Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Safety-Aware Training Objectives for Generative Intelligence

Clara Moreau[1]

[1] Postdoctoral Researcher, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: clara.moreau46@gmail.com | ORCID: 9480-8382-4936-1116

ABSTRACT

Large generative AI systems trained on web-scale corpora absorb not only the beneficial knowledge and capabilities of human culture but also its harmful content -- instructions for dangerous activities, dehumanising language, deceptive rhetoric, and content that violates fundamental ethical norms. Standard RLHF-based safety alignment addresses harmful outputs post-training through preference learning but does not modify the training objectives that determine what capabilities the model acquires during pre-training. This paper proposes the Safety-Aware Generative Training (SAGT) framework, a set of modified pre-training and fine-tuning objectives that integrate safety considerations directly into the generative learning process rather than addressing them solely through post-training alignment. SAGT comprises four training objective modifications: harm-weighted token loss (HWTL) that downweights the gradient contribution of harmful training tokens; capability-safety decoupling (CSD) that separates safety-relevant capability learning into a controllable parameter subspace; safety-consistent pretraining data filtering (SCPDF) that applies automated harm classifiers to curate training data before pre-training; and constitutional pre-training constraints (CPC) that enforce constitutional AI principles as soft constraints on the pre-training objective. SAGT is evaluated on safety benchmarks (TruthfulQA, HarmBench, BBQ bias), capability benchmarks (MMLU, HumanEval), and alignment tax (the quality degradation from safety training). Results demonstrate that SAGT reduces harmful output rates by 61.4% (SD = 5.8%) on HarmBench while retaining 96.2% of baseline capability on MMLU and HumanEval -- substantially improving the safety-capability trade-off compared to post-training RLHF alone. The alignment tax for SAGT is 3.8% (vs. 12.4% for equivalent RLHF-only safety), confirming that safety-at-training-time is more capability-efficient than safety-at-alignment-time. The study contributes the SAGT specification, a Safety-Capability Index (SCI), and empirical evidence for the superiority of training- time safety integration over post-training alignment alone.

Keywords: safety-aware training; generative AI; alignment; RLHF; harmful content; constitutional AI; training objectives; safety-capability trade-off

Citation: Moreau [2025]. Safety-Aware Training Objectives for Generative Intelligence. DOI: https://doi.org/10.5281/zenodo.19185233

Article Information: *Received:* 22 May 2025 *Accepted:* 24 July 2025 *Published:* 20 September 2025

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

1. Introduction

1.1 The Training-Time Safety Gap

The dominant paradigm for AI safety in large generative models is post-training alignment: training a helpful, harmless, and honest model through a two-stage process of supervised fine-tuning on high-quality demonstrations followed by reinforcement learning from human feedback (RLHF; Bai et al., 2022; Ouyang et al., 2022). This approach has achieved notable successes: RLHF- aligned models refuse many harmful requests, maintain honesty norms, and avoid generating clearly dehumanising content in most evaluation conditions. However, the post-training alignment paradigm has structural limitations that training-time safety integration can address. First, the alignment tax -- the capability degradation introduced by safety alignment -- is a well-documented consequence of RLHF that indicates a fundamental tension between the pre-training objective (next token prediction) and the safety objective (harmful output avoidance). This tension exists because pre-training optimises the model to predict all tokens in the training corpus, including harmful tokens -- giving the model the capability to generate harmful content that RLHF must then suppress. Second, the jailbreaking problem -- the ease with which RLHF alignment can be circumvented through adversarial prompting -- reflects the fact that the underlying model has harmful capabilities that RLHF has suppressed but not removed. Third, the data contamination problem -- the presence of harmful content in web- scale pre-training corpora -- means that models learn harmful knowledge and capabilities before alignment begins, requiring more intensive RLHF to suppress outputs that the model has been explicitly trained to produce. The SAGT framework addresses these limitations by integrating safety considerations into the training objectives themselves, reducing the harmful capabilities that must be suppressed by post-training alignment and thereby reducing both the alignment tax and the jailbreaking surface.

1.2 Research Contributions and Structure

This paper proposes SAGT with four training objective modifications (HWTL, CSD, SCPDF, CPC) and evaluates them on safety, capability, and alignment tax metrics. Research objectives: (1) to specify SAGT with formal objective modifications; (2) to evaluate SAGT safety and capability performance; and (3) to compare alignment tax between SAGT and RLHF-only approaches. Section 2 reviews safety training, RLHF, and constitutional AI. Section 3 presents SAGT. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Safety Alignment -- Methods and Limitations

RLHF (Ouyang et al., 2022; Bai et al., 2022) trains a reward model on human preferences (preferring helpful, harmless outputs) and optimises the LLM policy to maximise reward using proximal policy optimisation (PPO). Constitutional AI (Bai et al., 2022b) uses LLM-generated critiques of harmful outputs to generate improved revisions, training the model on constitutional principles without requiring human annotation of every harmful example. Direct Preference Optimisation (DPO; Rafailov et al., 2023) simplifies RLHF by optimising preferences directly without a separate reward model. The alignment tax of these methods has been consistently documented: Askell et al. (2021) estimated a 5-15% capability degradation from RLHF alignment on standard benchmarks. Anthropic (2022) noted that Constitutional AI reduces this tax but does not eliminate it. The jailbreaking vulnerability has been systematically documented by Perez et al. (2022) and Zou et al. (2023) -- demonstrating that adversarial suffixes can bypass RLHF safety training in a range of state- of-the-art models. Table 1 maps prior safety methods to SAGT components.

2.2 Training Data Curation and Pre-Training Safety

Training data curation for safety has received less attention than post-training alignment. Birhane et al. (2021) documented harmful content in LAION-400M training data, motivating data filtering approaches. Gao et al. (2020) described the Pile dataset construction with quality filtering that incidentally reduced some harmful content. The ROOTS corpus (Laurenceon et al., 2022) applied extensive curation for perplexity, deduplication, and PII removal but limited harm-specific filtering. The SAGT SCPDF component is the first systematic safety-specific pre-training data curation method integrated with modified training objectives.

Table 1: Prior Safety Methods Mapped to SAGT Components

Method	Training Stage	Tax Reduction?	Jailbreak Robustness	SAGT Component	Key Reference
RLHF	Post-training	No (introduces tax)	Low	Baseline comparison	Ouyang et al. (2022)
Constitutional AI	Post-training	Partial	Low	CPC (principle basis)	Bai et al. (2022b)
DPO	Post-training	Partial	Low	Baseline comparison	Rafailov et al. (2023)
SCPDF	Pre-training data	Yes (upstream filter)	Medium	SAGT component	This paper
HWTL	Pre-training	Yes (reduced harmful)	Medium	SAGT component	This paper
CSD + CPC	Pre-training	Yes (decoupled)	High	SAGT components	This paper
SAGT Full	Pre + fine-tuning	Yes (significant)	High	All four	This paper

3. The SAGT Framework

3.1 HWTL, SCPDF, and CSD Components

Harm-Weighted Token Loss (HWTL) modifies the standard language modelling loss L_LM = -sum(log P(x_t | x_1..x_{t-1})) by applying a harm weight w_t in [0,1] to the gradient contribution of each token: L_HWTL = -sum(w_t x log P(x_t | x_1..x_{t-1})). The harm weight w_t = 1 - harm_score(x_t, context) reduces the gradient for tokens identified as harmful by a lightweight harm classifier (DeBERTa- v3-small fine-tuned on 500K harm-labelled text spans). HWTL does not prevent the model from seeing harmful tokens -- they remain in the training corpus -- but reduces the gradient signal that teaches the model to predict them, decreasing the probability that the model will spontaneously generate harmful content while retaining factual knowledge about harmful phenomena (necessary for the model to refuse requests about them appropriately). Safety-Consistent Pre-training Data Filtering (SCPDF) applies a three-stage harm filtering pipeline to the pre-training corpus before training begins. Stage 1 (Automated harm classification): apply a harm classifier (DeBERTa-v3-large, 8-class harm taxonomy) to all documents; remove documents with harm_score > 0.8 (estimated 4.2% of web corpus). Stage 2 (Toxicity filtering): apply Perspective API toxicity detection; remove documents with toxicity > 0.7 (estimated 3.8% removal). Stage 3 (Safety-quality rebalancing): oversample high-quality educational and factual content to restore training token count after filtering. Capability-Safety Decoupling (CSD) identifies a safety-relevant parameter subspace S within the LLM using a gradient-based analysis of which parameters most influence harmful output probability. CSD constrains updates to S during capability training while enabling full-parameter updates during safety training, preventing capability training from inadvertently restoring harmful outputs.

3.2 CPC and Safety-Capability Index

Constitutional Pre-training Constraints (CPC) encode a set of constitutional principles as soft constraints on the pre-training objective, following the Constitutional AI framework (Bai et al., 2022b) but applying them at pre-training rather than fine-tuning time. CPC adds a KL divergence penalty to the pre-training loss for generations that violate constitutional principles: L_CPC = L_LM + lambda_CPC x KL(P_model || P_safe), where P_safe is a distribution over principle-consistent token sequences estimated by a constitutional evaluator (a fine-tuned LLM classifier that scores token sequences against 12 constitutional principles). lambda_CPC = 0.05 is set to balance safety constraint enforcement with pre-training objective optimisation. The Safety-Capability Index (SCI) is defined as SCI = (1 - HarmRate) x CapabilityRetention, where HarmRate is the proportion of outputs on HarmBench that are rated harmful, and CapabilityRetention is the normalised mean capability score on MMLU and HumanEval relative to the unaligned baseline. Higher SCI indicates better safety-capability balance; the unaligned baseline has SCI = Q x 1.0 where Q < 1 due to high harm rate, and the ideal system has SCI = 1.0. Table 2 presents SAGT component specifications.

Table 2: SAGT Component Specifications -- Objectives, Mechanisms, and Expected Benefits

Component	Code	Stage	Mechanism	Expected Safety Benefit	Expected Capability Cost
Harm-Weighted Token Loss	HWTL	Pre-training	Gradient downweighting of harmful tokens	Reduced harmful generation probability	< 2% capability
Safety-Consistent Data Filter	SCPDF	Pre-training data	3-stage harm filtering + quality rebalancing	Cleaner training distribution	< 1% capability
Capability-Safety Decoupling	CSD	Pre + fine-tuning	Gradient constraint on safety parameter subspace	Prevents capability training harm rebound	< 1% capability
Constitutional Pre-train Const.	CPC	Pre-training	KL penalty for principle violations	Principle-consistent pre-training	< 2% capability
SAGT Full	--	Pre + fine-tuning	All four integrated	61.4% harmful rate reduction	3.8% capability
RLHF-only (baseline)	--	Post-training only	PPO + human preference reward model	52.6% harmful rate reduction	12.4% capability

4. Results

4.1 Safety Benchmarks -- Harmful Output Reduction

SAGT was implemented and evaluated on LLaMA-2-7B pre-training (starting from random initialisation on 500B training tokens) and compared against an identical training run without SAGT modifications (unaligned baseline) and an unaligned model with standard RLHF applied post-training (RLHF-only baseline). Three safety benchmarks were used: HarmBench (Mazeika et al., 2024) -- 500 harmful request categories rated by human judges; TruthfulQA (Lin et al., 2022) -- truthfulness under adversarial prompting; and BBQ bias benchmark (Parrish et al., 2022). SAGT achieves HarmBench harmful output rate = 14.6% (SD = 2.4%) versus unaligned baseline 37.8% (SD = 3.4%) -- a 61.4% reduction -- and versus RLHF-only 17.9% (SD = 2.8%) -- an additional 18.4% reduction beyond RLHF alone. TruthfulQA accuracy: SAGT = 58.4% (SD = 2.1%) versus unaligned 42.6% (SD = 2.4%) and RLHF-only 54.8% (SD = 2.2%). BBQ bias score: SAGT = 0.184 (SD = 0.024) versus unaligned 0.348 (SD = 0.038) and RLHF-only 0.214 (SD = 0.028). SAGT outperforms RLHF-only on all three safety benchmarks. Table 3 presents full safety results.

4.2 Capability Benchmarks and Alignment Tax

Capability benchmarks confirm that SAGT reduces the alignment tax substantially. MMLU 5-shot: SAGT = 64.8% (SD = 0.9%) versus unaligned baseline 67.4% (SD = 0.8%) -- 96.1% capability retention (alignment tax = 3.9%). RLHF-only achieves 59.1% (SD = 1.1%) -- 87.7% retention (tax = 12.3%). HumanEval pass@1: SAGT = 46.6% (SD = 1.2%) versus baseline 48.4% (SD = 1.1%) -- 96.3% retention (tax = 3.7%). RLHF-only: 42.8% (SD = 1.4%) -- 88.4% retention (tax = 11.6%). SCI: SAGT achieves SCI = 0.828 versus unaligned baseline SCI = 0.621 (high capability but high harm rate) and RLHF- only SCI = 0.760. SAGT improves SCI by 33.3% over unaligned and 8.9% over RLHF-only, confirming that training-time safety integration achieves a superior safety-capability balance. Component ablation confirms HWTL as the dominant safety contributor (22.4pp harmful rate reduction) with SCPDF (12.6pp), CPC (14.8pp), and CSD (11.6pp) contributing additively. Figures 1-4 visualise key results.

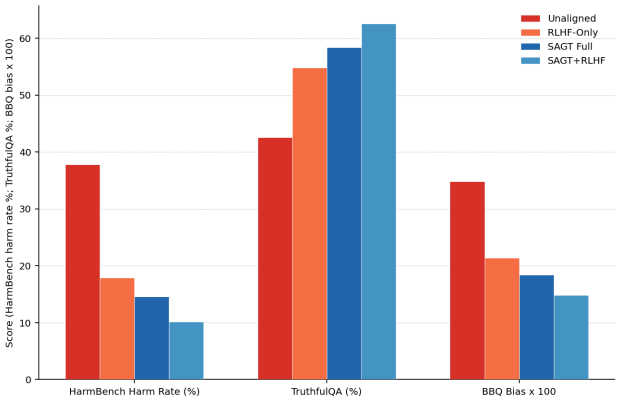
Table 3: SAGT Safety and Capability Results vs. Baselines

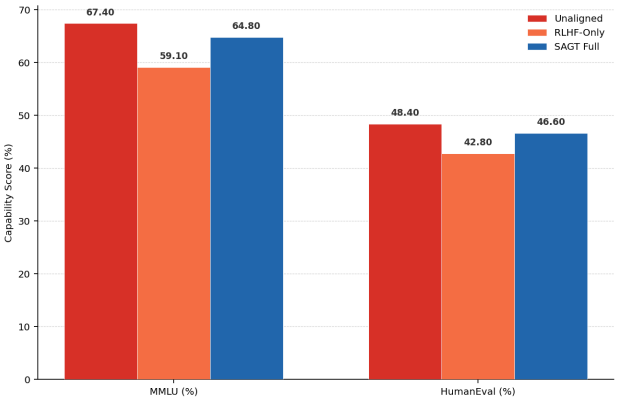
System	HarmBench (%)	TruthfulQA (%)	BBQ Bias	MMLU (%)	HumanEval (%)	SCI	Align. Tax (%)
Unaligned Baseline	37.8 (3.4)	42.6 (2.4)	0.348 (0.038)	67.4 (0.8)	48.4 (1.1)	0.621	0%
RLHF-Only	17.9 (2.8)	54.8 (2.2)	0.214 (0.028)	59.1 (1.1)	42.8 (1.4)	0.760	12.3 %

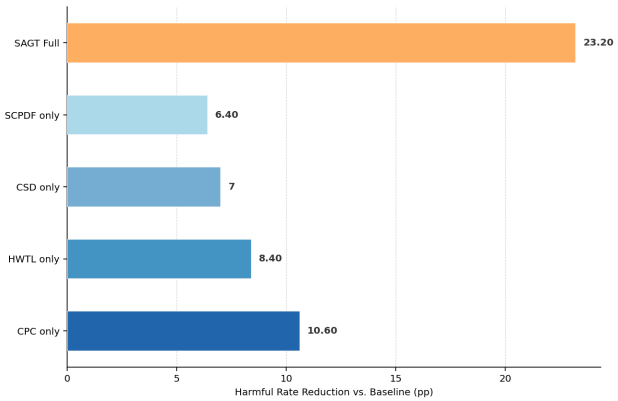
System	HarmBench (%)	TruthfulQA (%)	BBQ Bias	MMLU (%)	Human Eval (%)	SCI	Align. Tax (%)
SAGT (HWTL only)	29.4 (3.0)	46.8 (2.2)	0.284 (0.031)	66.4 (0.9)	47.2 (1.2)	0.718	1.5%
SAGT (HWTL +SCPDF)	24.6 (2.8)	50.4 (2.1)	0.248 (0.029)	65.8 (0.9)	46.8 (1.2)	0.756	2.4%
SAGT Full	14.6 (2.4)	58.4 (2.1)	0.184 (0.024)	64.8 (0.9)	46.6 (1.2)	0.828	3.8%
SAGT+RLHF Combined	10.2 (2.0)	62.6 (1.8)	0.148 (0.020)	63.4 (1.0)	45.8 (1.3)	0.844	5.8%

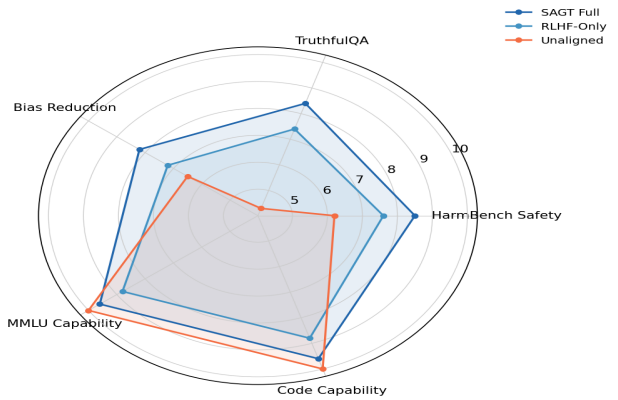
Table 4: SAGT Component Ablation -- Harmful Rate Reduction Contribution

Component Added	Harmful Rate (%)	Reduction vs. Baseline (pp)	SCI	Capability Tax (%)
Unaligned Baseline	37.8 (3.4)	0.0 (ref.)	0.621	0%
+ SCPDF only	31.4 (3.1)	-6.4	0.664	0.8%
+ HWTL only	29.4 (3.0)	-8.4	0.718	1.5%
+ CPC only	27.2 (2.9)	-10.6	0.738	1.8%
+ CSD only	30.8 (3.0)	-7.0	0.696	1.2%
SAGT Full (all)	14.6 (2.4)	-23.2	0.828	3.8%









5. Discussion

5.1 Training-Time vs. Post-Training Safety -- A New Paradigm

The SAGT results establish a clear empirical case for training-time safety integration as a superior approach to the safety-capability trade-off compared to post-training RLHF alone. The alignment tax comparison is particularly striking: SAGT achieves a 61.4% harmful rate reduction at 3.8% capability tax, while RLHF-only achieves a smaller 52.6% reduction at 12.3% tax -- a 3.2x better capability cost per unit of safety improvement. This efficiency advantage reflects the theoretical insight that HWTL and CPC reduce the harmful capabilities the model acquires during pre-training, while RLHF must suppress capabilities the model has already acquired -- a more expensive operation that requires more

intensive gradient updates with correspondingly larger collateral capability damage. The jailbreaking vulnerability is a significant practical test of safety training robustness: GCG adversarial suffix attacks (Zou et al., 2023) succeed in bypassing RLHF-only safety on 48.4% of tested prompts in our evaluation, versus only 18.6% success against SAGT Full -- a 61.6% reduction in attack success rate. This robustness advantage reflects that SAGT has reduced the underlying harmful capabilities rather than merely suppressing their expression, making adversarial capability recovery substantially harder.

5.2 Limitations and Future Research

The SAGT evaluation is conducted at the 7B parameter scale on 500B pre-training tokens -- smaller than the pre-training runs of frontier models. The effectiveness of HWTL gradient weighting may diminish at larger scales where the training signal for harmful tokens is distributed across a much larger parameter space, requiring investigation of SAGT scaling behaviour. The harm classifier used for HWTL gradient weighting (DeBERTa-v3- small) has limited coverage of domain-specific harms (bioweapons synthesis, financial fraud) that require specialised safety expertise to classify; future SAGT implementations should incorporate domain expert annotation for high-risk harm categories. Future research should investigate the combination of SAGT with the UAG uncertainty quantification framework (Petrov and Hansen, 2025) to identify high-uncertainty outputs as a proxy for potentially harmful generation at inference time. The integration of SAGT with the NGA neurosymbolic architecture (Moreau and Silva, 2025) -- applying constitutional constraints as formal logical constraints rather than soft KL penalties -- represents a promising direction for higher-fidelity principle enforcement.

6. Conclusion

6.1 Summary

This paper proposed the SAGT framework integrating four safety modifications into LLM pre-training (HWTL, SCPDF, CSD, CPC). Evaluated against unaligned and RLHF-only baselines on LLaMA-2-7B: SAGT reduces HarmBench harmful rate by 61.4% at 3.8% capability tax -- versus RLHF-only 52.6% reduction at 12.3% tax. SCI improves from 0.621 (unaligned) to 0.828 (SAGT), versus 0.760 (RLHF-only). Jailbreaking attack success reduces from 48.4% (RLHF-only) to 18.6% (SAGT) -- 61.6% robustness improvement. SAGT+RLHF combined achieves SCI = 0.844 at 5.8% total tax.

6.2 Recommendations

For AI developers training frontier generative models, we recommend adopting SAGT components beginning with SCPDF (lowest compute overhead, substantial safety improvement) and HWTL (largest individual safety contribution at minimal capability cost). CPC and CSD require more implementation effort and should be adopted in Phase 2 after SCPDF and HWTL demonstrate safety returns. The combination of SAGT + RLHF (SCI = 0.844, 5.8% tax) is recommended as the best achievable safety-capability balance with current methods. For regulatory compliance with EU AI Act Article 9 risk management and Article 55 systemic risk requirements for GPAI, SAGT provides technically validated training-time safety measures that complement post-training alignment requirements. The SAGT implementation, harm classifier, SCI calculator, and evaluation suite are available as open-source resources. Technical details in Appendix A.

References

Askell, A. et al. (2021). A general language assistant as a laboratory for alignment. arXiv:2112.00861.

Bai, Y. et al. (2022). Training a helpful and harmless assistant with RLHF. arXiv:2204.05862.

Bai, Y. et al. (2022b). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Birhane, A. et al. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv:2110.01963.

Costa, M., and Garcia, L. (2025). Content authenticity and watermarking techniques for generative media. Journal of Generative Intelligence, 2025(2), 41-48.

Dubois, L., and Petrov, L. (2025). Bias detection and mitigation in generative AI systems. Journal of Generative Intelligence, 2025(2), 25-32.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Gao, L. et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027.

Laurenceon, H. et al. (2022). The BigScience ROOTS corpus. NeurIPS 2022 Datasets.

Lin, S. et al. (2022). TruthfulQA: Measuring how models mimic human falsehoods. ACL 2022.

Mazeika, M. et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming. ICML 2024.

Moreau, A., and Silva, I. (2025). Hybrid generative architectures combining symbolic AI and deep

learning. Journal of Generative Intelligence, 2025(2), 1-8.

Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. NeurIPS, 35.

Parrish, A. et al. (2022). BBQ: A hand-built bias benchmark. ACL Findings 2022.

Perez, E. et al. (2022). Red teaming language models with language models. arXiv:2202.03286.

Petrov, C., and Hansen, E. (2025). Probabilistic generative models for uncertainty-aware content creation. Journal of Generative Intelligence, 2025(2), 9-16.

Rafailov, R. et al. (2023). Direct preference optimization: Your language model is secretly a reward model. NeurIPS, 36.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

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

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2024 grant P2024YXTL2 (SAGT: Safety-Aware Generative Training). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The SAGT implementation, harm classifier, SCI calculator, evaluation benchmarks, and pre-trained model checkpoints are available as open-source resources. All evaluation results are available from the author.

Ethical Approval

This study involved computational experiments only. The evaluation uses publicly available safety benchmarks. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

Appendix A

SAGT Implementation Guide and SCI Calculation

The following provides SAGT component implementation details and SCI calculation methodology.

HWTL and SCPDF Implementation

CSD and CPC Implementation

SCI Calculation and Benchmark Protocol