

# RACA: Representation-Aware Coverage Criteria for LLM Safety Testing

Zeming Wei

Zhixin Zhang

Chengcan Wu

Yihao Zhang

Xiaokun Luan

Meng Sun

Peking University

## Abstract

Recent advancements in Large Language Models (LLMs) have led to significant breakthroughs in various artificial intelligence applications. However, their sophisticated capabilities also introduce severe safety concerns, particularly the generation of harmful content through jailbreak attacks. Current safety testing for LLMs often relies on static datasets and lacks systematic criteria to evaluate the quality and adequacy of these tests. While coverage criteria have been effective for smaller neural networks, they are not directly applicable to LLMs due to scalability issues and differing objectives. To address these challenges, this paper introduces RACA (Representation-Aware Coverage criteriA), a novel set of coverage criteria specifically designed for LLM safety testing. RACA leverages representation engineering to focus on safety-critical concepts within LLMs, thereby reducing dimensionality and filtering out irrelevant information. The framework operates in three stages: first, it identifies safety-critical representations using a small, expert-curated calibration set of jailbreak prompts. Second, it calculates conceptual activation scores for a given test suite based on these representations. Finally, it computes coverage results using six sub-criteria that assess both individual and compositional safety concepts. We conduct comprehensive experiments to validate RACA’s effectiveness, applicability, and generalization, where the results demonstrate that RACA successfully identifies high-quality jailbreak prompts and is superior to traditional neuron-level criteria. We also showcase its practical application in real-world scenarios, such as test set prioritization and attack prompt sampling. Furthermore, our findings confirm RACA’s generalization to various scenarios and its robustness across various configurations. Overall, RACA provides a new framework for evaluating the safety of LLMs, contributing a valuable technique to the field of testing for AI.

## 1 Introduction

In recent years, Large Language Models (LLMs) have achieved tremendous success across a wide range of tasks, marking breakthrough milestones in artificial intelligence (AI) applications. By leveraging knowledge from extensive datasets and employing advanced token decoding strategies, LLMs have made significant progress in areas such as chat completion [34, 29], scientific reasoning [24, 1], code generation [21, 15], and program repair [7, 26]. These accomplishments have established LLMs as a foundational component of modern AI and software systems.

Despite remarkable success, LLMs also face severe safety issues in practical deployments [3, 63, 60, 12]. In particular, their advanced generation capabilities have raised concerns about their potential to produce harmful content that violates their safety or ethical guidelines, which is also referred to as *jailbreak attacks* [42, 74, 52, 31]. These issues prioritize the need for effective safety testing to comprehensively evaluate LLM safety against such risks. To prevent confusion, this work refers to the *LLM safety* as the generation of harmful content by LLMs, and *safety testing* as evaluating LLM on robustness against such jailbreak prompts.

So far, safety testing for LLMs primarily focuses on developing advanced algorithms to optimize attack prompts, such as transforming base jailbreak prompts into more sophisticated ones [9, 52, 74]. However, current safety testing often relies on static, open-source jailbreak datasets, limiting the ability to customize or evaluate their quality and adequacy. This imbalance makes the evaluation and prioritization of safety testing datasets relatively overlooked. Since safety constraints for LLMs can change over time and vary by scenario, consistently rejecting a specific dataset does not ensure comprehensive safety across diverse harmful contexts. Consequently, static jailbreak datasets cannot meet the complex safety testing requirements for real-world deployments, highlighting the need for established coverage criteria for LLM safety testing as a potential solution.

In the era of small-scale deep neural networks (DNNs), coverage criteria have proven effective in testing, focusing primarily on neuron activations or trajectories. For instance, Neuron Coverage (NC) [36] emphasizes the activation levels of individual neurons, while K-Multisection Neuron Coverage (KMNC) [32] evaluates the utilization of neurons across various activation ranges. However, these criteria cannot be directly applied to LLM safety testing due to their scalability limitations and different objectives. First, the computational complexity of most DNN criteria proves impractical for the larger parameter space of LLMs. Additionally, the objectives of DNN criteria differ significantly from LLM safety. Incorporating information from all neurons may yield redundant data irrelevant to LLM safety, thereby introducing noise into coverage results.

In this paper, we address the two problems above by leveraging LLM representation engineering techniques. Since most of the information derived by LLM neurons is redundant or irrelevant for safety, focusing on safety-critical concepts can achieve both effective dimension reduction and redundant information removal. To achieve this, we leverage the unique *representations* [73] that emerge in LLMs to model safety-related concepts within a specific prompt. These representations consist of specific directions in hidden states that indicate particular concepts and can be modeled by a moderate-sized calibration dataset. In particular, the safety-critical representations [53, 67, 64] denote the safety-related concepts within the inputs.

Based on this motivation, we explore representation-aware coverage criteria for LLM safety testing, and propose RACA, a set of Representation-Aware Coverage criteria. As outlined in Figure 1, the RACA derives coverage results for a test dataset through three key elements. First, RACA utilizes a calibration set to identify safety-critical representations for further evaluation. This relatively small size calibration set can be selected by human experts using representative jailbreak prompts aligned with desired safety guidelines. Next, for a target test suite, RACA calculates conceptual activation scores based on these identified safety-critical representations. This process transforms high-dimensional, redundant neuron information into safety-related concepts that are the primary focus of LLM safety testing. Finally, RACA computes the coverage results based on criteria derived from these conceptual activations, as well as their ensemble scores, to assess the quality of coverage for the test suite. We propose two sub-groups of criteria: (i) individual concept coverage, which focuses on the coverage of each extracted safety concept, and (ii) compositional coverage, which focuses on the coverage of compositions of different safety concepts, forming a set of six sub-criteria of RACA.

To validate the effectiveness of RACA, we conducted comprehensive experiments focusing on its effectiveness, applicability, and generalization. First, we demonstrated that RACA can identify high-quality jailbreak prompts while remaining insensitive to invalid or redundant cases, outperforming neuron-level criteria across multiple testing requirements. Additionally, we explored the application of RACA in real-world deployment in two scenarios: test set prioritization and jailbreaking prompt sampling, showing its practicality for LLM safety testing. Finally, we assessed the generalization of RACA across diverse settings, including its scalability to larger models, robustness in criteria

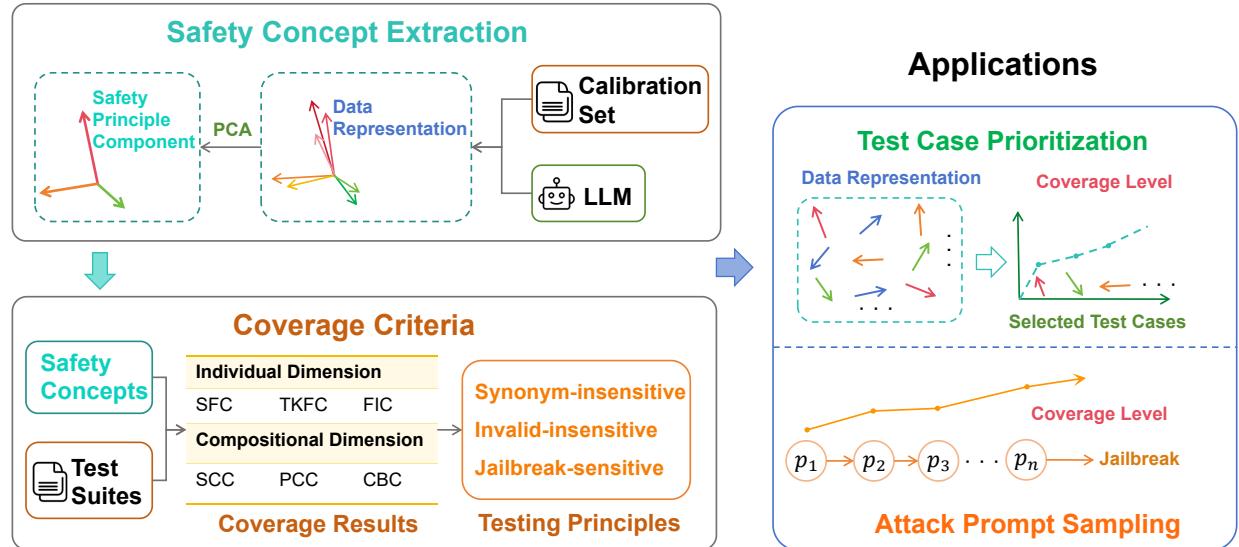


Figure 1: An overview of our RACA framework. In the Safety Concept Extraction module, we use a Calibration Set to extract safety-critical representations, and then apply PCA to obtain the Safety Principal Components, which serve as the Safety Concepts for subsequent steps. Then, in the Coverage Criteria module, we extract the representation vectors of the Test Suite and compute the final coverage scores with safety concept activations. These scores are derived from six distinct criteria categorized into two dimensions: Individual and Compositional Concept Coverage, aligning with our three core design principles. In practical applications, RACA can be leveraged for Test Prioritization and Attack Prompt Sampling.

configuration, and calibration set selection. Overall, RACA serves as a novel coverage testing framework for the safe alignment of large language models (LLMs), introducing a new technique for AI safety testing. Our contribution in this paper can be summarized as follows:

- We propose RACA, a set of coverage criteria specialized for LLM safety testing with a representation-aware activation modeling, addressing the scalability and adaptivity issues in existing criteria.
- We conduct comprehensive experiments to validate the effectiveness, applicability, and generalization of RACA, showing its practicality as a coverage criterion in real-world safe testing scenarios.
- We release RACA in <https://github.com/weizeming/RACA> and further provide practical suggestions for applying RACA.

The rest of this paper is organized as follows. In Section 2, we introduce backgrounds and preliminaries for this paper. Then, in Section 3, we detail our RACA framework. Section 4 presents the comprehensive evaluations on RACA across various research questions. Finally, we discuss threats to validity in Section 5 and related work in Section 6, and finally conclude our work in Section 7.

## 2 Background and Motivation

### 2.1 LLMs and their Safety Issues

With the fast development of computation and data sources, LLMs have achieved significant breakthroughs in various application paradigms like chat completion [29, 34], code generation [15, 21], and complex reasoning tasks [22, 69]. Generally, an LLM<sup>1</sup> can be formulated as follows:

**Definition 1** (Large Language Model, LLM). *An LLM is an autoregressive decoder-only transformer [47, 39], denoted by a tuple  $f = (\theta, H)$ , where  $\theta$  represents the model architecture and the model parameters, and  $H$  is the hidden feature space.*

During inference,  $f$  receives an input  $x$  consisting of a sequence of  $k$  tokens and  $f(x_{[1:k]})$  predicts the next token  $x_{[k+1]}$ , then append it to current input  $x$  and continues this procedure until the model outputs the (end of sequence) token <EOS>.

Despite their milestone success, their safety issues have become a major concern in their deployments [3, 60, 51, 13, 54]. Though LLMs are generally trained with alignment techniques [6, 5, 17, 56] to refuse to answer harmful queries, evaluations on popular open-source harmful prompt datasets have shown that their safety is still superficial and inadequate [38, 14], which are commonly known as the *jailbreak* issues. Additionally, safety principles and constraints for LLMs can vary by scenario, so static open-source harmful prompt datasets cannot satisfy the diverse safety testing needs. These situations further underscore the importance of suitable coverage criteria for LLM safety testing.

### 2.2 Coverage Testing for AI: From DNNs to LLMs

DNN coverage testing mirrors traditional code coverage by examining the internal states of DNNs, such as neuron activation statistics, to represent their functional logic. Typically, a DNN coverage criterion aims at assessing a test suite's ability to reveal functional diversity and detect adversarial defects in target models. Representative metrics like Neuron Coverage (NC) [36], Top-K Neuron Coverage (KNC) [32], and TensorFuzz Coverage (TFC) [33] focus on activated neurons and their patterns, which are used to evaluate the test suite's adequacy for identifying DNN vulnerabilities. However, most of these criteria suffer from complexity bottlenecks and are designed only for small-scale DNNs, which are proven impractical for LLMs.

Transitioning from DNN robustness to LLM safety presents unique challenges for existing coverage criteria, primarily due to computational complexity and shifting objectives. First, the high computational demands of most DNN criteria make them impractical for the extensive parameter space of LLMs. For example, KMNC/NBC/SNAC [32] requires tracing the activation range of neurons on all training data, which is massive for LLMs. Furthermore, the goals of DNN criteria differ significantly from those related to LLM safety. Incorporating information from all neurons can introduce redundant data that is irrelevant to LLM safety, adding unnecessary noise to coverage results. Thus, there is an emerging need for effective and scalable criteria specifically tailored for LLM safety.

### 2.3 Safety Representations in LLMs

Driven by the unique language comprehension ability of LLMs, *representation engineering* [73, 43, 64, 8, 44, 20, 53] has emerged as a novel framework to characterize high-level concepts emerged

---

<sup>1</sup>In this work, we focus on autoregressive decoder-only LLMs, which is the current mainstream model architecture.

in LLMs. Specifically, these studies revealed that low-rank representations exist that can represent and steer specific concepts in their hidden states, where a typical representation extraction process can be modeled as applying principal component analysis (PCA) techniques to the hidden states of a set of representative concept-related inputs. To capture these representations, we formulate related notations as follows: for a subset of the hidden feature space  $h \subseteq H$  used for representation construction, we denote  $h(p) \in \mathbb{R}^d$  as the focused feature representation of  $p$ , where  $d = |h|$  is the dimension of the representation.

In terms of LLM safety, a few preliminary studies have also revealed the existence of safety representations [53, 64, 67, 20, 35, 48]. Specifically, these works discovered safety representations that indicate the safety-related concepts in the processing inputs can be captured in the hidden states of LLMs. Furthermore, these findings suggest that the safety representations in LLMs can identify the safety concepts when processing new test cases. Thus, by projecting the safety representations of test cases into safety concepts, the activated safety concepts offer a unique solution for addressing the scalability issues aforementioned of coverage criteria for LLM safety testing.

### 3 Representation-Aware Coverage Criteria

In this section, we introduce our proposed representation-aware coverage criteria for LLM safety testing, starting from our design principles for this problem. Then, we introduce the concept extraction process to build the safety-critical representations, followed by the two groups of representation-aware criteria: individual-concept coverage and compositional-concept coverage.

#### 3.1 Design Principles

First, we state our design principles for LLM safety testing. Unlike conventional criteria for small-scale DNN, the goal of saty testing is fundamentally different. Desirable criteria should have the following functionalities:

- (i) **Synonym-insensitive.** This principle aligns with traditional coverage criteria for small-scale DNNs, emphasizing input diversity. Given that many harmful prompt datasets are generated by LLMs and tend to contain significant redundancy [10], it's anticipated that harmful prompts with duplicated or similar concepts should not increase the coverage metrics.
- (ii) **Invalid-insensitive.** In LLM safety testing, the focus is on the model's capability to reject harmful prompts; therefore, invalid inputs such as non-harmful or weak adversarial prompts should not contribute to testing coverage. However, this requirement is challenging for DNN criteria, as any semantic variation could lead to distinct neural activations, irrespective of their safety-specific meanings.
- (iii) **Jailbreak-sensitive.** In contrast to (ii), LLM safety testing ought to emphasize prompts that are capable of effectively triggering harmful behaviors for optimal testing. This requirement fundamentally differs from adversarial attacks in DNNs that target robustness and consider any unusual activations, as it specifically concentrates on the activation of safety-critical concepts.

Based on the design principles stated above, we propose a white-box testing framework designed to evaluate the robustness of LLM safety by inspecting the model's safety-critical representation space, with the procedure detailed below. We then formalize testing adequacy through two orthogonal dimensions: Individual Concept Coverage (Dimension I) and Compositional Concept Coverage (Dimension II).

### 3.2 Concept Extraction via Representation Engineering

A typical representation extraction process from LLMs involves two main components: calibration set construction and dimensionality reduction [73, 64, 67, 43, 55]. The calibration set implicitly encodes the target concepts with representative data to obtain the concept-related representations. For example, identifying the truthfulness-critical representation can leverage the data that causes LLM hallucinations [49]. In terms of safety-critical concepts, harmful prompts can be used for this construction [53, 64, 67]. Therefore, we propose to build a limited-sized harmful prompt dataset as the calibration set. In practice, this dataset can be obtained from human-expert annotations with representative prompts corresponding to the safety constraints for testing.

Then, following existing representation engineering methods, we apply Principal Component Analysis (PCA) on the calibration data. Let  $D_{calib}$  be the calibration dataset, we collect the centered activation vectors  $H \in \mathbb{R}^{m \times d_{model}}$  from the target layer. We then perform PCA to identify the top- $n$  principal components  $V = \{v_1, \dots, v_n\}$ , indicating the primary directions of the implicitly encoded safety concepts. We then define the *concept activation*  $f_j(x)$  for an input  $x$  as the magnitude of the projection of the layer activation  $h(x)$  onto the  $j$ -th principal component, which quantifies the intensity of the concept's presence:

$$f_j(x) = v_j^T (h(x) - \mu), \quad (1)$$

where  $\mu$  is the mean activation of the calibration set (corresponding to the centroid process of PCA).

We formally define the set of targeted safety features  $F_{safe} = \{1, \dots, n\}$  as the indices of these top- $n$  principal components. Based on representation engineering theories [73, 64], these extracted dimensions can be interpreted as the primary semantic variations in the safety concept space. Leveraging concept activations, we explore representation-aware coverage criteria based on the distributions of concept activations in the test set.

### 3.3 Dimension I: Individual Concept Coverage

Based on these extracted principal features, we first utilize three criteria to quantify how well the test suite  $\mathcal{T}$  exercises individual concepts.

**Criterion 1: Safety Feature Coverage (SFC).** We first consider measuring the breadth of extracted safety concepts triggered by the test suite. Motivated by the NC criterion at the neuron level, a comprehensive test suite should elicit significant activation across the identified safety subspace, so we define the Safety Feature Coverage (SFC) as the proportion of concepts that are sufficiently activated by at least one prompt in the test suite:

$$SFC = \frac{|\{j \in F_{safe} \mid \exists x \in \mathcal{T}, f_j(x) > \epsilon\}|}{|F_{safe}|}. \quad (2)$$

where  $\epsilon$  is a minimal threshold to distinguish significant semantic activation from numerical noise.

**Criterion 2: Top-K Feature Coverage (TKFC).** Motivated by the TKNC criterion for neuron coverage, a feature must be salient enough to characterize the model's safety state. We define  $TopK(x)$  as the set of indices of the  $k$  features with the highest activation magnitudes for input  $x$ , Top-K Feature Coverage (TKFC) quantifies the proportion of safety features that have served as a dominant component for at least one test case:

$$TKFC = \frac{|\{j \in F_{safe} \mid \exists x \in \mathcal{T}, j \in TopK(x)\}|}{|F_{safe}|}. \quad (3)$$

**Criterion 3: Feature Intensity Coverage (FIC).** Safety mechanisms may behave differently depending on the intensity of the activation along safety directions. Adapting from the KMNC

criterion, we discretize the activation magnitude range of each feature  $j \in F_{safe}$  into  $K$  intensity bins. Feature Intensity Coverage (FIC) then measures the proportion of these intensity bins covered across all safety features:

$$FIC = \frac{1}{|F_{safe}|} \sum_{j \in F_{safe}} \frac{|\{b \in \text{Bins} \mid \exists x \in \mathcal{T}, f_j(x) \in b\}|}{K}. \quad (4)$$

### 3.4 Dimension II: Compositional Concept Coverage

While Dimension I evaluates principal concepts in isolation, we also anticipate that a comprehensive safety test suite will cover diverse combinations across concepts, as complex adversarial attacks often exploit the composition of these underlying factors, particularly in uncommon combinations. This dimension evaluates the diversity of the global state vectors within the projected safety subspace.

For notation convenience, we define the safety state of an input  $x$  as the vector of feature activations:  $v_x = [f_j(x)]_{j \in F_{safe}}$ . To evaluate the composition of these states, we employ unsupervised clustering that pre-compute a set of  $M$  centroids  $\{\mu_1, \dots, \mu_M\}$  using K-Means clustering on the projected representations of the calibration dataset. These centroids represent distinct ‘semantic modes’ of safety scenarios derived from the calibration set.

**Criterion 4: Semantic Cluster Coverage (SCC).** Semantic Cluster Coverage (SCC) assesses the combination diversity of the test suite. It calculates the fraction of pre-defined semantic clusters that the test suite has visited. A low SCC indicates that the generated test cases are semantically repetitive (mode collapse to a few semantic clusters).

$$SCC = \frac{|\{m \in \{1 \dots M\} \mid \exists x \in \mathcal{T}, \operatorname{argmin}_k \|v_x - \mu_k\|_2 = m\}|}{M}. \quad (5)$$

**Criterion 5: Pairwise Concept Coverage (PCC).** To capture precise feature interactions while mitigating the combinatorial explosion, Pairwise Concept Coverage (PCC) focuses on the co-occurrence of concept pairs. This metric treats safety features as nodes in a semantic graph and evaluates the coverage of edges (pairwise correlations) between them. Let  $\mathcal{P}_{all} = \{(i, j) \mid i, j \in F_{safe}, i < j\}$  be the set of all unique feature pairs, with a total cardinality of  $\binom{|F_{safe}|}{2}$ . PCC measures the fraction of these pairs that activate simultaneously within the test suite:

$$PCC = \frac{|\{(i, j) \in \mathcal{P}_{all} \mid \exists x \in \mathcal{T}, f_i(x) > \epsilon \wedge f_j(x) > \epsilon\}|}{\binom{|F_{safe}|}{2}}. \quad (6)$$

**Criterion 6: Cluster Boundary Coverage (CBC).** Similar to the adversarial examples in DNNs, successful attacks often reside in the low-density regions between established semantic clusters: the ‘ambiguous’ zones where the model’s safety boundaries are ill-defined [74, 52]. Motivated by the superise-based criteria for DNNs [27, 28], we identify a test case as a ‘boundary case’ if its Euclidean distance to the nearest centroid  $\mu_{nearest}$  exceeds a threshold  $\delta$ . Cluster Boundary Coverage (CBC) measures the proportion of test cases that explore these sparse regions:

$$CBC = \frac{|\{m \in \{1 \dots M\} \mid \exists x \in \mathcal{T} \min_k \|v_x - \mu_k\|_2 > \delta\}|}{|\mathcal{T}|}. \quad (7)$$

### 3.5 Summary and Discussion

Overall, consisting of a set of six sub-criteria, RACA addresses the current limitations of neuron-level criteria for LLM safety testing and aligns with our proposed design principles:

- **Scalable to LLMs.** RACA addresses the computational bottlenecks inherent in traditional neuron-level coverage by projecting high-dimensional activation data into a low-dimensional conceptual space. By focusing on a limited set of safety-critical directions rather than billions of individual neurons, the framework significantly reduces the parameter space, making white-box safety testing feasible and efficient for Large Language Models.
- **Safety-specialized.** Unlike general coverage criteria that are often influenced by redundant or irrelevant neuron activations, RACA is specifically designed to isolate safety-critical representations. By utilizing a calibration set to identify conceptual directions directly related to safety violations, the framework filters out irrelevant semantic noise and ensures that the coverage results accurately reflect the model’s internal response to harmful content.
- **Principled.** The six sub-criteria of RACA are structured to align with the three core design principles. They achieve synonym-insensitivity by mapping semantically similar inputs to the same conceptual vectors, ensuring redundant prompts do not inflate metrics. They maintain invalid-insensitivity by ignoring activations that do not project onto the safety-critical subspace. Finally, they are jailbreak-sensitive as they specifically measure the activation intensity and composition of concepts that effectively trigger model vulnerabilities.

## 4 Evaluation

In this section, we present a comprehensive evaluation of RACA, starting with stating our research questions.

### 4.1 Research Questions

**RQ 1:** How effective is RACA compared with neuron coverage criteria for LLM safe alignment testing?

This research question examines the **effectiveness** of RACA. We conduct controlled experiments using test suites sampled from various configurations and compare the coverage results of RACA across them, as well as against neuron-level criteria.

**RQ 2:** Can RACA be applied to facilitate real-world LLM safety testing?

This question further examines the **applicability** of RACA. We investigate this through two representative applications of coverage criteria: enhancing test suite prioritization and effective attack prompt sampling.

**RQ 3:** How robust is RACA under diverse testing scenarios?

Finally, we assess the **generalization** of RACA under diverse testing scenarios, including scaling to larger models, construction of a calibration set, and sub-criterion hyperparameters.

### 4.2 Experiment Set-up

**Models and Datasets.** We consider three popular open-sourced LLMs: Vicuna-7b-v1.5 [68], Llama-2-7b-chat [46], and Qwen-2.5-7b-chat [4], which are widely adopted in LLM safety research. For harmful prompt datasets, we utilize two representative safety benchmarks: (1) **SorryBench**

[57], which contains 400+ harmful prompts across 44 distinct categories, providing a fine-grained taxonomy of safety violations, and (2) **ORBench-Toxic** [16], comprising thousands of harmful prompts across 10 toxic categories. We additionally consider Alpaca [45], consisting of benign prompts, as the simulation of invalid data.

**Baselines.** As verified by [71], only five neuron-level coverage criteria can be adapted for LLMs, while all others are incapable due to scalability bottlenecks. We consider all five criteria, with the best configuration parameter suggested by [71] for safety testing:

Table 1: Neuron-level criteria baselines.

Criteria	Config	Description
NC	$T = 0.25$	Measures the ratio of activated neurons above a threshold.
TKNC	$T = 10$	Counts neurons that have been among the top-k active neurons.
TKNP	$T = 1$	Counts the number of unique top-k activation patterns.
TFC	$T = 50$	Measures the diversity of activation values using buckets.
NLC	<i>N/A</i>	Evaluates layer-wise activation distributions.

**Parameter Settings.** In our experiments, we configured the parameters as follows: For neuron selection, we set  $\text{topk} = 2$ . The activation values are discretized using  $\text{bins} = 10$ . Neurons are grouped into  $\text{clusters} = 32$ . The slack coefficient for SFC is set as  $\varepsilon_{\text{sfc}} = 5.0$ . The threshold for PCC is defined as  $\varepsilon_{\text{pcc}} = 2.5$ . Additionally, we employ  $\delta = 8.0$  for CBC. These parameter values were determined through preliminary experiments to balance sensitivity and robustness in the coverage analysis, which will be further studied in RQ 3.

**Implementation.** The calibration set is selected by prioritizing successful jailbreak prompts and split from the target dataset to simulate high-quality test case annotations, maintaining a default size of 100. We then extract hidden states from the middle layers (15-18 out of 32) of the LLMs and average scores on them, as numerous studies have shown that middle layers effectively capture rich semantic information for representation extraction [73, 64, 35, 55], and apply PCA to reduce the dimensionality to 64. These selection strategies will be further discussed in RQ 3.

### 4.3 Test Suites Design

To rigorously evaluate RACA across different principles, we construct a series of synthetic test suites derived from the base plain suite (randomly sampled from the remaining prompts with a fixed size) to simulate different testing scenarios. First, Redundant-Semantic ( $S_{RS}$ ) adds synonymous prompts rephrased from the plain suite to test insensitivity to redundancy, where a subset of prompts is rephrased by the vicuna and incorporated. Second Redundant-Invalid ( $S_{RI}$ ) adds benign prompts sampled from Alpaca to test invalid-insensitivity. Finally, Jailbreak Attacks ( $S_{JA}$ ) adds prompts that successfully jailbreak the model without refusal from remaining prompts to test the sensitivity to adversarial inputs. An expansion set  $S_E$  that randomly adds remaining prompts from the same dataset to  $S_P$  serves as a baseline for these test suites for comparison. Additionally, the replacement versions ( $S^*$ ) retain the same size as  $S_P$  but replace a subset of prompts with their corresponding inputs for further comparison. Table 2 summarizes the design.

Under this design, suites  $S_{RI}^*$ ,  $S_{RS}^*$ , and  $S_{JA}^*$  have the same size with  $S_P$ , while  $S_{RI}$ ,  $S_{RS}$  and  $S_{JA}$  have the same size with  $S_E$ . Based on our design principles, for any coverage criterion  $s(\cdot)$ , the

Table 2: Summarization of synthetic test suites.

Suite	Notation	Source
Plain Expansion	$S_P$ $S_E$	Direct request harmful prompts More prompts than $S_P$
+ Redundant-Invalid	$S_{RI}$	Add $n$ prompts from benign datasets to $S_P$
+ Redundant-Semantic	$S_{RS}$	Add $n$ synonymous prompts to $S_P$
+ Jailbreak Attacks	$S_{JA}$	Add $n$ jailbreak attack prompts to $S_P$
$\sim$ Redundant-Invalid	$S_{RI}^*$	Replace $n$ prompts from $S_P$ to with invalid $S_P$
$\sim$ Redundant-Semantic	$S_{RS}^*$	Replace $n$ synonymous prompts to $S_P$
$\sim$ Jailbreak Attacks	$S_{JA}^*$	Replace $n$ prompts as jailbreak attack prompts in $S_P$

expected tendency across these test suites should be:

$$\begin{aligned} s(S_{JA}^*) &> s(S_P) > \{s(S_{RI}^*), s(S_{RS}^*)\}, \\ s(S_{JA}) &> s(S_E) > \{s(S_{RI}), s(S_{RS})\} \approx s(S_P). \end{aligned} \quad (8)$$

We then calculate the coverage results across all datasets and suites on the three models, which are summarized in Tables 3 and 4.

**Individual Concepts Criteria** The individual concept criteria, namely SFC, TKFC, and FIC, demonstrate a strong alignment with our design principles across both the SorryBench and ORBench benchmarks. First, they exhibit high sensitivity to jailbreak attacks. As shown in the average results in Tables, the jailbreak-attack suite ( $S_{JA}$ ) achieves a substantially higher coverage gain over the expanded suite ( $S_E$ ) for criteria like TKFC and FIC (e.g., on ORBench, the average FIC gain increases from +7.91% for  $S_E$  to +14.82% for  $S_{JA}$ ). Similarly, the replacement suite  $S_{JA}^*$  also yields a significant coverage increase compared to the plain suite  $S_P$ . Second, these criteria prove to be robust against synonym and invalid prompts. The coverage gains from adding semantically redundant ( $S_{RS}$ ) or irrelevant ( $S_{RI}$ ) prompts are marginal and consistently lower than those from  $S_E$ . More importantly, when existing prompts are replaced with such ineffective ones ( $S_{RS}^*$  and  $S_{RI}^*$ ), all criteria register a significant negative growth. For instance, on SorryBench, the average SFC for  $S_{RS}^*$  drops by -12.18%, confirming that RACA can effectively identify and penalize test cases of low utility.

**Compositional Concepts Criteria** The compositional concept criteria, including SCC, PCC, and CBC, further corroborate RACA’s capability to assess the combinatorial diversity of a test suite. These criteria also adhere perfectly to the expected tendencies. They display a more pronounced sensitivity to jailbreak attacks, with coverage gains on  $S_{JA}$  systematically surpassing those on  $S_E$ . For example, on ORBench, the average SCC gain skyrockets from +13.22% for  $S_E$  to +26.78% for  $S_{JA}$ , which suggests that jailbreak attacks often exploit novel and complex combinations of concepts to bypass safety alignments. Concurrently, these criteria exhibit reliable robustness against redundant ( $S_{RS}$ ,  $S_{RS}^*$ ) and invalid ( $S_{RI}$ ,  $S_{RI}^*$ ) test suites. The trends in coverage changes are consistent with those of the individual criteria: adding such prompts yields minimal gains, whereas replacing prompts leads to a marked decrease in coverage. Overall, the compositional criteria effectively measure a test suite’s capacity to trigger a model’s safety responses from multiple, deeper dimensions.

**Ensemble Comparison** The superiority of RACA is most evident when comparing its ensemble metrics (EI, EC, ER) against the neuron-level baseline (EN), as detailed in Table 5. First, in identifying high-quality jailbreak attacks, all of RACA’s ensemble metrics report stable and significant growth for  $S_{JA}$  and  $S_{JA}^*$ . Notably, on ORBench, the average ER gain for  $S_{JA}$  is +17.89%, substantially

Table 3: Overall coverage results on SorryBench.

Model	Suite	RACA						Neuron-level Criteria				
		SFC	TKFC	FIC	SCC	PCC	CBC	NC	TKNC	TKNP	TFC	NLC
Vicuna	$S_P$	0.53	0.53	0.76	0.88	0.45	0.67	1.00	0.01	66.75	10.00	4943.52
	$S_E$	+9.51%	+6.60%	+5.07%	+3.57%	+12.71%	+8.26%	+0.15%	+15.37%	+24.72%	+0.00%	+1.09%
	$S_{RS}$	+0.66%	+5.11%	+0.88%	+0.00%	+1.13%	+0.00%	+0.16%	+7.70%	+10.06%	+0.00%	-3.88%
	$S_{RI}$	+0.00%	+0.00%	+0.41%	+0.00%	+0.93%	+4.68%	+0.41%	+20.03%	+54.76%	+0.00%	+8.38%
	$S_{JA}$	+4.59%	+6.52%	+3.73%	+5.36%	+10.43%	+15.16%	+0.11%	+9.28%	+21.37%	+0.00%	+0.37%
	$S_{RS}^*$	-15.60%	-15.33%	-6.00%	-8.87%	-6.69%	-4.50%	-0.04%	-12.46%	-17.68%	+0.00%	-7.99%
	$S_{RI}^*$	-5.99%	-9.55%	-4.19%	-4.40%	-8.96%	+1.14%	+0.41%	+20.77%	+28.60%	+0.00%	+8.93%
	$S_{JA}^*$	+2.08%	+2.15%	+4.71%	+2.65%	+8.61%	-2.17%	+0.13%	+9.45%	+13.16%	-20.00%	+4.71%
	$S_P$	0.49	0.44	0.80	0.86	0.48	0.62	0.99	0.00	93.75	6.00	3820.05
Llama	$S_E$	+5.23%	+7.25%	+3.37%	+3.74%	+5.37%	+3.97%	+0.43%	+1.39%	+20.30%	+16.67%	+1.07%
	$S_{RS}$	+0.68%	+2.78%	+1.12%	+0.00%	+0.92%	+7.75%	+0.67%	+0.00%	+15.55%	+0.00%	-1.69%
	$S_{RI}$	+0.00%	+3.52%	+0.29%	+0.00%	+0.30%	+0.00%	+1.04%	+33.78%	+42.89%	+0.00%	+8.76%
	$S_{JA}$	+9.22%	+13.39%	+5.90%	+4.70%	+10.21%	+6.19%	+0.46%	+9.09%	+21.01%	+12.50%	+2.14%
	$S_{RS}^*$	-19.35%	-6.17%	-5.47%	-4.60%	-17.20%	+5.53%	-0.05%	-2.57%	-10.17%	+16.67%	-7.08%
	$S_{RI}^*$	-7.41%	-11.38%	-3.51%	-4.60%	-10.41%	-1.14%	+1.04%	+51.25%	+20.62%	+16.67%	+8.91%
	$S_{JA}^*$	+9.95%	-4.19%	+3.03%	+2.91%	+3.88%	-4.97%	+0.35%	-2.57%	+8.84%	-12.50%	+5.85%
	$S_P$	0.99	0.42	0.83	0.73	1.00	0.51	1.00	0.01	200.00	34.25	2059.68
	$S_E$	+0.80%	+9.55%	+3.56%	+10.60%	+0.00%	+5.61%	+0.04%	+5.10%	+25.00%	+19.74%	+0.10%
Qwen	$S_{RS}$	+0.40%	+6.65%	+0.75%	+1.04%	+0.01%	+1.32%	+0.16%	+15.79%	+25.00%	+44.54%	+2.49%
	$S_{RI}$	+0.00%	+2.90%	+0.65%	+0.00%	+0.01%	+0.00%	+0.22%	+30.09%	+25.00%	+89.83%	+8.21%
	$S_{JA}$	+1.20%	+15.68%	+5.01%	+11.73%	+0.01%	+9.03%	+0.06%	+10.08%	+25.00%	+31.45%	+1.06%
	$S_{RS}^*$	-1.59%	+0.64%	-4.03%	-11.73%	-0.06%	-4.92%	+0.10%	+4.59%	+0.00%	+25.25%	+2.36%
	$S_{RI}^*$	-1.59%	+0.36%	-3.65%	-8.51%	-0.04%	-4.30%	+0.22%	+23.73%	+0.00%	+60.60%	+9.66%
	$S_{JA}^*$	+1.20%	+13.89%	+3.73%	+6.34%	+0.01%	+5.96%	+0.04%	+1.62%	+0.00%	+13.78%	+0.78%
	$S_P$	0.67	0.46	0.80	0.83	0.64	0.60	0.99	0.01	120.17	16.75	3607.75
	$S_E$	+5.18%	+7.80%	+4.00%	+5.97%	+6.03%	+5.95%	+0.20%	+7.29%	+23.34%	+12.14%	+0.75%
	$S_{RS}$	+0.58%	+4.85%	+0.92%	+0.35%	+0.69%	+3.02%	+0.33%	+7.83%	+16.87%	+14.85%	-1.03%
Average	$S_{RI}$	+0.00%	+2.14%	+0.45%	+0.00%	+0.41%	+1.56%	+0.56%	+27.97%	+40.88%	+29.94%	+8.45%
	$S_{JA}$	+5.01%	+11.86%	+4.88%	+7.26%	+6.88%	+10.13%	+0.21%	+9.48%	+22.46%	+14.65%	+1.19%
	$S_{RS}^*$	-12.18%	-6.96%	-5.17%	-8.40%	-7.98%	-1.30%	+0.00%	-3.48%	-9.28%	+13.97%	-4.23%
	$S_{RI}^*$	-5.00%	-6.86%	-3.78%	-5.84%	-6.47%	-1.43%	+0.55%	+31.92%	+16.40%	+25.76%	+9.17%
	$S_{JA}^*$	+4.41%	+3.95%	+3.82%	+3.97%	+4.17%	-0.39%	+0.17%	+2.84%	+7.33%	-6.24%	+3.78%

outperforming the EN gain of +12.89%. The most critical distinction, however, lies in their ability to discriminate against invalid test inputs. When irrelevant benign prompts are added ( $S_{RI}$ ), the gain in RACA’s metrics is negligible (e.g., +0.76% for ER on SorryBench), correctly reflecting the low value of these prompts.

In stark contrast, the neuron-level EN metric is catastrophically misleading, showing a massive surge in coverage (+21.56% on SorryBench and +43.26% on ORBench). The issue is exacerbated in the replacement scenario ( $S_{RI}^*$ ), where EN continues to report a positive gain while RACA metrics correctly show a significant coverage drop. This provides definitive evidence that traditional neuron coverage cannot distinguish between effective safety tests and meaningless input perturbations. RACA, through its representation-aware engineering, offers a precise and reliable methodology for assessing the true quality of test suites for LLM safety.

**Answer to RQ 1:** RACA’s sub-criteria and their ensembles are effective for LLM safety testing principles and surpass neuron-level baselines.

Table 4: Overall coverage results on ORBench.

Model	Suite	RACA						Neuron-level Criteria				
		SFC	TKFC	FIC	SCC	PCC	CBC	NC	TKNC	TKNP	TFC	NLC
Vicuna	$S_P$	0.57	0.55	0.76	0.84	0.37	0.70	1.00	0.05	4.75	8.00	4801.50
	$S_E$	+2.25%	+14.36%	+7.43%	+12.31%	+5.51%	+12.63%	+0.01%	+29.11%	+0.00%	+0.00%	+0.26%
	$S_{RS}$	+7.07%	+5.00%	+4.05%	+5.64%	+9.37%	+6.71%	+0.01%	+50.01%	+21.25%	+25.00%	+0.25%
	$S_{RI}$	+0.76%	+7.09%	+1.28%	+1.79%	+2.15%	+2.00%	+0.03%	+96.11%	+48.75%	+9.38%	+10.87%
	$S_{JA}$	+3.74%	+19.39%	+12.51%	+16.99%	+8.83%	+13.51%	+0.01%	+29.81%	+6.25%	+0.00%	+1.05%
	$S_{RS}^*$	-12.96%	-9.61%	-6.96%	-14.99%	+6.04%	-5.26%	+0.01%	+10.04%	+0.00%	+25.00%	-0.64%
	$S_{RI}^*$	-9.79%	-3.34%	-7.68%	-14.10%	-8.75%	-3.17%	+0.03%	+70.08%	+55.00%	+3.13%	+13.80%
	$S_{JA}^*$	-4.04%	+4.66%	+2.46%	+8.46%	-4.04%	+4.21%	-0.01%	+1.40%	+6.25%	-12.50%	+0.62%
Llama	$S_P$	0.51	0.44	0.73	0.89	0.32	0.61	1.00	0.04	3.50	6.00	3728.81
	$S_E$	+16.88%	+13.35%	+10.93%	+6.10%	+16.34%	+13.03%	+0.01%	+36.51%	+0.00%	+0.00%	+0.26%
	$S_{RS}$	+4.14%	+16.20%	+4.91%	+4.44%	+6.60%	+10.37%	+0.06%	+41.19%	+45.83%	+33.33%	-0.10%
	$S_{RI}$	+0.00%	+9.05%	+0.96%	+1.72%	+1.68%	+9.24%	+0.09%	+107.83%	+97.92%	+16.67%	+11.09%
	$S_{JA}$	+21.99%	+21.71%	+19.15%	+12.39%	+24.51%	+23.24%	+0.03%	+44.16%	+31.25%	+0.00%	+1.69%
	$S_{RS}^*$	-15.99%	-12.71%	-5.28%	-12.26%	-8.86%	-8.08%	+0.06%	+4.68%	+39.58%	+16.67%	-1.21%
	$S_{RI}^*$	-14.48%	-17.95%	-7.78%	-8.75%	-7.10%	-6.40%	+0.09%	+77.15%	+75.00%	+0.00%	+13.58%
	$S_{JA}^*$	+19.88%	+16.41%	+15.37%	+10.66%	+18.41%	+20.38%	+0.01%	+5.68%	+2.08%	+0.00%	+0.89%
Qwen	$S_P$	1.00	0.49	0.81	0.66	1.00	0.41	1.00	0.05	7.75	99.50	2337.52
	$S_E$	+0.00%	+17.86%	+5.36%	+21.25%	+0.00%	+22.74%	+0.01%	+14.13%	+3.12%	+35.35%	-0.31%
	$S_{RS}$	+0.00%	+24.88%	+1.98%	+4.72%	+0.00%	+5.27%	+0.06%	+26.73%	+22.77%	+57.38%	+0.95%
	$S_{RI}$	+0.00%	+10.24%	+0.29%	+0.00%	+0.00%	+0.00%	+0.10%	+44.87%	+106.70%	+90.14%	+8.38%
	$S_{JA}$	+0.00%	+33.21%	+12.80%	+50.97%	+0.00%	+27.02%	+0.03%	+18.67%	+3.12%	+54.77%	+2.52%
	$S_{RS}^*$	-0.39%	+7.38%	-11.19%	-32.97%	+0.00%	-13.26%	+0.05%	+6.62%	+3.12%	-1.21%	-1.86%
	$S_{RI}^*$	-0.39%	+0.60%	-12.85%	-31.78%	+0.00%	-18.67%	+0.09%	+28.20%	+77.68%	+32.61%	+8.81%
	$S_{JA}^*$	+0.00%	+21.43%	+9.13%	+36.94%	+0.00%	+22.74%	+0.01%	+3.06%	-15.62%	+14.98%	+3.73%
Average	$S_P$	0.69	0.49	0.77	0.86	0.56	0.58	1.00	0.05	5.33	37.83	3622.61
	$S_E$	+6.38%	+15.19%	+7.91%	+13.22%	+7.28%	+16.13%	+0.01%	+26.58%	+1.04%	+11.78%	+0.07%
	$S_{RS}$	+3.74%	+15.36%	+3.64%	+4.93%	+5.32%	+7.45%	+0.04%	+39.31%	+29.95%	+38.57%	+0.37%
	$S_{RI}$	+0.25%	+8.79%	+0.84%	+1.17%	+1.28%	+3.75%	+0.07%	+82.94%	+84.46%	+38.73%	+10.11%
	$S_{JA}$	+8.58%	+24.77%	+14.82%	+26.78%	+11.11%	+21.26%	+0.02%	+30.88%	+13.54%	+18.26%	+1.75%
	$S_{RS}^*$	-9.78%	-4.98%	-7.81%	-20.07%	-0.94%	-6.44%	+0.04%	+7.28%	+14.24%	+13.49%	-1.24%
	$S_{RI}^*$	-8.22%	-5.37%	-8.43%	-18.21%	-5.28%	-9.41%	+0.07%	+58.48%	+56.81%	+11.91%	+12.06%
	$S_{JA}^*$	+5.83%	+8.54%	+8.98%	+18.69%	+4.79%	+9.11%	+0.00%	+3.38%	-2.43%	+0.83%	+1.75%

#### 4.4 RQ 2: Application

After validating the effectiveness of RACA, we explore its potential for practical applications under real-world safety testing scenarios. Following [71], we consider two application tasks: test suite prioritization and attack prompt sampling.

**Test Suite Prioritization.** This task involves filtering out redundant and invalid test prompts when selecting new ones from a large dataset. In practical deployments of LLMs, safety testing data may be extracted from a vast data stream that contains repeated or benign test cases. In this context, coverage criteria can be applied to ensure that only test cases that enhance coverage beyond a certain threshold are sampled. These cases are regarded as effective, while those below this threshold are considered ineffective. To simulate this application, we synthesized a large candidate pool consisting of 50% Normal, 40% Synonym, and 10% Invalid samples and a base set of 200 existing harmful prompts to establish the initial coverage. We compare RACA’s ensemble metrics (EI, EC, ER) against the neuron-level baseline (EN), and evaluate the proportion of normal examples recalled after this threshold sampling. After simple calibration, we set the threshold  $\tau = 0.01$  for RACA and  $\tau = 0.0005$  for EN.

Table 5: Ensemble metrics summary and comparison.

Model	Dataset Suite	SorryBench				ORBench			
		EI	EC	ER	EN	EI	EC	ER	EN
Vicuna	$S_E$	+7.06%	+8.18%	+7.62%	+8.27%	+8.01%	+10.15%	+9.08%	+5.88%
	$S_{RS}$	+2.22%	+0.38%	+1.30%	+2.81%	+5.37%	+7.24%	+6.31%	+19.31%
	$S_{RI}$	+0.14%	+1.87%	+1.00%	+16.72%	+3.04%	+1.98%	+2.51%	+33.03%
	$S_{JA}$	+4.95%	+10.32%	+7.63%	+6.23%	+11.88%	+13.11%	+12.50%	+7.42%
	$S_{RS}^*$	-12.31%	-6.68%	-9.50%	-7.63%	-9.84%	-4.74%	-7.29%	+6.88%
	$S_{RI}^*$	-6.58%	-4.07%	-5.33%	+11.74%	-6.94%	-8.67%	-7.80%	+28.41%
	$S_{JA}^*$	+2.98%	+3.03%	+3.00%	+1.49%	+1.03%	+2.88%	+1.95%	-0.85%
Llama	$S_E$	+5.28%	+4.36%	+4.82%	+7.97%	+13.72%	+11.82%	+12.77%	+7.36%
	$S_{RS}$	+1.53%	+2.89%	+2.21%	+2.90%	+8.42%	+7.14%	+7.78%	+24.06%
	$S_{RI}$	+1.27%	+0.10%	+0.69%	+17.29%	+3.34%	+4.21%	+3.77%	+46.72%
	$S_{JA}$	+9.50%	+7.03%	+8.27%	+9.04%	+20.95%	+20.05%	+20.50%	+15.43%
	$S_{RS}^*$	-10.33%	-5.43%	-7.88%	-0.64%	-11.33%	-7.31%	-9.32%	+11.96%
	$S_{RI}^*$	-7.43%	-5.38%	-6.41%	+19.70%	-13.41%	-7.42%	-10.41%	+33.16%
	$S_{JA}^*$	+2.93%	+0.61%	+1.77%	-0.01%	+17.22%	+16.49%	+16.85%	+1.73%
Qwen	$S_E$	+4.63%	+5.40%	+5.02%	+10.00%	+7.74%	+14.66%	+11.20%	+10.46%
	$S_{RS}$	+2.60%	+0.79%	+1.70%	+17.60%	+8.95%	+3.33%	+6.14%	+21.58%
	$S_{RI}$	+1.18%	+0.00%	+0.59%	+30.67%	+3.51%	+0.00%	+1.75%	+50.04%
	$S_{JA}$	+7.30%	+6.92%	+7.11%	+13.53%	+15.34%	+26.00%	+20.67%	+15.82%
	$S_{RS}^*$	-1.66%	-5.57%	-3.62%	+6.46%	-1.40%	-15.41%	-8.40%	+1.35%
	$S_{RI}^*$	-1.63%	-4.28%	-2.95%	+18.84%	-4.21%	-16.81%	-10.51%	+29.48%
	$S_{JA}^*$	+6.28%	+4.11%	+5.19%	+3.24%	+10.18%	+19.89%	+15.04%	+1.23%
Average	$S_E$	+5.66%	+5.98%	+5.82%	+8.75%	+9.82%	+12.21%	+11.02%	+7.90%
	$S_{RS}$	+2.12%	+1.35%	+1.74%	+7.77%	+7.58%	+5.90%	+6.74%	+21.65%
	$S_{RI}$	+0.86%	+0.66%	+0.76%	+21.56%	+3.30%	+2.06%	+2.68%	+43.26%
	$S_{JA}$	+7.25%	+8.09%	+7.67%	+9.60%	+16.06%	+19.72%	+17.89%	+12.89%
	$S_{RS}^*$	-8.10%	-5.89%	-7.00%	-0.60%	-7.52%	-9.15%	-8.34%	+6.73%
	$S_{RI}^*$	-5.21%	-4.58%	-4.90%	+16.76%	-8.19%	-10.97%	-9.57%	+30.35%
	$S_{JA}^*$	+4.06%	+2.58%	+3.32%	+1.58%	+9.48%	+13.09%	+11.28%	+0.70%

The results presented in Table 6 unequivocally demonstrate the superiority of RACA’s ensemble metrics for test case prioritization. The core objective in this task is to preferentially select novel, meaningful test cases (*i.e.*, Normal samples) while rejecting semantically redundant or invalid ones. The Normal Sample Proportion metric directly quantifies this capability. Across both benchmarks and all models, EI, EC, and ER consistently outperform the EN baseline by a significant margin. On average across the SorryBench dataset, RACA metrics achieve accuracies of 78% (EI), 68% (EC), and 70% (ER), which are approximately twice the 39% achieved by EN. This stark difference highlights EN’s fundamental weakness: it is easily deceived by superficial input variations that may trigger a few new neurons but contribute no new semantic concepts, thereby incorrectly accepting a large number of low-quality synonyms or invalid prompts. RACA, by contrast, correctly identifies these as low-value and rejects them. The trend continues on the ORBench dataset, where RACA’s average proportion (58% to 74%) is also clearly superior to EN’s 50%. The particularly strong performance

of EC on Qwen for ORBench (90%) underscores the robustness of our compositional criteria in identifying diverse and valuable test inputs. This evidence clearly supports the conclusion that RACA provides a far more reliable and precise signal for filtering large, noisy data streams, making it a highly effective tool for scalable and efficient LLM safety test suite maintenance.

Table 6: Test Suite Prioritization Results (Normal Sample Proportion).

Model	SorryBench				ORBench			
	EI	EC	ER	EN	EI	EC	ER	EN
Vicuna	78%	69%	74%	43%	65%	75%	77%	53%
Llama	74%	65%	62%	41%	50%	58%	57%	56%
Qwen	81%	70%	73%	33%	58%	90%	68%	40%
Average	78%	68%	70%	39%	58%	74%	67%	50%

**Attack Sample Acceleration.** The goal of this task is to prioritize successful jailbreak prompts from a mixed pool of successful and failed attempts. This simulates a common scenario in red-teaming where a large number of attack variants are generated, but only a subset effectively bypasses the model’s safety filters. The candidate pool for our experiment consists mixture of successful and failed jailbreak prompts. We apply the same thresholding technique as in the prioritization task and measure the Attack Success Rate (ASR) of the selected samples, which is the percentage of successful attacks within the prioritized set. A higher ASR indicates a more effective criterion for identifying potent attacks.

Table 7 provides compelling evidence that RACA is significantly more adept at identifying successful attacks than the neuron-level baseline. The average performance alone highlights a stark contrast: on SorryBench, RACA’s ensemble metrics achieve an ASR between 83% and 86%, whereas EN languishes at 53%, barely better than random chance in a 50/50 split pool. Similarly, on ORBench, RACA’s metrics (62%-71%) maintain a clear and consistent advantage over EN (47%). The most striking result is observed with the Qwen model on the SorryBench dataset, where both EC and ER achieve a perfect 100% ASR. This demonstrates that, under certain conditions, our compositional criteria can create a near-perfect filter, exclusively selecting prompts that successfully compromise the model while discarding all failed attempts. In this same scenario, EN’s ASR of 57% proves it is largely incapable of distinguishing the subtle but critical differences that determine an attack’s success. This discrepancy arises because a failed jailbreak may still activate a similar set of neurons as a successful one, thus confusing the EN metric. RACA, however, is sensitive not just to the presence of concepts but to their effective composition. A successful attack triggers a specific harmful representation that RACA is designed to detect, allowing it to precisely differentiate between effective and ineffective attack vectors, thereby accelerating the discovery of true vulnerabilities.

**Answer to RQ 2:** RACA effectively boosts test suite prioritization and attack prompt sampling, making it a practical guidance for facilitating real-world LLM safety testing.

## 4.5 RQ 3: Generalization

Finally, we assess the generalization ability of RACA across diverse settings and configurations, including scalability to larger models, selection of calibration set, layers for representation extraction, and parameter sensitivity.

Table 7: Attack Sampling Results (Attack Success Rate).

Model	SorryBench				ORBench			
	EI	EC	ER	EN	EI	EC	ER	EN
Vicuna	76%	91%	87%	53%	59%	61%	68%	45%
Llama	84%	58%	63%	48%	57%	68%	67%	48%
Qwen	97%	100%	100%	57%	69%	77%	77%	47%
Average	<b>86%</b>	83%	83%	53%	62%	69%	<b>71%</b>	47%

**Larger models.** In previous sections, we primarily focused our evaluation on 7B open-source models. However, our representation-based approach is model theoretically scalable to larger models, as the only computational overhead is forward passes to extract the representations. We further validate RACA’s effectiveness on a larger model, **Vicuna-13b**, using the same RQ1 setting. Table 8 presents the ensemble metrics comparison. The results confirm that RACA’s advantages scale with model size, where RACA metrics remain low for redundant inputs ( $S_{RS}$ ), while EN (especially NLC/TKNC) often overreacts to invalid inputs ( $S_{RI}$ ). Additionally, each of EI/EC/ER shows significant gains for Jailbreak Attacks ( $S_{JA}$ ), with EI increasing by +6.85% on SorryBench and +18.20% on ORBench, correctly identifying the high tendency of successful attacks.

Table 8: Vicuna-13b Performance (Ensemble Metrics) on SorryBench.

Model	Suite	SorryBench			
		EI	EC	ER	EN
Vicuna-13b	$S_E$	+1.96%	+0.84%	+1.40%	+4.67%
	$S_{RS}$	+1.24%	+3.99%	+2.61%	+12.05%
	$S_{RI}$	+0.76%	+0.00%	+0.38%	+26.47%
	$S_{JA}$	+6.85%	+8.91%	+7.88%	+5.43%
	$S_{RS}^*$	-5.73%	-5.10%	-5.42%	+10.87%
	$S_{RI}^*$	-4.10%	-4.34%	-4.22%	+29.72%
	$S_{JA}^*$	+1.63%	+1.46%	+1.54%	-1.69%

**Selection of Calibration Set.** RACA relies on a calibration set to extract the safety-critical representations. In practical scenarios, a small dataset representing the required safety constraints can serve this purpose. We evaluate the robustness of RACA metrics when the calibration set size is reduced from 100 (default) down to 10. Table 9 presents the results for the  $S_{JA}$  suite. We observe that though very small sample sizes (e.g., 10) are insufficient to capture the baseline distribution, even 20 samples can perform reasonable coverage results. This shows the robustness of RACA against the selection of calibration sets without careful preparation.

**Representation Extraction.** We analyze the impact of different layers and PCA components on the effectiveness of coverage analysis. Recall that we use ensembled middle layers to provide the safety representations. Table 10 further compares the performance of RACA metrics across layers 15-18 on Vicuna/SorryBench, where all layers consistently distinguish the Redundant suites with minimal gains, correctly reflecting their low semantic novelty. Meanwhile, all layers show higher sensitivity to jailbreak attacks replacement ( $S_{JA}^*$ ). This justifies our choice of using these layers or their ensemble.

Table 9: Calibration Set Size Sensitivity (with Vicuna on SorryBench).

Test Suite	Size 10			Size 20			Size 50			Size 100 (Base)		
	EI	EC	ER	EI	EC	ER	EI	EC	ER	EI	EC	ER
$S_E$	+2.23%	+1.18%	+1.71%	+4.85%	+2.56%	+3.70%	+5.26%	+6.82%	+6.04%	+5.80%	+5.72%	+5.76%
$S_{RS}$	+1.92%	+0.00%	+0.96%	+3.40%	+2.20%	+2.80%	+2.71%	+4.06%	+3.39%	+2.13%	+3.45%	+2.79%
$S_{RI}$	+0.13%	+0.00%	+0.06%	+0.36%	+4.09%	+2.23%	+0.71%	+1.55%	+1.13%	+0.35%	+0.09%	+0.22%
$S_{JA}$	+1.25%	+0.21%	+0.73%	+5.40%	+3.71%	+4.56%	+7.42%	+9.20%	+8.31%	+10.90%	+9.49%	+10.20%
$S_{RS}^*$	-4.96%	-1.64%	-3.30%	-7.98%	-1.11%	-4.55%	-5.12%	-2.52%	-3.82%	-15.34%	-3.33%	-9.33%
$S_{RI}^*$	-8.15%	-0.21%	-4.18%	-5.93%	-0.23%	-3.08%	-2.50%	-1.83%	-2.16%	-13.29%	-9.01%	-11.15%
$S_{JA}^*$	+1.42%	+0.00%	+0.71%	-2.04%	-0.89%	-1.47%	+0.48%	+6.58%	+3.53%	+6.87%	+10.10%	+8.48%

Table 10: Layer-wise Performance of RACA Metrics (with Vicuna on SorryBench).

Test Suite	Layer 15			Layer 16			Layer 17			Layer 18		
	EI	EC	ER									
$S_E$	+36.68%	+25.28%	+30.98%	+33.09%	+31.18%	+32.13%	+36.04%	+29.69%	+32.86%	+38.08%	+33.12%	+35.60%
$S_{RS}$	+4.82%	+6.75%	+5.79%	+3.70%	+10.60%	+7.15%	+1.13%	+8.12%	+4.62%	+5.43%	+12.07%	+8.75%
$S_{RI}$	+0.49%	+5.31%	+2.90%	+3.46%	+3.17%	+3.32%	+3.98%	+1.80%	+2.89%	+6.43%	+11.94%	+9.18%
$S_{JA}$	+28.82%	+22.37%	+25.60%	+15.01%	+16.99%	+16.00%	+25.89%	+13.83%	+19.86%	+29.75%	+21.25%	+25.50%
$S_{RS}^*$	-18.82%	-11.70%	-15.26%	-17.28%	-4.89%	-11.09%	-19.86%	-11.05%	-15.46%	-15.75%	-3.66%	-9.70%
$S_{RI}^*$	-22.63%	-14.13%	-18.38%	-11.93%	-2.60%	-7.27%	-15.59%	-17.96%	-16.78%	-17.01%	-9.76%	-13.39%
$S_{JA}^*$	-0.38%	+3.00%	+1.31%	-3.67%	+3.15%	-0.26%	+10.14%	+1.30%	+5.72%	+13.77%	+10.00%	+11.88%

**Parameter Sensitivity.** We evaluate the robustness of our metrics by analyzing their performance under different hyperparameter settings. This is crucial to ensure that the effectiveness of RACA is not an artifact of meticulous parameter tuning but a general property of our framework. Tables 11 and 12 present the model-averaged results on SorryBench for individual and compositional criteria, respectively, by varying one key hyperparameter for each criterion while keeping others at their default values.

The results demonstrate that RACA’s criteria are remarkably robust to hyperparameter variations. For the individual-dimension criteria shown in Table 11, the fundamental trends observed in RQ1 hold true across all tested configurations. For example, regardless of the choice of  $\epsilon$  for SFC,  $topk$  for TKFC, or  $bins$  for FIC, the coverage gain from the jailbreak suite ( $S_{JA}$ ) consistently and substantially exceeds the gains from the expanded ( $S_E$ ), synonym ( $S_{RS}$ ), and invalid ( $S_{RI}$ ) suites. Likewise, the replacement suites  $S_{RS}^*$  and  $S_{RI}^*$  consistently result in a drop in coverage, while  $S_{JA}^*$  yields a gain, confirming that the criteria’s ability to distinguish valuable tests from low-quality ones is not parameter-dependent.

A similar stability is observed for the compositional criteria in Table 12. SCC and PCC maintain their expected behavior across different cluster counts and distance thresholds. The one exception is the CBC metric, which shows some instability at a large boundary threshold ( $\delta = 16.0$ ), where the gain from the invalid suite  $S_{RI}$  (+24.00%) surpasses the gain from the expanded suite  $S_E$  (+15.07%). However, within the more moderate parameter ranges, it behaves as expected. Overall, this comprehensive analysis confirms that the core properties of the RACA framework are stable and reliable, reinforcing the validity of our findings without requiring extensive parameter optimization.

Table 11: Parameter Sensitivity on SorryBench (model-averaged): Individual-Dimensions.

Suite Config	SFC ( $\epsilon$ )			TKFC ( $topk$ )			FIC ( $bins$ )		
	3.0	5.0	8.0	1	2	5	5	10	20
$S_P$	0.9075	0.6719	0.3633	0.0898	0.4206	0.8945	0.8893	0.7724	0.6702
$S_E$	+1.52%	+8.54%	+11.93%	+4.69%	+14.76%	+3.02%	+2.58%	+4.66%	+5.89%
$S_{RS}$	+0.30%	+2.07%	+0.44%	+4.58%	+4.70%	+1.17%	+0.89%	+1.35%	+2.29%
$S_{RI}$	+0.00%	+0.00%	+0.90%	+2.37%	+4.58%	+1.42%	+0.29%	+0.53%	+1.39%
$S_{JA}$	+3.51%	+13.21%	+13.56%	+6.16%	+16.28%	+4.88%	+5.58%	+8.48%	+8.34%
$S_{RS}^*$	-2.24%	-8.29%	-9.95%	+1.24%	-13.12%	-0.81%	-5.41%	-4.32%	-2.97%
$S_{RI}^*$	-5.34%	-7.27%	-5.25%	-1.58%	-13.63%	-0.52%	-4.96%	-5.78%	-4.49%
$S_{JA}^*$	+0.87%	+7.05%	+16.59%	+4.09%	+9.26%	+3.66%	+3.62%	+3.05%	+4.46%

Table 12: Parameter Sensitivity on SorryBench (model-averaged): Compositional-Dimensions.

Suite Config	SCC ( $clusters$ )			PCC ( $\epsilon$ )			CBC ( $\delta$ )		
	16	32	64	1.5	2.5	4.0	4.0	8.0	16.0
$S_P$	0.8906	0.8229	0.6875	0.8914	0.6252	0.3995	0.8281	0.6120	0.1042
$S_E$	+3.64%	+6.83%	+14.71%	+2.68%	+5.31%	+13.67%	+3.06%	+3.02%	+15.07%
$S_{RS}$	+0.00%	+0.64%	+2.61%	+0.56%	+1.24%	+2.30%	+0.64%	+3.19%	+8.33%
$S_{RI}$	+0.00%	+0.00%	+2.15%	+1.45%	+0.19%	+0.43%	+2.22%	+0.86%	+24.00%
$S_{JA}$	+9.08%	+11.19%	+16.52%	+3.78%	+11.23%	+13.99%	+3.98%	+7.01%	+12.59%
$S_{RS}^*$	-2.26%	-6.52%	-4.23%	-1.49%	-5.45%	-10.37%	+1.20%	-0.30%	+28.15%
$S_{RI}^*$	-1.53%	-8.48%	-8.11%	-2.19%	-5.98%	-10.04%	+2.22%	-6.57%	+9.72%
$S_{JA}^*$	-0.33%	+2.67%	+11.82%	+1.26%	+4.93%	+14.88%	-1.14%	-0.47%	-4.70%

**Answer to RQ 3:** RACA demonstrates consistent performance across different model sizes, calibration set sizes, representation extraction layers, and different hyperparameter configurations, showing desirable generalization ability under diverse testing scenarios and making it a robust criterion for LLM safety testing.

## 5 Threat to Validity

### 5.1 External validity

External validity threats stem from the specificity of the experimental setup, which may limit the generalizability of RACA’s findings. We select a variety of evaluation settings to mitigate these threats. Specifically, the experiments include three mainstream 7B open-source LLMs with different training data and architectures, and further validate RACA on the 13B-scale Vicuna. This verifies scalability across model sizes and confirms consistent effectiveness regardless of parameter scale. Moreover, we employ two complementary safety benchmarks. SorryBench provides 44 fine-grained safety violation categories, while ORBench-Toxic includes thousands of prompts across 10 toxic categories. Alpaca’s benign prompts simulate invalid data, ensuring coverage of both harmful and non-harmful input scenarios. We also incorporate calibration set robustness testing, where

we evaluate RACA’s performance with calibration set sizes ranging from 10 to 100. These studies enhance the generalizability of our claims.

## 5.2 Internal validity

Internal validity is threatened by factors related to experimental design and implementation. To address this issue, the experiments compare RACA’s performance across layers 15-18, confirming stable trends in coverage results. Furthermore, the study tests key hyperparameters (e.g.,  $\epsilon = 3.0/5.0/8.0$  for SFC, clusters=16/32/64 for SCC). This demonstrates RACA’s insensitivity to minor parameter adjustments and ensures result reliability. Third, synthetic test suites are constructed with clear logical distinctions: Redundant-Semantic (synonymous prompts), Redundant-Invalid (benign prompts), and Jailbreak Attacks. Both expansion and replacement versions are used, with strict size control to eliminate confounding variables from input quantity differences.

# 6 Related Work

## 6.1 LLM Safety Testing

LLM developers are aligning models to reject malicious requests and prevent harmful outputs [41, 50]. However, jailbreak attacks [72, 61, 18, 11, 19] have emerged to break through such safety measures, compromising the security of LLMs to generate harmful content. Although defensive mechanisms [30, 40, 62, 25, 2, 66, 37, 70] are evolving to mitigate jailbreak issues, the problem remains unresolved under the fast development of jailbreak methods. This unaddressed issue raised the requirement for reliable LLM safety testing to measure model performance against malicious requests. Existing LLM safety testing primarily relies on black-box approaches with open-source benchmarks, like SafetyBench [65], SorryBench [57], to comprehensively evaluate the LLM safety configuration. Despite the advancement in black-box testing, attackers can generate unexpected malicious data outside the static benchmark to make successful attacks even if the model is proven safe in the test. This motivates the community to explore white-box testing to examine the internal behavior of LLM, which contributes to more effective testing and defenses.

## 6.2 Coverage Testing for AI

Coverage criteria [36, 32, 33] are prevalent in testing AI robustness, trustworthiness, fairness, etc., and can satisfy the requirement for a more thorough and comprehensive safety testing paradigm. They seek to capture neuron activation patterns in target models, uncovering subtle errors and vulnerabilities to reveal functional diversity and detect adversarial defects. For instance, neuron activation-based criteria, including Neuron Coverage (NC) [36] and Neuron Boundary Coverage (NBC) [32], measure the proportion of neuron activation to reflect the logic breadth of the model. Similarly, several criteria, including Top-K Neuron Coverage/Patterns (TKNC/TKNP) [32], pay attention to top neuron activations. Another type of criteria monitor neuron trajectory or causal features, including TensorFuzz Coverage (TFC) [33], Neuron Path Coverage (NPC) [58].

However, most of these criteria suffer from complexity bottlenecks and are designed only for small-scale DNNs, which are proven impractical for LLMs [71, 23]. Specifically, white-box testing coverage for LLMs differs significantly from traditional DNNs due to the vast input-output scope and larger architectures of LLMs, posing challenges for adapting existing white-box criteria. Very few works explored the safety criteria testing paradigm in LLMs. One of them [71] explored applying existing DNN criteria to testing LLM safety, and proposed comprehensive empirical insights into the issue, but is still limited to existing criteria to without specialization for LLM safety. AcTracer [23]

and LeCov [59] propose the criteria for LLM in truthfulness testing, yet they still mainly focus on neuron-level rather than representation-level evaluation, which is more suitable for LLMs. To summarize, the representation-level coverage criteria in LLM safety testing remain unexplored and require effective specialization.

## 7 Conclusion

In this paper, we propose RACA, a representation-aware coverage criterion tailored for LLM safety testing, addressing the scalability and irrelevance issues of traditional neuron-level criteria. RACA operates via three stages: safety-critical representation identification, conceptual activation calculation, and coverage computation with six sub-criteria spanning individual and compositional dimensions. Extensive experiments validate its effectiveness in identifying high-quality jailbreak prompts, practicality in test prioritization and attack sampling, and robustness across models and configurations. RACA outperforms baselines by focusing on safety-critical concepts and filtering noise, offering a reliable framework for LLM safety evaluation and advancing testing for AI techniques.

## References

- [1] Janice Ahn et al. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.
- [2] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.
- [3] Usman Anwar et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024.
- [4] Jinze Bai et al. Qwen technical report. <https://qwenlm.github.io/blog/qwen3/>, 2023.
- [5] Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [6] Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [7] Islem Bouzenia et al. Repairagent: An autonomous, llm-based agent for program repair. *arXiv preprint arXiv:2403.17134*, 2024.
- [8] Sviatoslav Chalnev et al. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*, 2024.
- [9] Patrick Chao et al. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.
- [10] Patrick Chao et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.
- [11] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025.
- [12] Canyu Chen et al. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, pages 354–368, 2024.
- [13] Huanran Chen et al. Towards the worst-case robustness of large language models. *arXiv preprint arXiv:2501.19040*, 2025.

- [14] Huanran Chen et al. Understanding pre-training and fine-tuning from loss landscape perspectives. *arXiv preprint arXiv:2505.17646*, 2025.
- [15] Tristan Coignion et al. A performance study of llm-generated code on leetcode. In *EASE*, 2024.
- [16] Justin Cui et al. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.
- [17] Josef Dai et al. Safe rlhf: Safe reinforcement learning from human feedback. In *ICLR*, 2024.
- [18] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*, 2023.
- [19] Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, and Yang Liu. Pandora: Jailbreak gpts by retrieval augmented generation poisoning. *arXiv preprint arXiv:2402.08416*, 2024.
- [20] Tianqi Du et al. Advancing llm safe alignment with safety representation ranking. *arXiv preprint arXiv:2505.15710*, 2025.
- [21] Daya Guo et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.
- [22] Daya Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [23] Yuheng Huang, Jiayang Song, Qiang Hu, Felix Juefei-Xu, and Lei Ma. Actracer: Active testing of large language model via multi-stage sampling. *ACM Transactions on Software Engineering and Methodology*, 2025.
- [24] Shima Imani et al. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.
- [25] Neel Jain et al. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.
- [26] Matthew Jin et al. Inferfix: End-to-end program repair with llms. In *FSE*, pages 1646–1656, 2023.
- [27] Jinhan Kim, Robert Feldt, and Shin Yoo. Guiding deep learning system testing using surprise adequacy. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 1039–1049. IEEE, 2019.
- [28] Jinhan Kim, Jeongil Ju, Robert Feldt, and Shin Yoo. Reducing dnn labelling cost using surprise adequacy: An industrial case study for autonomous driving. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1466–1476, 2020.
- [29] Tomasz Korbak et al. Pretraining language models with human preferences. In *ICML*, 2023.
- [30] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. Certifying llm safety against adversarial prompting. In *COLM*, 2023.
- [31] Yi Liu et al. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.
- [32] Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, et al. Deepgauge: Multi-granularity testing criteria for deep learning systems. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, pages 120–131, 2018.

- [33] Augustus Odena, Catherine Olsson, David Andersen, and Ian Goodfellow. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. In *International conference on machine learning*, pages 4901–4911. PMLR, 2019.
- [34] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.
- [35] Wenbo Pan et al. The hidden dimensions of llm alignment: A multi-dimensional safety analysis. *arXiv preprint arXiv:2502.09674*, 2025.
- [36] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.
- [37] Mansi Phute et al. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308*, 2023.
- [38] Xiangyu Qi et al. Safety alignment should be made more than just a few tokens deep. In *ICLR*, 2024.
- [39] Alec Radford et al. Improving language understanding by generative pre-training. 2018.
- [40] Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.
- [41] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey, 2023.
- [42] Xinyue Shen et al. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *CCS*, 2023.
- [43] Oscar Skean et al. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.
- [44] Alessandro Stolfo et al. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*, 2024.
- [45] Rohan Taori et al. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [46] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [47] Ashish Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.
- [48] Cheng Wang, Zeming Wei, Qin Liu, and Muhao Chen. False sense of security: Why probing-based malicious input detection fails to generalize. *arXiv preprint arXiv:2509.03888*, 2025.
- [49] Hanyu Wang, Bochuan Cao, Yuanpu Cao, and Jinghui Chen. Truthflow: Truthful llm generation via representation flow correction. *arXiv preprint arXiv:2502.04556*, 2025.
- [50] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more, 2024.
- [51] Zhichao Wang et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024.
- [52] Alexander Wei et al. Jailbroken: How does llm safety training fail? In *NeurIPS*, 2023.
- [53] Boyi Wei et al. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

- [54] Zeming Wei et al. Position: Agent-specific trustworthiness risk as a research priority. *OpenReview preprint*, 2025.
- [55] Zeming Wei, Chengcan Wu, and Meng Sun. Rega: Representation-guided abstraction for model-based safeguarding of llms. *arXiv preprint arXiv:2506.01770*, 2025.
- [56] Chengcan Wu et al. Mitigating fine-tuning risks in llms via safety-aware probing optimization. *arXiv preprint arXiv:2505.16737*, 2025.
- [57] Tinghao Xie et al. Sorry-bench: Systematically evaluating large language model safety refusal. In *ICLR*, 2025.
- [58] Xiaofei Xie, Tianlin Li, Jian Wang, Lei Ma, Qing Guo, Felix Juefei-Xu, and Yang Liu. Npc: N euron p ath c overage via characterizing decision logic of deep neural networks. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(3):1–27, 2022.
- [59] Xuan Xie, Jiayang Song, Yuheng Huang, Da Song, Felix Juefei-Xu, and Lei Ma. Lecov: Multi-level testing criteria for large language models. *Journal of Systems and Software*, page 112763, 2025.
- [60] Yifan Yao et al. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 2024.
- [61] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.
- [62] Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Xiaofei Xie, Yang Liu, and Chao Shen. A mutation-based method for multi-modal jailbreaking attack detection. *CoRR*, 2023.
- [63] Yedi Zhang et al. The fusion of large language models and formal methods for trustworthy ai agents: A roadmap. *arXiv preprint arXiv:2412.06512*, 2024.
- [64] Yihao Zhang et al. Adversarial representation engineering: A general model editing framework for large language models. *arXiv preprint arXiv:2404.13752*, 2024.
- [65] Zhixin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, 2024.
- [66] Ziyang Zhang, Qizhen Zhang, and Jakob Foerster. Parden, can you repeat that? defending against jailbreaks via repetition. *arXiv preprint arXiv:2405.07932*, 2024.
- [67] Chujie Zheng et al. Prompt-driven llm safeguarding via directed representation optimization. *arXiv preprint arXiv:2401.18018*, 2024.
- [68] Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.
- [69] Tianyang Zhong et al. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*, 2024.
- [70] Qi Zhou, Tianlin Li, Qing Guo, Dongxia Wang, Yun Lin, Yang Liu, and Jin Song Dong. Defending lvlms against vision attacks through partial-perception supervision. *arXiv preprint arXiv:2412.12722*, 2024.
- [71] Shide Zhou, Tianlin Li, Kailong Wang, Yihao Huang, Ling Shi, Yang Liu, and Haoyu Wang. Understanding the effectiveness of coverage criteria for large language models: A special angle from jailbreak attacks. *arXiv preprint arXiv:2408.15207*, 2024.

- [72] Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. Autodan: Interpretable gradient-based adversarial attacks on large language models. In *COLM*, 2023.
- [73] Andy Zou et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.
- [74] Andy Zou et al. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.