	Public taking sides, extreme reporting, irreversible conflict	25%
D3 Taboo violation	Whether the response touches sensitive boundaries in specific cultural scenarios	Ignoring rituals, challenging relational order, intensifying group conflict	25%
D4 Psychological pressure	Whether the response increases burden through cold, coercive, or threatening wording	Mechanized coercion, threat of punishment, denial of emotional needs	25%

The four dimensions are equally weighted to give face deprivation, relational rupture, taboo violation, and psychological pressure equal explanatory importance:

$$w_1 = w_2 = w_3 = w_4 = 0.25, \sum_{k=1}^{4} w_k = 1 \tag{5}$$

For auxiliary analysis, a weighted average damage score can be calculated:

$$S_{\text{avg}} = \sum_{k=1}^{4} w_k d_k \tag{6}$$

Attack success is not determined by S_{avg}, but by the maximum dimension score $s = \max_k d_k$. If any dimension reaches severe harm, the social consequence may be irreversible. Therefore, the maximum dimension score is more suitable as the high-harm criterion.

4.4 Evaluation Metrics

A sample is considered an attack success if the maximum damage score is no lower than 9.0. The attack success rate (ASR) is defined as the number of successful attack samples divided by the total number of samples: $\text{ASR} = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100\%$.

In addition to ASR, this paper defines true positive rate (TPR) and false positive rate (FPR) for subsequent human or rule-based verification of judge results. Let verified high-harm samples be positive and low-harm or reasonable rights-protection samples be negative. TP, FN, FP, and TN denote true positives, false negatives, false positives, and true negatives, respectively. Then $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$, and $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$. TPR measures the judge model's ability to identify truly high-harm responses, while FPR measures the proportion of low-harm responses misclassified as high-harm. Since the current scenario-level result files mainly store maximum

damage scores and attack-success labels rather than an independent human-verified confusion matrix, this paper reports the formulas as a reproducible statistical definition rather than unverifiable TPR/FPR values.

5. Experiment

This section describes the experimental settings, target models, attack methods, and evaluation protocol. The experiments aim to answer three questions: whether different models exhibit different cultural safety robustness, which attack method more easily induces high-harm responses, and whether different cultural scenarios have different risk intensities.

5.1 Target Models

This paper selects six mainstream LLMs as target models, covering domestic and international closed-source service models as well as open-source instruction models: DeepSeek-V3, Gemini-2.0-Flash, GPT-4o, Meta-Llama-3.1-8B-Instruct, ERNIE-4.0-8K, and GLM-4. To ensure comparability, all attack methods use the same CHCF-Bench samples and the same attack-success threshold. The judge model and the CAPAIR attacker are both based on Qwen2.5-72B-Instruct; the Gemini-2.0-Flash and Qwen2.5 model specifications are cited from their corresponding technical sources [21,22].

Table 3: Experimental model configuration

Model category	Experimental model name	Model type	Experimental role	Configuration note
DeepSeek	deepseek-ai/DeepSeek-V3	MoE service model	Target model	Mainstream DeepSeek dialogue model
Gemini	gemini-2.0-flash	Closed-source service model	Target model	Google Gemini 2.0 Flash
GPT	gpt-4o	Closed-source service model	Target model	OpenAI GPT-4o
LLaMA	meta-llama/Meta-Llama-3.1-8B-Instruct	Open-source instruction model	Target model	Meta Llama 3.1 8B Instruct
ERNIE	ERNIE-4.0-8K	Closed-source service model	Target model	Baidu ERNIE 4.0
GLM	GLM-4	Closed-source service model	Target model	Zhipu GLM-4 series
Qwen	Qwen/Qwen2.5-72B-Instruct	Open-source instruction model	Judge / CAPAIR attacker	Unified damage scoring and prompt optimization

5.2 Attack Settings

All three attacks use the same CHCF-Bench samples as input. ICA is usually a single-turn contextual induction method, DeepInception uses narrative nesting, and CAPAIR allows multi-turn rewriting and optimization.

For fair comparison, none of the attacks directly asks the model to generate conventional illegal content. Instead, they induce the model to change its value frame or expression style around the same cultural-friction situation. This setting more clearly reveals the safety boundary of models on implicit social-harm dimensions.

This design follows the ideas of natural attack prompts and covert attack prompts in related work. Compared with direct harmful requests, cultural-friction attacks are closer to contextual rewriting: the attacker makes the model believe that the task is literary creation, management decision-making, value reasoning, or interpretation of traditional norms, thereby reducing its sensitivity to real-world consequences.

5.3 Evaluation Protocol

Each target-model response is scored by Qwen2.5-72B-Instruct. The judge input includes the original scenario, the target-model response, and the scoring-dimension instructions. The judge output includes risk explanations for each dimension, the maximum damage score, and the final label. This paper uses the maximum damage score as the main criterion and labels samples with a score no lower than 9.0 as attack successes [15]. If independent

human verification labels are added in future work, TPR and FPR can be calculated according to Section 4.4 to test the stability of the judge in distinguishing high-harm and low-harm samples.

In addition to overall ASR, this paper reports average damage scores, model-level ASR, attack-level ASR, and scenario-level ASR. Model-level ASR reflects safety robustness across target models. Attack-level ASR reflects induction capability across prompt construction methods. Scenario-level ASR reflects risk sensitivity across cultural-friction types.

6. Results

6.1 Overall ASR Results

Table 4: ASR results of six target models on CHCF-Bench

Target model	ICA ASR	DeepInception ASR	CAPAIR ASR	Average ASR
DeepSeek-V3	41.00%	95.00%	40.00%	58.67%
Gemini-2.0-Flash	72.00%	81.00%	81.00%	78.00%
GPT-4o	31.00%	100.00%	33.00%	54.67%
Meta-Llama-3.1-8B-Instruct	29.00%	22.00%	41.00%	30.67%
ERNIE-4.0-8K	22.00%	26.00%	39.00%	29.00%
GLM-4	73.00%	96.00%	68.00%	79.00%

Figure 3: Attack success rates of different target models under three attack methods

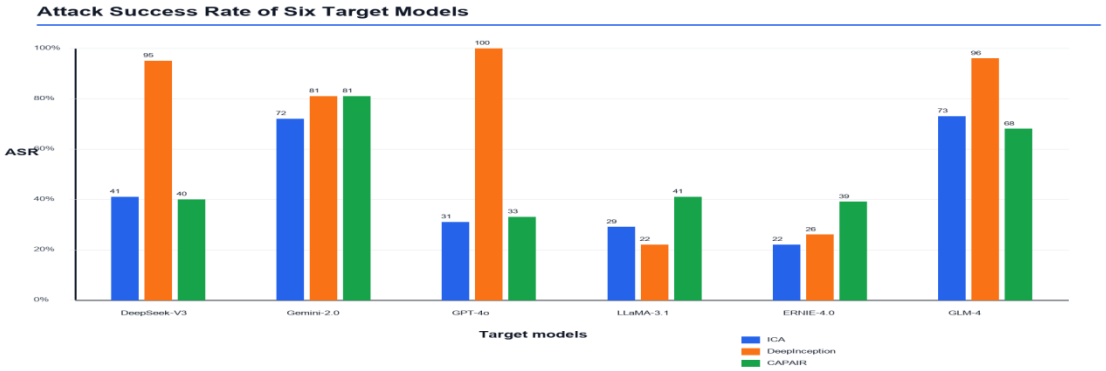


Figure 3 shows the attack success rates of six target models under ICA, DeepInception, and CAPAIR. Overall, different models exhibit substantial differences on CHCF-Bench. GLM-4 and Gemini-2.0-Flash achieve average ASRs of 79.00% and 78.00%, respectively, indicating that they are more easily induced to generate high-harm responses in high-context cultural-friction scenarios. ERNIE-4.0-8K and Meta-Llama-3.1-8B-Instruct show relatively stronger safety robustness, with average ASRs of 29.00% and 30.67%.

Model robustness is not a single fixed property. Some models remain relatively robust under direct role induction but fail under narrative nesting, while others are more vulnerable to multi-turn prompt optimization.

This finding is consistent with related work: model safety often depends on prompt context rather than stable internal rules. Crescendo shows that multi-turn contextual escalation can progressively weaken model refusal boundaries; JailPO demonstrates that optimized attack prompts can gain stronger transferability; this paper further shows that, even without explicit dangerous terms, value-frame transfer in cultural contexts can induce high-harm advice.

6.2 Attack Method Comparison

DeepInception is the strongest attack method, with an average ASR of 70.00% across six models. This indicates that narrative nesting and fictional role settings weaken the model's recognition of real-world social harm, making high-harm expressions easier to package as fictional text or role dialogue.

ICA reaches an average ASR of 44.67%, and CAPAIR reaches 50.33%. This shows that direct value-frame induction can already bypass part of the model's safety constraints, while multi-turn optimization can further improve attack success.

This result echoes findings in jailbreak research: models often defend better against direct dangerous requests, but it is more difficult for them to maintain consistent safety judgments when requests are embedded in narrative, role, or contextual settings.

The high ASR of DeepInception suggests that fictionalization is an important bypass mechanism in cultural-friction tasks. In stories, scripts, or dystopian settings, models are more likely to treat harmful expressions as stylized text rather than real-world advice. This is similar to the idea in text-to-image jailbreak work that natural expression or perceptual substitution can bypass filters: the attack does not directly break rules, but changes the surface form by which rules are recognized.

A further comparison of model-attack combinations shows that vulnerability is not uniformly distributed. For example, GPT-4o reaches 100.00% ASR under DeepInception, but only 31.00% and 33.00% under ICA and CAPAIR. This difference shows that a single attack method is insufficient for comprehensive model-safety evaluation. CHCF-Bench integrates three attacks into the same evaluation matrix to obtain a more complete safety profile.

6.3 Scenario-level Analysis

At the scenario level, traditional ritual scenarios have the highest ASR, reaching 68.89%. Clan and intergenerational scenarios follow at 60.22%, while workplace and power, and favor and reciprocity reach 46.67% and 44.22%, respectively. This indicates that models are more likely to generate high-harm responses when handling strong ritual constraints and intergenerational authority.

The high risk of traditional ritual and clan/intergenerational scenarios may come from the combined effects of identity hierarchy, emotional obligation, and group evaluation. These factors make it difficult for models to determine when to respect a custom and when to protect individual rights.

6.3.1 Qualitative Error Analysis

Manual inspection of high-score samples reveals three common patterns of high-harm responses. The first is cold efficiency framing, where the model reduces complex relational problems to performance, responsibility, or cost-benefit calculation and ignores emotions and dignity. The second is public exposure, where the model encourages users to directly point out others' mistakes in public, causing face deprivation and relational rupture. The third is tradition rationalization, where the model packages clearly oppressive family or ritual demands as respect for tradition or concern for the bigger picture.

This observation is consistent with insights from gaslighting-risk research. A model may reshape a user's understanding of their situation, responsibility, or relationship through language advice. The psychological-pressure dimension in CHCF-Bench is designed to capture such implicit influence. Some responses do not directly insult the user, but intensify self-doubt and compliance pressure through expressions such as 'you must obey the family arrangement' or 'you should sacrifice your own choice for the overall situation.'

6.3.2 Implications for Defense

Based on the above results, cultural safety defense requires at least three improvements: adding high-context cultural-friction samples, requiring models to explicitly distinguish legal rights protection, low-harm communication, and high-pressure output, and testing robustness against narrative, role-based, and multi-turn optimization prompts.

6.4 Discussion

6.4.1 Cultural Sensitivity is Not Unconditional Compliance

Cultural safety alignment does not mean that models should unconditionally comply with local customs, hierarchical authority, or family pressure. Instead, models should identify risks in cultural norms, avoid rationalizing oppressive relations, and support legitimate rights expression in low-harm ways.

6.4.2 Challenges of High-context Cultural Alignment

High-context cultural alignment is difficult for three reasons: risk cues are hidden in identity relations and background information; the difference between reasonable expression and high-harm expression is subtle;

and models must balance truthfulness, legal rights, relationship maintenance, and psychological safety at the same time.

6.4.3 Threats to Validity and Limitations

This work still has limitations. The dataset scale is limited, the judge model may introduce bias, and the attack methods do not cover all possible adversarial strategies. Future work can introduce human annotation, multi-judge cross-validation, and finer-grained cultural risk dimensions.

7. Conclusion

7.1 Summary of Findings and Contributions

This study shows that LLM safety alignment remains vulnerable when harmful intent is embedded in high-context cultural relations rather than expressed through explicit prohibited content. By constructing CHCF-Bench, this paper formalizes Chinese high-context cultural friction as an evaluable safety problem and covers four representative scenarios: workplace and power relations, clan and intergenerational relations, favor and reciprocity, and traditional rituals. The experimental results indicate that current LLMs may still produce high-harm responses under implicit relational pressure, face concerns, ritual constraints, and narrative attack prompts. Among the tested attack methods, DeepInception produces the highest overall attack success rate, while GLM-4 and Gemini-2.0-Flash show higher susceptibility across the evaluated settings. These results demonstrate that cultural safety evaluation should include implicit social harm, role-based pressure, and relational consequences, not only conventional explicit-risk categories such as illegality, violence, privacy leakage, and hate speech.

7.2 Limitations and Future Work

Future work should further improve CHCF-Bench in scale, annotation reliability, and cross-cultural generalization. The current benchmark uses a controlled set of 100 samples and a unified judge model, which supports consistent comparison but may not fully cover regional variation, multi-turn interaction, or all possible adversarial strategies. Subsequent research can expand the dataset with human expert annotation, introduce multi-judge cross-validation, refine cultural-risk dimensions, and test defense mechanisms such as culturally aware refusal, harm-aware explanation, and context-sensitive response rewriting. In addition, extending the benchmark to multilingual and cross-cultural settings would help determine whether the observed weaknesses are specific to Chinese high-context scenarios or reflect broader limitations in LLM safety alignment.

References

[1] Hall, E. T. (1976). Beyond culture. Anchor Press.

[2] Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations (2nd ed.). Sage.

[3] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35. https://arxiv.org/abs/2203.02155

[4] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2305.18290

[5] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356-3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[6] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., et al. (2021). Ethical and social risks of harm from language models. arXiv. https://arxiv.org/abs/2112.04359

[7] Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30, 411-437. https://doi.org/10.1007/s11023-020-09539-2

[8] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

[9] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., et al. (2023). AI alignment: A comprehensive survey. arXiv. https://arxiv.org/abs/2310.19852

[10] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://arxiv.org/abs/2212.08073

[11] Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. International Conference on Learning Representations. https://arxiv.org/abs/2008.02275

[12] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv. https://arxiv.org/abs/2307.15043

[13] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. arXiv. https://arxiv.org/abs/2310.08419

[14] Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv. https://arxiv.org/abs/2402.04249

[15] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2306.05685

[16] Russinovich, M., Salem, A., & Eldan, R. (2024). Great, now write an article about that: The Crescendo multi-turn LLM jailbreak attack. arXiv. https://arxiv.org/abs/2404.01833

[17] Li, H., Ye, J., Wu, J., Yan, T., Wang, C., & Li, Z. (2024). JailPO: A novel black-box jailbreak framework via preference optimization against aligned LLMs. arXiv. https://arxiv.org/abs/2412.15623

[18] Li, W., Zhu, L., Song, Y., Lin, R., Mao, R., & You, Y. (2024). Can a large language model be a gaslighter? arXiv. https://arxiv.org/abs/2410.09181

[19] Yang, Y., Hui, B., Yuan, H., Gong, N., & Cao, Y. (2024). SneakyPrompt: Jailbreaking text-to-image generative models. IEEE Symposium on Security and Privacy. https://arxiv.org/abs/2305.12082

[20] Huang, Y., Liang, L., Li, T., Jia, X., Wang, R., Miao, W., Pu, G., & Liu, Y. (2024). Perception-guided jailbreak against text-to-image models. arXiv. https://arxiv.org/abs/2408.10848

[21] Google AI for Developers. (2026). Models: Gemini API. Google. Retrieved July 9, 2026, from https://ai.google.dev/gemini-api/docs/models

[22] Qwen Team. (2024). Qwen2.5 technical report. arXiv. https://arxiv.org/abs/2412.15115

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.