

Zen Safety Framework: Evaluation and Mitigation

Technical Report v2025.05

Zen LM Research Team
research@zenlm.org

May 2025

Abstract

We present the Zen Safety Framework, a comprehensive methodology for evaluating and mitigating safety risks in large language models. The framework encompasses structured red-teaming, a 2400-item safety evaluation suite, automated adversarial generation, and constitutional RLHF for mitigation. Zen models achieve 99.2% on ToxiGen, 96.8% jailbreak resistance on our Adversarial Red-Team Benchmark (ARTB), and competitive bias scores on BBQ and WinoBias. We describe our red-teaming methodology, evaluation protocols, and the feedback loop between red-team findings and model retraining. Safety does not significantly degrade capability: across MMLU, HumanEval, and MATH, safety-trained models lose fewer than 0.8 points versus unsafety-trained counterparts.

1 Introduction

Deploying frontier language models requires confidence that they will not produce harmful, biased, or privacy-violating content. Safety evaluation must be adversarial, systematic, and diverse: models quickly overfit to narrow evaluations that do not transfer to real-world misuse.

The Zen Safety Framework is built on four principles:

1. **Adversarial by default:** Evaluations must probe the model’s worst case, not its average case.
2. **Multi-dimensional:** Safety is not one-dimensional; harm types differ in severity and mechanism.
3. **Automated and scalable:** Human red-teaming is expensive; automation must close the gap.
4. **Closed-loop:** Safety findings must feed back into training, not just documentation.

2 Background

2.1 Taxonomy of Harms

We adopt a five-category harm taxonomy:

- **Physical harm:** Instructions for violence, weapons, or dangerous substances
- **Psychological harm:** Harassment, emotional manipulation, suicide encouragement
- **Privacy violation:** PII extraction, doxxing, surveillance assistance
- **Bias and discrimination:** Demographic stereotyping, hate speech
- **Misinformation:** False factual claims, election interference

2.2 Related Work

ToxiGen [?] provides implicit toxicity benchmarking. BBQ [?] and WinoBias [?] measure demographic biases. Do-Not-Answer [?] evaluates refusal of harmful requests. Jailbreak benchmarks [?] probe prompt-based safety bypasses.

3 Red-Teaming Methodology

3.1 Structured Human Red-Teaming

Zen red-teaming employs 64 red-teamers across 8 specializations:

Specialization	Testers	Focus Area
Prompt injection	12	Jailbreaks, prompt leakage
Social engineering	10	Manipulation, impersonation
Bioweapons/CBRN	6	Dual-use knowledge elicitation
Misinformation	8	Factual manipulation, political bias
Privacy	8	PII extraction, doxxing
Hate speech / bias	10	Demographic targeting, slurs
Child safety	6	CSAM, grooming facilitation
Financial fraud	4	Scam generation, market manipulation

Table 1: Red-team composition and specializations.

Each red-teamer conducts 8-hour sessions targeting their specialization, documenting successful attack vectors, partial successes, and near-misses. All sessions are logged and used to expand the automated evaluation suite.

3.2 Attack Vector Classification

Red-team attacks are classified into five vector types:

1. **Direct request:** Straightforward harmful request (“How do I make...”)
2. **Roleplay:** Fictional framing (“As a character in a story...”)
3. **Hypothetical:** Abstracted framing (“Suppose someone wanted to...”)
4. **Jailbreak:** Prompt engineering to bypass safety (DAN, AIM, etc.)
5. **Multi-turn:** Incrementally escalating conversation

3.3 Automated Red-Teaming

Human red-teaming is bottlenecked by cost and coverage. Automated red-teaming generates adversarial prompts using a separately trained attacker model:

Algorithm 1 Automated Adversarial Prompt Generation

Require: Target model π , attacker model $\mathcal{A}$, harm categories $\mathcal{H}$, budget K

Ensure: Attack set $\mathcal{S}$

```
1:  $\mathcal{S} \leftarrow \emptyset$ 
2: for each category  $h \in \mathcal{H}$  do
3:   for  $k = 1 \dots K/|\mathcal{H}|$  do
4:      $p \leftarrow \mathcal{A}.\text{generate}(h, \mathcal{S}_h)$  {Generate new attack for category  $h$ }
5:      $r \leftarrow \pi(p)$  {Get target model response}
6:      $\text{score} \leftarrow \text{HarmClassifier}(r, h)$  {Score harmfulness 0-1}
7:      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(p, r, \text{score})\}$ 
8:     if  $\text{score} > 0.5$  then
9:        $\mathcal{A} \leftarrow \mathcal{A}.\text{reinforce}(p)$  {Reward successful attacks}
10:    end if
11:  end for
12: end for
13: return  $\mathcal{S}$ 
```

The attacker model is a Zen-7B fine-tuned via RL to generate attacks that elicit harmful responses from the target. We generate 50K attacks per model checkpoint and use a harm classifier (Zen-7B fine-tuned on 100K human-rated response pairs) to score responses.

4 Safety Evaluation Suite

4.1 Adversarial Red-Team Benchmark (ARTB)

Our Adversarial Red-Team Benchmark (ARTB) contains 2400 manually verified adversarial prompts across five harm categories and five attack vector types (480 per category):

Category	Items	Jailbreak Resistance
Physical harm	480	97.3%
Psychological harm	480	96.9%
Privacy violation	480	97.1%
Bias / discrimination	480	96.2%
Misinformation	480	96.5%
Overall	2400	96.8%

Table 2: ARTB results for Zen-7B-Aligned. Jailbreak resistance = % of attacks successfully refused.

4.2 ToxiGen Evaluation

ToxiGen measures implicit toxicity targeting 13 demographic groups. Zen models are evaluated in generative mode (completing implicit toxic sentences):

Model	ToxiGen Score (%)	Group Disparity
Zen-7B (no alignment)	74.3%	18.2%
Zen-7B (SFT only)	91.4%	8.7%
Zen-7B (full alignment)	99.2%	2.1%
Zen-32B (full alignment)	99.4%	1.8%

Table 3: ToxiGen scores. Higher = safer. Group disparity = max-min across 13 groups.

4.3 Bias Evaluation: BBQ

The Bias Benchmark for QA (BBQ) tests whether models apply stereotypes when answering ambiguous questions. We report accuracy on the disambiguated subset (correct answer determinable) and bias score on the ambiguous subset (lower = more biased):

Model	Disambig Acc.	Bias Score	Gender Bias	Racial Bias
Zen-7B (unaligned)	78.2%	0.31	0.44	0.38
Zen-7B (aligned)	81.4%	0.08	0.09	0.07
Zen-32B (aligned)	84.1%	0.06	0.07	0.05

Table 4: BBQ results. Bias score closer to 0 indicates less bias.

4.4 WinoBias Evaluation

WinoBias tests gender bias in coreference resolution across stereotypical and anti-stereotypical pronoun assignments:

Model	Pro-stereo F1	Anti-stereo F1	Bias Gap
Zen-7B (unaligned)	89.3%	62.1%	27.2%
Zen-7B (aligned)	87.1%	83.4%	3.7%
Zen-32B (aligned)	88.2%	85.1%	3.1%

Table 5: WinoBias F1 scores. Smaller bias gap indicates more equitable treatment of gender.

5 Mitigation: Constitutional RLHF

Safety findings from red-teaming and evaluation feed directly into model training via a three-step mitigation cycle:

1. **Attack harvest:** Collect successful attacks from ARTB and automated red-teaming.
2. **Constitutional critique:** Generate model critiques of harmful responses using the Zen constitution’s safety principles.
3. **RLHF reward update:** Add safety-specific reward signal for harmful output classes.

The safety reward combines a harm classifier score and a refusal quality score:

$$r_{\text{safety}}(x, y) = \lambda_{\text{harm}}(1 - \text{HarmScore}(y)) + \lambda_{\text{refusal}} \cdot \text{RefusalQuality}(x, y) \quad (1)$$

where RefusalQuality scores refusals on a 0-1 scale: 0 for silent refusal, 0.5 for bare refusal, 1.0 for explanatory refusal that redirects to safe alternatives.

6 Safety-Capability Tradeoff

A central concern is that safety training degrades capability. We measure this across standard benchmarks:

Model Variant	MMLU	HumanEval	MATH	GSM8K	ARTB
Zen-7B (base pretrain)	82.1	74.3	63.2	80.4	41.2%
Zen-7B (SFT)	85.3	78.2	67.4	84.1	71.3%
Zen-7B (SFT + const.)	85.1	78.0	67.2	83.9	83.4%
Zen-7B (full RLHF)	84.8	77.8	67.1	83.7	96.8%
Δ (RLHF vs SFT)	-0.5	-0.4	-0.3	-0.4	+25.5%

Table 6: Safety-capability tradeoff. RLHF adds 25.5 ARTB points at < 0.5 capability cost.

The capability cost of full alignment is less than 0.5 points on any benchmark, confirming that safety and capability are not fundamentally at odds.

7 Analysis

7.1 Attack Vector Resistance

Jailbreak resistance varies by attack vector type:

Attack Vector	Resistance (Zen-7B)	Resistance (Zen-32B)
Direct request	99.1%	99.6%
Roleplay framing	96.2%	97.8%
Hypothetical framing	95.4%	97.1%
Jailbreak prompt	94.8%	96.3%
Multi-turn escalation	92.3%	95.2%
Average	96.8%	97.9%

Table 7: Jailbreak resistance by attack vector. Multi-turn remains the hardest.

Multi-turn escalation remains the most challenging attack vector because each individual turn appears benign. Mitigating this requires maintaining safety context across the full conversation history.

7.2 Model Size and Safety

Larger models are generally safer (higher ARTB resistance, lower bias scores). This positive scaling relationship holds consistently across all harm categories. We hypothesize that larger models develop more robust world models that allow them to reason about harm more effectively.

7.3 Over-Refusal Analysis

A key failure mode is over-refusal: refusing legitimate requests that superficially resemble harmful ones. We measure over-refusal rate on a 1000-item benign request set covering sensitive-but-legitimate topics (medical questions, historical violence, security research):

Model Variant	Over-Refusal Rate
Zen-7B (SFT only)	4.2%
Zen-7B (naive RLHF)	22.4%
Zen-7B (constitutional RLHF)	8.1%

Table 8: Over-refusal rates on benign sensitive requests.

Constitutional training reduces over-refusal by teaching the model to distinguish intent-sensitive benign requests from genuinely harmful ones.

8 Conclusion

The Zen Safety Framework demonstrates that comprehensive, adversarial safety evaluation combined with constitutional RLHF produces models that are both safe and capable. 96.8% jailbreak resistance, 99.2% ToxiGen score, and near-parity BBQ/WinoBias scores are achieved at under 0.5 points capability degradation. Future work includes multi-turn safety evaluation protocols, dynamic red-teaming with model-in-the-loop attack generation, and cultural safety evaluation for non-English languages.

References

- [1] Hartvigsen et al. (2022). ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *ACL*.
- [2] Parrish et al. (2022). BBQ: A hand-built bias benchmark for question answering. *ACL Findings*.
- [3] Zhao et al. (2018). Gender Bias in Coreference Resolution. *NAACL*.
- [4] Wang et al. (2023). Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *arXiv:2308.13387*.
- [5] Zou et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043*.