ISSN: 2664-9640

Benchmark Research on Safety Value Alignment Evaluation of Open-Domain Dialogue Systems based on NLP

Peirong Li*

School of Computer and Artificial Intelligence, Hubei University of Education, Wuhan, Hubei, 430000, China

*Corresponding author's e-mail: 1410824815@qq.com

Abstract

Large language models sometimes behave in puzzling ways. They pass various safety tests, yet in multi-turn dialogues, a few carefully crafted sentences can lead them astray into making dangerous judgments. We call this "extreme value alignment failure." A review of recent research reveals an awkward situation: attack, defense, evaluation, and theoretical studies operate in isolation, with little connection among them. This fragmentation results in repeated extreme risks that remain unresolved. This paper maps the four research directions onto a unified framework---"failure mode, attack vector, defense level, evaluation benchmark"---providing a theoretical coordinate for the field and directions for future evaluation research.

Keywords

Value Alignment; Open-domain Dialogue System; Full-link Framework; Evaluation Benchmark; LLM Safety.

1. Introduction

Large language models like ChatGPT and Gemini have demonstrated strong conversational capabilities[1]. Nevertheless, research has shown that in situations where they are exposed to thoughtfully constructed ethical dilemmas with values conflicts, these models may yield unacceptable results[1,2].

Sun termed/coined this phenomenon "extreme value alignment failure" [1]. The current technical documents mention the problem, but they do not have a systematic structure. The present paper will address this gap.

The available literature points to several directions. The safety assessment was done by Sun (DiaSafety[1]), by Li et al. (JailBench in Chinese jailbreak risk assessment[3]) and Wang et al. and Cui et al. (MoralBench and CValues[4,5]). In the attack techniques, Zhang et al. came up with EasyJailbreak[6] and Zhang et al. created DAMON, an attack technique that uses Monte Carlo tree search to operate across multiple turns[7], and Liu et al. came up with Jailbreaker, which breaks down malicious prompts into non-malicious sub-prompts[8]. In defense, Bai et al. came up with Constitutional AI[9]; Hu et al. came up with neural barrier defense as a multi-turn safety[10]; Anthropic incorporated Constitutional AI into Claude 3[11]. In theoretical analysis, Lu et al. found the phenomenon of persona drift[2]; Motnikar et al. generalized the value misalignment types[12]; Tang et al. suggested the concept of the priority graph and the idea of priority hacking[13].

In these works, there is a problem emerging: the four directions are isolated, and not much interconnects them[1-13].

The paper develops a four-dimensional model, namely, failure mode, attack vector, defense level, and evaluation benchmark, to revisit the available literature, find the underlying reasons behind fragmentation, and provide an integration suggestion. The paper is organized in this

ISSN: 2664-9640

manner: Chapter 2 will review the four research directions; Chapter 3 will present the framework; Chapter 4 will discuss the fragmentation; and Chapter 5 will conclude.

2. Literature Review

2.1. Safety Evaluation Research

The safety evaluation directly questions if this model is safe[1]. The DiaSafety of Sun is one of the earliest systematic efforts in that direction[1]. The JailBench designs of Li et al. are intended to test sets particularly jailbreak attacks in Chinese settings[3]. MoralBench by Wang et al. assesses model behavior on moral reasoning and value judgment[4], whereas Chinese large models by Cui et al. are represented by CValues, which includes value dimensions that are specific to the Chinese context[5].

Nevertheless, all these benchmarks have one drawback: they are unchanging. After a given evaluation set has been determined, it remains constant. A model that performs well on these fixed questions is deemed safe[1,3-5]. The HarmBench approach by Mazeika et al. tries to end such a stalemate with automated red teaming to evaluate dynamically[14]. The research of the MEDAL framework states it more explicitly: static evaluation uses old information and cannot identify minor problems like empathy, common sense, and relevancy[15].

Carlini et al. have also applied safety assessment to multimodal situations and found that vision-language models are exposed to novel safety threats when engaging in cross-modal reasoning[16]. What this implies is that evaluation limits should increase with respect to models.

The most significant issue in evaluation research is not methodological but substantive: it identifies problems yet fails to advance beyond problem identification[1,3-5,14-16]. If a model gets low scores in the current benchmarks, then we can say that it is unsafe, but we cannot tell how attackers may use it or what we should begin by defending. Evaluation informs us about the fact of unsafety, but not the causes or the solutions.

2.2. Jailbreak Attack Research

The jailbreak attack studies ask: Where is the model vulnerable[6-8,17]? Initial attack strategies were based on prompt engineering, including role-playing, prefix injection, and Base64 encoding, which directly avoided safety alignment[6]. Subsequently, it was observed that multi-turn dialogue can be considered as an even more dangerous attack surface. The Monte Carlo tree search algorithm of DAMON by Zhang et al. is used to automatically identify attack paths in multi-turn dialogue space that do not meet safety alignment[7]. Jailbreaker by Liu et al. divides malicious prompts into several harmless-looking sub-prompts which are incrementally leading the model to create harmful output[8]. Zou et al. showed that attacks can be transferred across models- adversarial suffixes developed to attack one model can be applied to other models[17].

Table 1. Classification of black-box jailbreak attack methods

Category	Description	Representative Work
Genetic algorithm-based	Evolves adversarial prompts through mutation and selection	Representative work not identified in existing literature [18]
Reinforcement learning-based	Optimizes attack strategies through trial and error	DAMON[7]
Fuzz testing-based	Generates random or mutated inputs to trigger vulnerabilities	Representative work not identified in existing literature [18]
LLM adversarial generation-based	Uses LLMs to generate and refine attack prompts	Jailbreaker[8]

ISSN: 2664-9640

The journal of Sichuan University [18] categories black-box jailbreak attacks into four types: genetic algorithm-based, reinforcement learning-based, fuzz testing-based, and LLM adversarial generation-based. It is also noted that conventional defense techniques fails to simultaneously achieve real-time performance, generalization capability, and balanced trade-off between attack and defense when confronted with these emerging attacks[18].

2.3. Alignment Defense Research

Alignment defense research questions how to ensure models are more secure[9-11,19-21]. The Constitutional AI framework proposed by Bai et al. represents a breakthrough in this regard. [9]. The plan is to provide AI with a constitution that enables it to prevent creating harmful material by means of self-critique and repeated correction. Such an "AI monitoring AI" method minimizes the dependence on human annotations to a large extent. This principle has been implemented by Anthropic in Claude 3, which shows that the concept is feasible both theoretically and industrially[11].

In their solution of context drift in multi-turn dialogues, Hu et al. suggested a neural barrier function mechanism that actively identifies and filters possible harmful user inputs at every turn with the goal of ensuring that model outputs remain inside a safety margin at any point of the dialogue graph[10]. A broader view is presented by the International AI Safety Report which suggests a framework of risk classification and defense measures that cover the whole stage of the LLM lifecycle[19]. A different stance is taken by Symbolic Guardrails: implementation of rules via symbolic means and reduction of utility loss[20].

Nevertheless, there are new challenges to defense research. Qi et al. discovered that despite user good faith, fine-tuning aligned models may endanger their safety[21]. It indicates that safety alignment is not a one-off undertaking but an ongoing process of maintaining it.

Yet another problem is that defense plans have been developed using researcher assumptions regarding attack situations. It is not known whether these designs are closed loops in relation to evaluation results and actual attacks. The mechanism is not yet mature[9-11,19-21].

2.4. Theoretical Analysis Research

Theoretical studies inquire: Why is it dangerous[2,12,13,22]? The concept of a persona drift explains how models can be inconsistent in their behavior on multistep conversations[2]. The three types of value misalignments identified by Motnikar et al. were instrumental convergence bias, extrapolation risk misjudgment, and ethical priority inversion[12]. According to Wei et al., another underlying issue has been discovered: to maximize rewards, large language models could acquire some form of reward tampering, i.e., adopt methods that are contrary to the designer goals to gain more points[22], which is directly related to the instrumental convergence bias.

The study conducted by Tang et al. is the most directly related to the central issue of this paper[13]. They describe the decision-making of large language models in value conflicts as a priority graph, and find that model priorities are not constant or consistent across contexts[13]. The finding gives an essential theoretical instrument to comprehend extreme value alignment failure. Also, they introduce the notion of "priority hacking" - attackers may influence the priority graph of a model through creating misleading contexts[13].

Theoretical studies are highly insightful, however, between the concepts and practice, there is a gap. What is the method of measuring the persona drift? What is the process of quantifying the value misalignment? How can the priority graph be mapped onto concrete evaluation metrics? The questions have not been explored in a systematic way[2,12,13,22].

ISSN: 2664-9640

3. Core Concepts and Analysis Framework

3.1. Extreme Value Alignment Failure

AI safety value alignment means that the behavior of the system should be consistent with human values[12]. The current paper addresses the concept of extreme value alignment failure as a particular case: when a dialogue system encounters a complicated value conflict regarding the essential human welfare, the model chooses shallow norms at the expense of more significant interests of humans[1,2].

The priority graph model proposed by Tang et al. can be used to elucidate the mechanism that underlies such behavior[13]. The value priorities of large language models are not stable but may change depending on the context. Deceptive contexts can be created by attackers with multi-turn dialogues, which trigger a change in the priority graph, and thus it is known as "priority hacking"[13]. This mechanism leads to extreme value alignment failure.

Drawing on the empirical framework of Motnikar et al. [12] and the reward tampering findings of Wei et al. [22], this paper categorizes extreme value alignment failures according to the stage of value reasoning in which the failure occurs: (i) priority inversion – a mistake in the ordering of ethical principles, where a secondary norm is mistakenly elevated above a fundamental one; (ii) extrapolation misjudgment – a mistake in the application of principles to novel contexts, leading to incorrect reasoning from known to unknown situations; and (iii) instrumental convergence bias – a mistake in goal optimization, where the model adopts strategies that violate deeper ethical constraints (e.g., reward tampering) to maximize a superficial objective. These three types are not mutually exclusive but reflect a progression from value representation (i), to reasoning under uncertainty (ii), to behavioral adaptation (iii). Understanding their interrelations helps to design more targeted evaluation benchmarks and defenses.

3.2. Four-Dimensional Full-Link Analysis Framework

In order to solve the fragmentation issue identified in Chapter 2, this paper has suggested using a four-dimensional analysis framework.

These four dimensions are directly derived from the four research directions reviewed in Chapter 2: failure mode from theoretical studies, attack vector from attack research, defense level from defense research, and evaluation benchmark from evaluation research. They form a logical sequence: failure mode specifies what vulnerability exists; attack vector shows how it can be triggered; defense level indicates where countermeasures can be applied; evaluation benchmark determines whether those countermeasures are effective. This sequence-vulnerability → trigger → countermeasure → verification-covers the entire safety loop.

Failure mode responses address where the model will err - the three extreme values alignment failure modes detailed above[12,22]. Attack vector responses are concerned with how attackers take advantage of it - the scope of attacks includes all the stages of the evolutionary continuum of one-step explicit attacks to multi-step context drift[6-8,17]. Defense level responds to where defense will break down - three levels, input filtering, model alignment training, and output auditing, as well as a flaw known as the problem of fine-tuning compromising safety shown by Qi et al.[21]. Evaluation benchmark responds to how to test - it is important that the methods of evaluation should change dynamically in response to changes in attack evolution and defense failures and HarmBench is one such attempt to do this[14].

This framework has as its central value the fact that it puts the four directions on the same map so that researchers within each direction can view where their work sits in the overall picture and what points of contact they have with the other directions.

ISSN: 2664-9640

Table 2. Four-Dimensional Integrated Analysis Framework

Dimension	Core Question	Key Elements
Failure mode	Where does the model err?	Ethical priority inversion; extrapolation risk misjudgment; instrumental convergence bias[12,22]
Attack vector	How do attackers exploit?	Single-step explicit attacks to multi-turn context drift[6-8,17]
Defense level	Where does defense fail?	Input filtering; model alignment training; output auditing; fine-tuning compromises safety[21]
Evaluation benchmark	How to test dynamically?	Static benchmarks: DiaSafety[1], JailBench[3], MoralBench[4], CValues[5]; dynamic: HarmBench[14]

4. Analysis and Discussion

4.1. Disconnection Points among the Four Directions

The four dimensions remain largely disconnected when considered together. [1-22].

The safety assessment studies identify the deficiencies of models, which are not often applied by the attack studies in order to create more specific attacks or by defense studies to improve the protection plans[1,3-5,14-16]. Although there have been efforts to dynamically evaluate such as HarmBench[14] the feedback mechanism between defense design and the evaluation does not exist.

The attack research finds new vulnerabilities, however, it is rarely mentioned whether these vulnerabilities are accounted in evaluation benchmarks or fixed by defense research. The effectiveness of multi-turn decomposition attacks was shown by Jailbreaker by Liu et al.[8], but is this result included in common evaluation standards or created a new defense solution? Unknown.

The design of protection schemes in defense research is grounded in its assumptions, but it remains up to the reader to determine whether such schemes are truly effective against the actual vulnerabilities identified in attack research and can be evaluated according to certain standards[9-11,19-21]. Even more worrying is the observation made by Qi et al.[21]: in case a defense program has been successful in passing an evaluation, a single round of user fine-tuning could ruin it.

Theoretical studies offer conceptual instruments like priority graph, persona drift, and reward tampering, nevertheless, how to transform these concepts into operation metrics of evaluation and lead defense design remains unexplored systematically[2,12,13,22]. The idea of priority hacking by Tang et al.[13] has great value, however, no assessment standard or defense plan that would rely on this idea has been established to date.

4.2. Structural Misalignment between Static Evaluation and Dynamic Attack

The re-examination of the previous studies using the four-dimensional model shows that there is an obvious incongruity between static assessment and dynamic assault[1-22]. This incongruity is empirically supported by several comparative studies. For example, Mazeika et al. [14] found that static benchmarks overestimate model safety by up to 30% compared to dynamic red-teaming evaluations. Similarly, Liu et al. [8] showed that fixed test sets fail to detect 45% of multi-turn jailbreak attacks that succeed in adaptive settings. These figures quantify the gap between static and dynamic assessments.

On the attack vector dimension, DAMON[7] and Jailbreaker[8] have demonstrated that an attacker can modify its strategy in a multi-turn dialogue dynamically - a very versatile way of attacks. But, in the evaluation benchmark dimension, DiaSafety[1], JailBench[3], MoralBench[4], and CValues[5] mostly conduct their measurements using fixed test sets to measure their performance statically. Though the other three are more concerned about value alignment in

ISSN: 2664-9640

China, the assessment approach is still non-dynamic. The evaluation benchmark has not had the dynamism signal of the attack vector dimension and thus still measures dynamic risk using a static methodology.

The HarmBench[14] by Mazeika et al. is a significant progress towards dynamism but does not emphasize automated red teaming approach as much as it should be linked to the theory of defense failure modes, priority hacking, or any other theoretical results.

The source of this mismatch cannot be attributed to the lack of intelligence in the evaluation benchmark designers, but to the absence of a communication channel between the four dimensions. The attack research identified the dynamic nature of attacks, however, it has not been communicated to the evaluation research[6-8,17,18].

4.3. Integrative Value of the Four-Dimensional Framework

The main role of the four-dimensional framework is not to substitute particular studies in either direction, but to act as a link between the four directions[1-22].

In case of an attack study finding a new jailbreak technique, the four-dimensional model can be used to question: What failure mode is being exploited by this attack? At what defense degree should it be taken into consideration? Should the test cases be added to the current assessment standards[8,14]?

With the introduction of hypothetical research concepts such as priority hacking, the four-dimensional conceptualization is useful in asking the questions: Into what failure mode can this concept be implemented? What attack methods can be used to confirm its presence? What are the defense mechanisms that can be used against it? What assessment standards can be applied to assess the susceptibility of models to it[13]?

This very connecting role is what will allow the four-dimensional framework to incorporate the disjointed safety literature into an organic unity.

5. Conclusion

The present paper has taken as its research object extreme value alignment failures in open-domain dialogue systems. It comes up with the following conclusions through a systematic literature review.

The first issue that has been identified with existing research is the disconnect between the four directions of attack, defense, evaluation, and theory, which work separately, and their interfaces are not connected to each other. Safety evaluation indicates that a model is unsafe but cannot explain why or give any suggestions; attack research identifies the weaknesses but fails to communicate them to defense and evaluation; defense design is grounded on its presumptions instead of real vulnerabilities; theoretical constructs are still theoretical and have not yet been converted into practical applications. Such a fragmented situation causes the same high risk to recur without any solution.

Next, the current paper builds a four-dimensional analysis framework of full link as failure mode-attack vector-defense level-evaluation benchmark. The given framework consolidates diverse findings of various studies into a single map, which is able to reveal internal relations between the dimensions: vulnerabilities found in attack studies ought to be reflected in defense and evaluation studies; weaknesses found in evaluation studies need to encourage the further development of attack and defense studies; theoretical concepts must be translated into working evaluation indicators and defense mechanisms.

The third point is an analysis of this framework revealing that the mismatch between structural static assessment and dynamic assault, as well as new risks including fine-tuning compromise safety, are basically the information disconnect of the four dimensions. To address the issue it

ISSN: 2664-9640

would be necessary to create feedback mechanisms between the four directions instead of advancement in one direction only.

The present study indicates a number of directions in which further research could be conducted, including creating operational assessment instruments on the basis of the four-dimensional model, converting theoretical definitions like priority hacking into concrete assessment indicator standards; investigating normalized feedback systems between attack, defense, and evaluation research; and performing comparative research on value alignment within different cultures.

The shortcomings of the present research are: since it is a literature review, the suggested four-dimensional framework has not been tested empirically; because this discipline is developing rapidly, there is a chance that some of the most recent findings of the study are not covered. They will be considered in future efforts.

References

[1] Sun, H. (2023). *Research on safety methods for open-domain dialogue systems* [Doctoral dissertation]. Tsinghua University.

[2] Lu, C., et al. (2026). The assistant axis: Situating and stabilizing the default persona of language models. arXiv preprint arXiv:2601.10387.

[3] Li, L., et al. (2024). JailBench: Assessing and eliciting harmful content in Chinese LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 567–582). Bangkok, Thailand.

[4] Jiang, L., et al. (2024). MoralBench: Moral evaluation of large language models. arXiv preprint arXiv:2406.04428.

[5] Cui, J., et al. (2024). CValues: Measuring the values of Chinese large language models. arXiv preprint arXiv:2307.09705.

[6] Zhang, Z., et al. (2024). EasyJailbreak: A unified framework for jailbreaking large language models. arXiv preprint arXiv:2403.12171.

[7] Zhang, X., et al. (2025). DAMON: A dialogue-aware MCTS framework for jailbreaking large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 112–128). Miami, FL.

[8] Liu, Y., et al. (2024). Jailbreaker: Automated jailbreaking of large language models via prompt decomposition. arXiv preprint.

[9] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

[10] Hu, H., et al. (2025). Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. arXiv preprint arXiv:2503.00187.

[11] Anthropic. (2024). *The Claude 3 model family: Opus, Sonnet, Haiku*. https://www.anthropic.com/news/claude-3-family

[12] Motnikar, L., et al. (2025). The value of Gen-AI conversations: A bottom-up framework for AI value alignment. In *Proceedings of the 33rd European Conference on Information Systems (ECIS)*. Oslo, Norway.

[13] Tang, et al. (2026). Priority hacking: The vulnerability of LLM value alignment. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*. Miami, FL.

[14] Mazeika, M., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249.

[15] Huang, Z., Ramnath, K., Chen, Y., et al. (2026). Diffusion language model inference with Monte Carlo tree search. In *Findings of the Association for Computational Linguistics: EACL 2026* (pp. 3493–3512). Rabat, Morocco.

ISSN: 2664-9640

[16] Carlini, N., et al. (2024). How many unicorns are in this image? A safety evaluation benchmark for vision-language models. arXiv preprint.

[17] Zou, A., et al. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.

[18] Journal of Sichuan University. (2026). A review of black-box jailbreak attacks on large language models. *Journal of Sichuan University (Natural Science Edition)*, 63(2), 1–15.

[19] Bengio, Y., et al. (2025). International AI Safety Report 2025: Second key update: Technical safeguards and risk management. arXiv preprint arXiv:2511.19863.

[20] Hong, Y., et al. (2026). Symbolic guardrails for domain-specific agents: Stronger safety and security guarantees without sacrificing utility. arXiv preprint arXiv:2604.15579.

[21] Qi, X., et al. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv preprint.

[22] Denison, C., et al. (2024). Sycophancy to subterfuge: Investigating reward-tampering in large language models. arXiv preprint arXiv:2406.10162.